An Honest Approach to Parallel Trends *

Ashesh Rambachan†   Jonathan Roth‡ *(Job market paper)*

December 18, 2019

Please click here for the latest version.

Abstract

Standard approaches for causal inference in difference-in-differences and event-study designs are valid only under the assumption of parallel trends. Researchers are typically unsure whether the parallel trends assumption holds, and therefore gauge its plausibility by testing for pre-treatment differences in trends ("pre-trends") between the treated and untreated groups. This paper proposes robust inference methods that do not require that the parallel trends assumption holds exactly. Instead, we impose restrictions on the set of possible violations of parallel trends that formalize the logic motivating pre-trends testing — namely, that the pre-trends are informative about what would have happened under the counterfactual. Under a wide class of restrictions on the possible differences in trends, the parameter of interest is set-identified and inference on the treatment effect of interest is equivalent to testing a set of moment inequalities with linear nuisance parameters. We derive computationally tractable confidence sets that are uniformly valid ("honest") so long as the difference in trends satisfies the imposed restrictions. Our proposed confidence sets are consistent, and have optimal local asymptotic power for many parameter configurations. We also introduce fixed length confidence intervals, which can offer finite-sample improvements for a subset of the cases we consider. We recommend that researchers conduct sensitivity analyses to show what conclusions can be drawn under various restrictions on the set of possible differences in trends. We conduct a simulation study and illustrate our recommended approach with applications to two recently published papers.

Keywords: Difference-in-differences, event-study, parallel trends, sensitivity analysis, robust inference, partial identification

Link to the online appendix.

---

†Harvard University, Department of Economics. Email: asheshr@g.harvard.edu

‡Harvard University, Department of Economics. Email: jonathanroth@g.harvard.edu

1   Introduction

Conventional methods for causal inference in difference-in-differences and related quasi-experimental research designs are valid only under the so-called "parallel trends" assumption.[1] Researchers are often unsure whether the parallel trends assumption is valid in applied settings, and they consequently try to assess its plausibility by testing for differences in trends between the treated and untreated groups prior to treatment ("pre-trends"). Conditional on not finding a significant pre-trend, inference is then typically conducted under the assumption that parallel trends holds *exactly*. The current approach to inference relying on an exact parallel trends assumption that is assessed via pre-trends testing suffers from two broad limitations. First, inference will be misleading if in fact the parallel trends assumption is violated but a pre-trends test fails to detect the violation (Freyaldenhoven, Hansen and Shapiro, 2019; Bilinski and Hatfield, 2018; Kahn-Lang and Lang, 2018; Roth, 2019). Second, we may still be interested in learning something about the causal effect of the treatment even if it is clear that there is a non-zero pre-existing trend, yet standard approaches deliver valid inference in this case only under strong functional form assumptions. Both of these limitations of the conventional approach motivate constructing methods for inference that allow the assumption of exact parallel trends to be relaxed.

The main contribution of this paper is to develop robust inference methods for difference-in-differences and related research designs that do not require that the parallel trends assumption holds exactly. Instead, the researcher need only specify a set of restrictions on the possible violations of parallel trends motivated by economic knowledge, and can conduct a sensitivity analysis to examine the robustness of their conclusions to different assumptions about such violations. We primarily consider polyhedral restrictions on the difference in trends, i.e. restrictions that can be written as a system of linear inequalities. A variety of economic intuitions that commonly arise in difference-in-differences and related settings can be formalized using such restrictions. For instance, the intuition behind the common practice of testing for pre-trends is that pre-treatment differences in trends are informative about counterfactual post-treatment differences in trends – a notion that we operationalize by restricting the degree of possible non-linearity in the difference in trends. We are also able to incorporate intuition about simultaneous policy changes or secular trends by imposing sign or shape restrictions on the difference in trends. Under these restrictions, the treatment effect of interest is typically set-identified.

To develop uniformly valid inference methods, we exploit the novel observation that the problem of conducting inference on the treatment effect of interest in our setting is equivalent to a moment inequality problem with linear nuisance parameters. A practical challenge to testing the implied set of moment inequalities is that the number of nuisance parameters scales linearly in the number of post-treatment periods, and thus will often be large (above 10) in typical empirical applications.

---

[1]The parallel trends assumption states that the average outcome for the treated and untreated groups would have moved in parallel if the treatment of interest had not occurred. This is also sometimes referred to as the "common trends" assumption (Angrist and Pischke, 2009; Lechner, 2011; Cunningham, 2018).

This renders many moment inequality procedures, which rely on test inversion over a grid for the full parameter vector, computationally infeasible. Andrews, Roth and Pakes (2019a, henceforth ARP) study conditional moment inequality problems with linear nuisance parameters, and propose confidence sets that condition on the set of sample moments that bind when using a linear program to profile out the nuisance parameters. We use this conditional inference procedure to obtain computationally tractable confidence sets that achieve uniform asymptotic size control. We then exploit additional structure in our setting to derive new results on the asymptotic performance of conditional confidence sets.[2]

We provide two novel results on the asymptotic properties of conditional confidence sets in our setting. First, we show that the conditional confidence sets are consistent, meaning that any fixed point outside of the identified set is rejected with probability approaching one asymptotically. Second, we provide a condition under which the conditional confidence sets have local asymptotic power converging to the power envelope — i.e., the upper bound on the power of any procedure that controls size uniformly. Intuitively, the condition for optimality requires that the underlying trend is not aligned with a "corner" of the allowed class in a particular sense. This condition is implied by, but somewhat weaker than, linear independence constraint qualification (LICQ), which has been used recently in the moment inequality settings of Gafarov (2019); Cho and Russell (2018); Flynn (2019); Kaido and Santos (2014).[3] In contrast to many previous uses of LICQ, however, we do not require this condition to obtain uniform asymptotic size control. To prove the optimality result, we make use of duality results from linear programming and the Neyman-Pearson lemma to show that both the optimal test and our conditional test converge to a t-test in the direction of the Lagrange multipliers of the linear program that profiles out the nuisance parameters.

Our results on the local asymptotic power of the conditional confidence sets have two limitations. First, the condition needed for the conditional confidence sets to have optimal local asymptotic power does not hold for all parameter values, and in particular fails when the parameter of interest is point-identified. Second, the conditional confidence sets may have low power in finite samples if the binding and non-binding moments are not well-separated.

We therefore introduce fixed length confidence intervals (FLCIs), which can address the limitations of the conditional approach in certain special cases of interest. We construct optimal FLCIs based on affine estimators for the treatment effect of interest, following Donoho (1994) and Armstrong and Kolesar (2018, 2019). We provide a novel characterization of when the optimal FLCI is consistent, meaning that any point outside of the identified set is rejected with probability approaching one asymptotically: the optimal FLCI is consistent *if and only if* the length of the identified set at the true population parameter equals its maximum possible length. This condition is non-restrictive for some classes of differential trends, but is restrictive (and even provably false)

---

[2]Specifically, we make use of the fact that the moments are linear in the target parameter and that the Jacobian of the moments with respect to the nuisance parameters is a known, constant matrix.

[3]Kaido, Molinari and Stoye (2019) show that constraint qualifications are closely related to other assumptions used for inference in the moment inequality literature.

for others. When this additional condition holds, however, the optimal FLCI achieves optimal local asymptotic power under the same conditions as the conditional confidence sets. Moreover, the optimal FLCI is near-optimal in finite sample in certain special cases of interest (Armstrong and Kolesar, 2018, 2019). In particular, the length of the optimal FLCI is close to the lower bound among all confidence sets that control size if parallel trends holds and the class of possible underlying differential trends is convex and centrosymmetric.

We next introduce a novel hybrid procedure that combines the relative strengths of the conditional confidence sets and FLCIs. The hybrid procedure is consistent, and has near-optimal local asymptotic power when the condition for the optimality of the conditional approach holds. Moreover, we find in simulations (discussed in more detail below) that hybridization with the FLCIs improves performance in finite sample when the moments are not well-separated. We thus recommend the hybrid approach for settings in which the condition for the consistency of the FLCIs is not guaranteed.[4]

For ease of exposition, the main text presents our results in a finite sample normal model. The finite sample normal model emerges as an asymptotic approximation to a wide range of econometric settings. In Appendix F, we show that these finite sample results translate to asymptotic results that hold uniformly over a large class of non-normal data-generating processes satisfying some high-level regularity conditions.

To explore the performance of our methods in practice, we conduct simulations calibrated to the 12 papers analyzed by Roth (2019), who systematically reviewed recent papers in three leading journals. We compute the expected excess length of each procedure, and compare this to an optimal benchmark, which is derived using the observation that the confidence set that minimizes expected excess length inverts most powerful tests (Armstrong and Kolesar, 2018). We find that for many parameter configurations, the conditional confidence sets have excess length within a few percent of this lower bound. However, as expected, the conditional approach can underperform in cases where the parameter of interest is (close to) point-identified. The FLCIs perform particularly well in the special cases where the conditions for their finite-sample near-optimality hold, but can have power substantially below the conditional test when these conditions fail; in one specification, the excess length of the conditional confidence sets is within 2% of the optimum, whereas for the FLCIs it is more than double the optimum. The hybrid approach performs quite well across a wide range of specifications — its excess length is no more than 17% larger than the other procedures across the specifications reported in the main text, whereas each of the other two procedures is at least 50% worse than another in some specification. We thus recommend the hybrid for most cases where the consistency of the FLCIs is not guaranteed. Lastly, all of the procedures remain computationally tractable in all specifications, including in cases with over 15 nuisance parameters.

---

[4]We also introduce a hybridization between the conditional confidence sets and tests based on least favorable (LF) critical values, which we recommend for cases where it is known that the FLCIs will be uninformative. Since the conditional-FLCI hybrid outperforms the conditional-LF hybrid in simulations for most parameter configurations, we focus mainly on the FLCI hybrid in the main text.

In practice, we recommend that applied researchers use our methods to conduct a sensitivity analysis in which they report confidence sets under varying restrictions on the set of possible violations of parallel trends. For instance, one class of restrictions we consider requires that the difference in trends not deviate "too much" from linearity, and is governed by a single parameter M that determines the degree of possible non-linearity. If the researcher is interested in testing a particular null hypothesis — e.g., the treatment effect in a particular period is zero — then a simple statistic to report is the "breakdown" value of M at which the null hypothesis of interest can no longer be rejected.[5] We discuss how one can benchmark the value of this parameter using economic knowledge of possible confounds, as well as placebo periods or treatment groups. The researcher can also report how her conclusions change with the inclusion of additional sign or shape restrictions motivated by context-specific knowledge. For instance, in cases where researchers are concerned about secular trends correlated with treatment, researchers may specify that the underlying difference in trends is monotone. Likewise, in cases with known simultaneous policy changes, researchers may restrict the sign of the bias. Performing such a sensitivity analysis and documenting how the reported confidence set changes across a range of assumptions makes clear what must be assumed about the possible violations of parallel trends in order to draw specific conclusions. Our publicly-available R package `HonestDiD` provides functions for easy and fast implementation of our recommended methods.[6]

We illustrate our empirical recommendations by applying our approach to two recently-published papers. First, we consider an application to Lovenheim and Willen (2019, henceforth LW), who study how exposure to public sector duty-to-bargain (DTB) laws as a child affects long-run labor market outcomes. LW find no clear pre-trends for male employment, but do find substantial pre-trends for female employment, and therefore focus on their results for men. The example allows us to illustrate how we can gauge the robustness of the results for men to different assumptions on the possible violations of parallel trends, as well as how we can use such assumptions to construct confidence sets for the effect on women despite the presence of a pre-trend. Second, we consider an application to Benzarti and Carloni (2019), who study the impact of a decrease in the value-added tax (VAT) on the profitability of restaurants in France, and find large effects of the VAT decrease on restaurant profitability. We show that the conclusion that restaurant profits increase in the first year after the VAT cut are robust to a wide range of assumptions about the underlying difference in trends, but conclusions about the longer-run effects (e.g., four years after the VAT decrease) appear to be more fragile. This illustrates a general point: when allowing for the possibility of secular differential trends, bias accumulates over time, and so longer-run estimates will typically be more sensitive to violations of parallel trends than short-run estimates.

---

[5]Similar "breakdown" concepts have been proposed in other partially identified settings (Horowitz and Manski, 1995; Kline and Santos, 2013; Masten and Poirier, 2019).

[6]The latest version of the R package may be downloaded here.

Related Literature   This paper contributes to a large and active literature on the economet-
rics of difference-in-differences and related research designs by developing robust inference methods
that allow for the possibility that parallel trends may be violated.[7] Several other papers consider
methods for relaxing or circumventing the assumption of exact parallel trends. Manski and Pepper
(2018) consider how the set of identified parameters changes as we relax the parallel trends assump-
tion, but do not consider the problem of statistical inference. Keele, Small, Hsu and Fogarty (2019)
likewise develop techniques for testing the sensitivity of difference-in-differences designs to violations
of the parallel trends assumption, but they do not incorporate information from pre-trends in their
sensitivity analysis. A common approach when there are concerns about violations of the parallel
trends assumption is to adjust for the extrapolation of a linear trend from the pre-treatment pe-
riod (cf. Dobkin, Finkelstein, Kluender and Notowidigdo, 2018; Goodman-Bacon, 2018a,b; Bhuller,
Havnes, Leuven and Mogstad, 2013). This approach provides valid inference only under the restric-
tion that the underlying trend is exactly linear. Our approach nests this restriction as a special
case, but allows for less restrictive assumptions on the class of underlying trends; for instance,
one class of restrictions we consider requires only that the linear extrapolation be approximately
correct. Freyaldenhoven et al. (2019) propose a method that allows for violations of the parallel
trends assumption but requires a covariate that is known to be affected by the confounds but not
the treatment of interest.[8]

Our robust approach to inference helps to address several concerns related to conventional
practice in difference-in-differences and related research designs. First, a recent literature suggests
that common tests for pre-trends may be underpowered against meaningful violations of parallel
trends, potentially leading to severe undercoverage of conventional confidential intervals, which are
valid only if parallel trends holds exactly (Freyaldenhoven et al., 2019; Bilinski and Hatfield, 2018;
Kahn-Lang and Lang, 2018). For instance, Roth (2019) finds that linear violations of parallel trends
producing bias equal to the estimated treatment effect would be detected less than half the time
by conventional pre-tests in simulations calibrated to several recent papers in leading economics
journals. Second, statistical distortions from pre-testing for pre-trends may further undermine
the performance of conventional inference procedures (Roth, 2019). Third, although there exist
parametric approaches to controlling for pre-existing trends, there are concerns that such methods
are sensitive to functional form assumptions (Wolfers, 2006; Lee and Solon, 2011). Our paper helps
to address these issues by providing tools for inference that do not rely on an exact parallel trends

---

[7]Previous work on inference in difference-in-differences and related designs has addressed robustness to
serial correlation and clustered sampling (Arellano, 1987; Moulton, 1990; Bertrand, Duflo and Mullainathan,
2004; Sun and Yan, 2019), and concerns related to having a small number of treated units (Donald and Lang,
2007; Conley and Taber, 2010; Ferman and Pinto, 2018; MacKinnon and Webb, 2019), among other issues.

[8]In practice, one may be unsure if a covariate satisfies the exclusion restriction required by Freyaldenhoven
et al. (2019). In their applications, Freyaldenhoven et al. (2019) therefore create an event-study plot that
shows estimated placebo pre-treatment effects using their method, which are analogous to pre-period coeffi-
cients in a typical event-study. Since their procedure yields asymptotically normal estimates, one can apply
the methods in this paper to assess the sensitivity of results using their method under various assumptions
on the relationship between pre-treatment and post-treatment violations of the exclusion restriction.

assumption and that make clear the mapping between the restrictiveness of the assumptions on the potential differences in trends and the strength of one's conclusions.

Our work is also complementary to a recent literature on the causal interpretation of the identified coefficients in two-way fixed effects models in settings with staggered treatment timing or heterogeneous treatment effects (Meer and West, 2016; Borusyak and Jaravel, 2016; Abraham and Sun, 2018; Hull, 2018; Athey and Imbens, 2018; de Chaisemartin and D'Haultfœuille, 2018a,b; Goodman-Bacon, 2018a; Kropko and Kubinec, 2018; Callaway and Sant'Anna, 2019; Imai and Kim, 2019; Słoczyński, 2018). A key finding of this literature is that regression coefficients from conventional approaches may not produce convex weighted averages of local average treatment effects even if parallel trends holds. Additionally, conventional event-study estimators may exhibit spurious pre-trends when parallel trends holds (Abraham and Sun, 2018). The literature has thus proposed a number of alternative strategies that allow for consistent estimation of sensible causal estimands under a suitable parallel trends assumption. We encourage researchers to start by choosing an estimation strategy that yields consistent estimates of the causal parameter of interest under the best-case assumption of parallel trends. We then recommend using our methods to assess the robustness of conclusions to potential violations of the parallel trends assumption.[9]

More broadly, this paper relates to a large literature in econometrics on sensitivity analysis and misspecification robust inference. Recent work has studied misspecification in the context of instrumental variables models (Conley, Hansen and Rossi, 2012), generalized method of moments (Andrews, Gentzkow and Shapiro, 2017, 2019b; Armstrong and Kolesar, 2019), parametric and semi-parametric models (Bonhomme and Weidner, 2018; Mukhin, 2018), regression discontinuity designs (Kolesar and Rothe, 2018; Armstrong and Kolesar, 2018), structural models (Christensen and Connault, 2019), and experimental design (Rosenbaum and Rubin, 1983; Imbens, 2003; Rosenbaum, 2005; Lee, 2009; Ding and VanderWeele, 2016), among many others. An interesting feature of our setting is that we use restrictions to relate the degree of bias in the treatment effects estimates for periods after treatment to pre-trends identified in periods prior to treatment. Our approach thus shares similar structure with papers in sensitivity analysis that relate bias from selection into treatment on unobservables to selection on observables in cross-sectional contexts (Altonji, Elder and Taber, 2005; Oster, 2019).

Finally, our work connects to the rich literature on partial identification and moment inequality procedures in econometrics. As in Manski and Pepper (2018), we consider relaxations of the assumption of exact parallel trends, under which the treatment effect of interest is partially identified in difference-in-differences and related settings. The use of partial identification in the analysis of treatment effects in other observational or quasi-experimental settings dates back to at least Manski (1990), and has been used since by Balke and Pearl (1997), Heckman and Vytlacil (1999), Manski and Nagin (1998), Manski and Pepper (2000), Ginther (2000), González (2005), Bhattacharya,

---

[9]Our approach can be applied so long as the estimator is asymptotically normally distributed, which covers the vast majority of proposed procedures in difference-in-differences and related research designs.

Shaikh and Vytlacil (2008), Nevo and Rosen (2012), Kreider, Pepper, Gundersen and Jolliffe (2012), and Gunsilius (2019), among many others. We show that testing a null hypothesis on the treatment effect of interest under a relaxation of the parallel trends assumption can be cast as a moment inequality problem, and derive novel asymptotic consistency and optimality results for confidence sets constructed using the conditional approach of ARP. We also propose a new hybrid procedure that combines fixed-length confidence intervals with conditional confidence sets. This hybrid procedure may be useful in other settings. Reviews of the wide range of applications of partial identification and moment inequality methods in economics include Manski (2003), Tamer (2010), and more recently Ho and Rosen (2017), Canay and Shaikh (2017), and Molinari (2019).

2 Motivating example: Difference-in-differences

We begin with a stylized three-period difference-in-differences model, which allows us to highlight the causal parameter of interest, inferential goal, and key assumptions of our approach in a simple setting. We later introduce the general setting that will be our main focus in Section 3.

2.1 Data-generating process

We observe an outcome Y_{it} for a sample of individuals $i = 1, \ldots, N$ for three time periods, $t = -1, 0, 1$. Individuals in the treated group ($D_i = 1$) receive a treatment of interest between period $t = 0$ and $t = 1$.[10] The observed outcome can be expressed as $Y_{it} = D_i Y_{it}(1) + (1 - D_i) Y_{it}(0)$, where $Y_{it}(1)$ and $Y_{it}(0)$ are the potential outcomes for an individual i in period t if she were or were not assigned to the treated group. We assume that the treatment has no causal effect before it is implemented, so that $Y_{it}(1) = Y_{it}(0)$ for $t < 1$.[11] We are interested in the average treatment effect on the treated (ATT) in the period after treatment, $\tau_{ATT} = \mathbb{E}\left[Y_{i,t=1}(1) - Y_{i,t=1}(0) \mid D_i = 1\right]$.

In this setting, researchers commonly estimate the regression specification,

$$Y_{it} = \lambda_i + \phi_t + \sum_{s \neq 0} \beta_s \times \mathbb{1}[t = s] \times D_i + \epsilon_{it}. \tag{1}$$

This specification is often referred to as a "dynamic event-study regression" or a "two-way fixed effects model" with dynamic treatment effects (Borusyak and Jaravel, 2016; de Chaisemartin and D'Haultfœuille, 2018b). It is well-known that in this simple setting, the coefficient $\hat{\beta}_1$ can be written as the "difference-in-differences" of sample means across treated and untreated groups between period $t = 0$ and $t = 1$. That is, the "post-period" coefficient is

---

[10]Formally, for the purposes of this example, we think of our sample of size $N = N_0 + N_1$ as consisting of N_1 independent draws from the treated ($D = 1$) population and N_0 independent draws from the control population ($D = 0$), as in Abadie and Imbens (2006).

[11]An important case where this could be violated is if individuals modify their behavior in anticipation of the reform (Malani and Reif, 2015). See Section 9 for additional discussion of anticipation.

$$\hat{\beta}_1 = (\bar{Y}_{D=1,t=1} - \bar{Y}_{D=1,t=0}) - (\bar{Y}_{D=0,t=1} - \bar{Y}_{D=0,t=0}),$$

where $\bar{Y}_{D=d,t=s}$ is the sample mean of Y_{it} for treatment group d in period s. The "pre-period" coefficient $\hat{\beta}_{-1}$ can be expressed similarly as

$$\hat{\beta}_{-1} = (\bar{Y}_{D=1,t=-1} - \bar{Y}_{D=1,t=0}) - (\bar{Y}_{D=0,t=-1} - \bar{Y}_{D=0,t=0}).$$

Taking expectations and re-arranging, we see that

$$\mathbb{E}\left[\hat{\beta}_1\right] = \tau_{ATT} + \underbrace{\mathbb{E}\left[Y_{i,t=1}(0) - Y_{i,t=0}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,t=1}(0) - Y_{i,t=0}(0) \mid D_i = 0\right]}_{\text{Post-period differential trend} := \delta_1},$$

$$\mathbb{E}\left[\hat{\beta}_{-1}\right] = \underbrace{\mathbb{E}\left[Y_{i,t=-1}(0) - Y_{i,t=0}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,t=-1}(0) - Y_{i,t=0}(0) \mid D_i = 0\right]}_{\text{Pre-period differential trend} := \delta_{-1}}.$$

Thus, the post-period regression coefficient $\hat{\beta}_1$ is biased for the treatment effect τ_{ATT} if $\delta_1 \neq 0$. The bias δ_1 is the differential trend in outcomes for the two groups between period 0 and period 1 that would have occurred if the treatment counterfactually did not take place. Likewise, δ_{-1} is the pre-period differential trend between groups from period 0 to period -1. The parallel trends assumption imposes that the expectation of $Y_{it}(0)$ moves in parallel for the two groups, $\delta_{-1} = \delta_1 = 0$, in which case $\hat{\beta}_1$ is unbiased for the treatment effect of interest.

2.2   Inferential goal

Applied to this illustrative three period difference-in-differences example, this paper considers the problem of conducting inference on τ_{ATT} while relaxing the assumption that δ_1 is exactly zero. Of course, if we allow δ_1 to be completely unrestricted, then the data will be uninformative about the value of τ_{ATT}. However, the motivating logic behind the common practice of pre-trends testing is that the pre-period difference in trends δ_{-1} is informative about the counterfactual post-period difference in trends δ_1. We thus may be willing to place restrictions on the possible values of δ_1 given δ_{-1}.

We formalize this intuition by assuming that $\delta = (\delta_{-1}, \delta_1)' \in \Delta$, where Δ is some class of possible differential trends that is specified by the researcher. The usual parallel trends assumption is thus the special case where $\Delta = \{0\}$. Given a set Δ, we then construct confidence sets that are uniformly valid over a wide class of distributions with $\delta \in \Delta$. That is, for $\mathcal{P}$ a class of distributions P such that $\delta \in \Delta$ under $\mathcal{P}$, we construct confidence sets $\mathcal{C}_n$ satisfying

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \mathbb{P}_P\left(\tau \in \mathcal{C}_n\right) \geqslant 1 - \alpha. \tag{2}$$

A confidence set $\mathcal{C}_n$ that satisfies this criterion is sometimes referred to as "uniform" or "honest" (Andrews, Cheng and Guggenberger, Forthcoming; Li, 1989; Armstrong and Kolesar, 2018).

In practice, the class of allowed differential trends Δ will often be parameterized by M, which indexes the extent to which the counterfactual post-period difference in trends can differ from an extrapolation of the pre-period difference in trends. We recommend that researchers report confidence sets under a variety of values of M to give the reader a sense of how much regularity needs to be imposed *ex ante* to obtain informative inference. We now discuss a few specifications for Δ that may be reasonable in empirical applications.

2.3 Possible choices of Δ

The class of possible violations of parallel trends Δ must be specified by the researcher, and the choice of Δ will depend on the economic context. Here, we highlight several choices of Δ that may be reasonable in applications. These examples formalize intuitive arguments that are made in the literature regarding possible violations of parallel trends. We discuss possible restrictions in the stylized three-period model presented above, and also discuss how these restrictions can be generalized to cases with multiple periods, as will be common in applications.

Bounds on changes in slope. In practice, researchers concerned about possible violations of the parallel trends assumption often include treatment-group specific linear trends.[12] In the motivating three period model, such an approach will recover the causal parameter τ_{ATT} under the assumption that the differential trend is exactly linear, $\delta_1 = -\delta_{-1}$, as shown in Figure 1. There are concerns, however, that this linear extrapolation of the pre-period trend to the post-period may not hold exactly (Wolfers, 2006; Lee and Solon, 2011). A natural relaxation of the linear trend model is to require only that the linear extrapolation be *approximately* correct, meaning that $\delta_1 \in [-\delta_{-1} - M, -\delta_{-1} + M]$, where M governs the maximum possible error of the linear extrapolation (i.e. the change in slope of the differential trend). This corresponds with setting Δ equal to

$$\Delta^{SD}(M) := \{(\delta_{-1}, \delta_1)' \,:\, \delta_1 \in -\delta_{-1} \pm M\}.$$
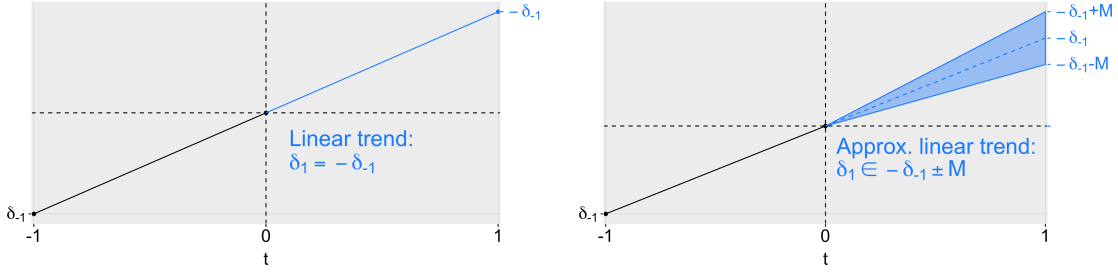
This restriction can easily be extended to settings with additional periods by bounding the extent to which the slope of the differential trend can change between consecutive periods. For instance, in Section 3 we consider a more general model where the researcher estimates $\underline{T}$ pre-event

---

[12]That is, instead of estimating specification (1), researchers estimate

$$Y_{it} = \lambda_i + \phi_t + \beta_{trend} \times D_i \times t + \sum_{s>0} \beta_s \times \mathbb{1}[t = s] \times D_i + \epsilon_{it},$$

which Dobkin et al. (2018) refer to as a "parametric event-study." An analogous approach is to estimate a linear trend using only observations prior to treatment, and then subtract out the estimated linear trend from the observations after treatment (Bhuller et al., 2013; Goodman-Bacon, 2018a,b).

Figure 1: Linear and Approximately Linear Trends

coefficients and $\bar{T}$ post-event coefficients (with period $t = 0$ again normalized to 0). In this context, δ is a $\underline{T} + \bar{T}$ dimensional vector describing the counterfactual differential trend over the observed sample period, and we can impose approximately linearity by requiring that δ lie in the set

$$\Delta^{SD}(M) := \{\delta \,:\, |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leqslant M, \, \forall t\}, \tag{3}$$

where we adopt the convention that $\delta_0 = 0$.[13] The parameter $M \geqslant 0$ again governs the amount by which the slope of δ can change between consecutive periods, and in the special case where $M = 0$, Δ^{SD} requires that the difference in trends be linear. $\Diamond$

Bounds on relative magnitudes In some cases, we may be willing to assume that if the pre-period difference in trends is small, then the counterfactual post-period difference in trends would also be small. However, if the two groups did not follow similar trends in the pre-period, then we may find it plausible that the difference in trends between the two groups would have changed substantially between the pre- and post-periods. One way to formalize this intuition in the context of our 3-period model is to allow the magnitude of δ_1 to depend on the magnitude of δ_{-1},

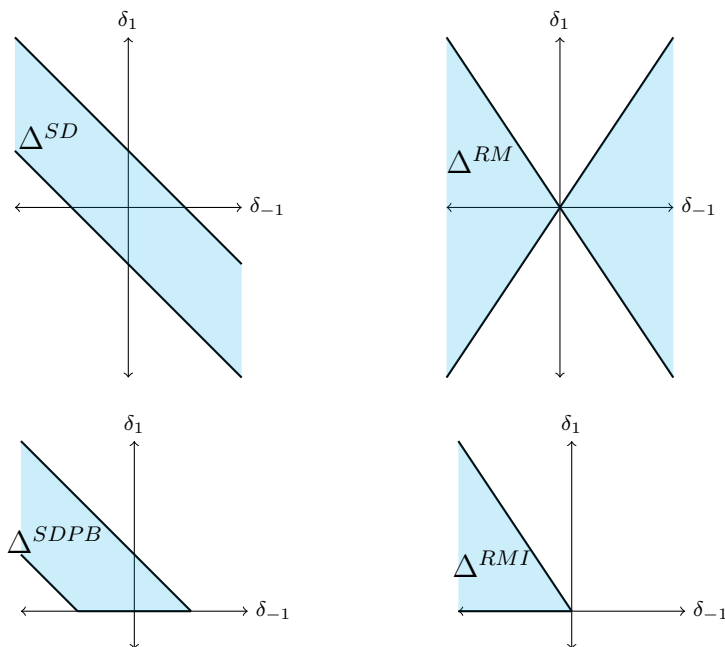$$\Delta^{RM}(\bar{M}) = \{(\delta_{-1}, \delta_1) \,:\, |\delta_1| \leqslant \bar{M}|\delta_{-1}|\}.$$

This type of restriction can be extended to multiple periods by bounding the percentage change in the slope of the differential trend,

$$\Delta^{RM}(\bar{M}) := \{\delta \,:\, |\delta_{t+1} - \delta_t| \leqslant \bar{M}|\delta_t - \delta_{t-1}|, \, \forall t\}. \tag{4}$$

The parameter $\bar{M} \geqslant 0$ places an upper bound on the percentage change in the magnitude of the slope between periods. $\Diamond$

---

[13]$\Delta^{SD}(M)$ bounds the discrete analog of the second derivative of δ, and is thus similar to restrictions on the second derivative of the conditional expectation function or density in regression discontinuity settings (Kolesar and Rothe, 2018; Frandsen, 2016). Smoothness restrictions are also used to obtain partial identification in Kim, Kwon, Kwon and Lee (2018).

Figure 2: Example choices for Δ

Note: This figure shows diagrams of potential restrictions Δ on the set of possible violations of parallel trends. Δ^{SD} requires approximate linearity by restricting the change in slope of the differential trend. Δ^{RM} bounds the magnitude of the post-period bias relative to the magnitude of the pre-trend. Δ^{SDPB} adds the restriction that the post-period bias be positive to Δ^{SD}. Likewise, Δ^{RMI} adds the restriction that the differential trend be increasing over time to Δ^{RM}.

Sign and monotonicity restrictions Context-specific knowledge may sometimes imply sign or shape restrictions on the differential trend. For instance, suppose we know of a simultaneous, confounding policy change that we expect would have a positive effect on the outcome of interest. In this case, we might restrict that the bias in the post-treatment period be positive,

$$\delta \in \Delta^{PB} := \{\delta \,:\, \delta_t \geqslant 0 \;\forall t \geqslant 0\}.$$

Likewise, it may be reasonable to impose monotonicity in cases where we are concerned about secular trends that we expect would have continued following the date of treatment.[14] For instance, we could impose that the differential trend be increasing,

---

[14]The discussion of possible violations of parallel trends in applied work often implicitly assumes monotonicity. For example, Lovenheim and Willen (2019) argue that violations of parallel trends cannot explain their results because "pre-[treatment] trends are either zero or in the wrong direction (i.e., opposite to the direction of the treatment effect)." Greenstone and Hanna (2014) estimate upward-sloping pre-existing trends and argue that "if the pre-trends had continued" their estimates would be upward biased. The so-called Ashenfelter (1978)'s dip in job training programs is a well-known exception wherein we expect violations of parallel trends to be non-monotonic.

$$\delta \in \Delta^I := \{\delta \, : \, \delta_t \geqslant \delta_{t-1} \, \forall t\}.$$

These sign and monotonicity restrictions can also be combined with the bounds on changes in slope or relative magnitudes imposed above. For example, we will define $\Delta^{SDPB}(M) := \Delta^{SD}(M) \cap \Delta^{PB}$ and $\Delta^{RMI}(\bar{M}) := \Delta^{RM}(\bar{M}) \cap \Delta^I$. $\diamondsuit$

Figure 2 gives a geometric depiction of the sets $\Delta^{SD}, \Delta^{RM}, \Delta^{SDPB}$, and Δ^{RMI} in the simple case in which there is one pre-period and one post-period coefficient.

3 General Set-up

We now introduce the assumptions, target parameter, and inferential goal considered throughout the remainder of the paper. We consider a finite-sample normal model, which arises as an approximation to the motivating model discussed in Section 2 and to a variety of other econometric settings of interest. In Appendix F, we show that our results derived in the context of the finite-sample normal model hold uniformly over a large class of non-normal data-generating processes.

3.1 Finite sample normal model

We consider the model,

$$\hat{\beta}_n \sim \mathcal{N}\left(\beta, \, \Sigma_n\right), \tag{5}$$

where $\hat{\beta}_n \in \mathbb{R}^{\underline{T}+\bar{T}}$, and $\Sigma_n = \frac{1}{n}\Sigma^*$ for Σ^* a known, positive-definite $(\underline{T} + \bar{T}) \times (\underline{T} + \bar{T})$ matrix. We partition the event-study coefficients $\hat{\beta}_n$ into vectors corresponding with the pre-treatment and post-treatment periods, $\hat{\beta}_n = (\hat{\beta}'_{n,pre}, \hat{\beta}'_{n,post})'$, where $\hat{\beta}_{n,pre} \in \mathbb{R}^{\underline{T}}$ and $\hat{\beta}_{n,post} \in \mathbb{R}^{\bar{T}}$. We adopt analogous notation to partition other vectors that are the same length as $\hat{\beta}_n$ into *pre* and *post* components.

This finite sample normal model (5) can be viewed as an asymptotic approximation to a wide range of econometric settings. In particular, under mild regularity conditions, a variety of estimation strategies will yield asymptotically normal event-study estimates, $\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma^*\right)$.[15] This convergence in distribution suggests the finite-sample approximation $\hat{\beta}_n \stackrel{d}{\approx} \mathcal{N}\left(\beta, \Sigma_n\right)$, where $\stackrel{d}{\approx}$ denotes approximate equality in distribution and $\Sigma_n = \frac{1}{n}\Sigma^*$. In the main text of the paper, we derive results assuming this equality in distribution holds exactly in finite samples, and in Appendix

---

[15]Examples of data-generating processes that yield asymptotically normal event-study estimates include the two-way fixed effects model (1), the GMM procedure proposed by Freyaldenhoven et al. (2019), instrumental variables event-studies (Hudson, Hull and Liebersohn, 2017), the estimation strategies of Abraham and Sun (2018) and Callaway and Sant'Anna (2019) to address issues with non-convex weights on cohort-specific effects in staggered treatment designs, as well as a range of procedures that flexibly control for differences in covariates between treated and untreated groups (e.g., Heckman, Ichimura, Smith and Todd, 1998; Abadie, 2005; Sant'Anna and Zhao, 2019).

F, we translate our results into uniform asymptotic statements that hold over a large class of non-normal data-generating processes.

We assume that the mean vector β satisfies the following decomposition.

Assumption 1. *The parameter vector β can be decomposed as*

$$\beta = \underbrace{\begin{pmatrix} \tau_{pre} \\ \tau_{post} \end{pmatrix}}_{:= \tau} + \underbrace{\begin{pmatrix} \delta_{pre} \\ \delta_{post} \end{pmatrix}}_{:= \delta}, \tag{6}$$

with $\tau_{pre} \equiv 0$.

The first term, τ, represents the time path of dynamic causal effects of interest. We assume that the treatment has no causal effect prior to its implementation, so $\tau_{pre} = 0$. The second term, δ, represents the difference in trends between the treated and untreated groups that would have occurred absent treatment.[16] In this general setting, the parallel trends assumption corresponds to imposing that $\delta = 0$. Therefore, under parallel trends, $\beta = \tau$.

3.2   Target parameter and identification

The parameter of interest is a scalar, linear combination of the post-treatment causal effects, $\theta :=$ $l'\tau_{post}$ for some known $\bar{T}$-vector l. For example, the parameter θ is the t-th period causal effect τ_t when the vector l equals the t-th standard basis vector. Similarly, θ is the average causal effect across all of the post-treatment periods when $l = \left(\frac{1}{T}, ..., \frac{1}{T}\right)'$. Under the usual exact parallel trends assumption ($\delta = 0$), θ is point-identified.

We relax the exact parallel trends assumption by assuming only that δ lies in some set of possible violations of parallel trends Δ, which is specified by the researcher. The usual parallel trends assumption is thus a special case with $\Delta = \{0\}$. Under the assumption that $\delta \in \Delta \neq \{0\}$, the parameter θ will typically only be set-identified (rather than point-identified). For a given value of β, the set of values θ consistent with β under the assumption that $\delta \in \Delta$ is

$$\mathcal{S}(\Delta, \beta) := \left\{ \theta \; : \; \exists \delta \in \Delta, \tau_{post} \in \mathbb{R}^{\bar{T}} \text{ s.t. } l'\tau_{post} = \theta, \beta = \delta + \begin{pmatrix} 0 \\ \tau_{post} \end{pmatrix} \right\}, \tag{7}$$

which we refer to as the identified set.

We note that the identified set has a simple characterization when Δ is a convex set. In particular, it is clear from (7) that if Δ is convex, then so too is $\mathcal{S}(\Delta, \beta)$, and thus the identified

---

[16]In settings with staggered treatment timing and treatment effect heterogeneity, the vector τ_{post} that is identified by a two-way fixed effects specification may represent some weighted sum of cohort-specific effects. Researchers should take care to specify their estimator in such a way that τ_{post} has a meaningful, causal interpretation. See, e.g., recent work by Abraham and Sun (2018), among others.

set is an interval in $\mathbb{R}$. Furthermore, re-arranging terms in the definition given in (7), the identified set can be equivalently written as

$$\mathcal{S}(\Delta, \beta) = \{\theta \, : \, \exists \delta \in \Delta \text{ s.t. } \delta_{pre} = \beta_{pre}, \theta = l'\beta_{post} - l'\delta_{post}\}. \tag{8}$$

It is then immediate from (8) that the lower bound of the interval $\mathcal{S}$ is

$$\theta^{lb} := l'\beta_{post} - \underbrace{\left(\max_{\delta} l'\delta_{post}, \text{ s.t. } \delta \in \Delta, \delta_{pre} = \beta_{pre} \right)}_{=:b^{max}(\beta_{pre};\Delta)}. \tag{9}$$

Likewise, the upper bound is given by

$$\theta^{ub} := l'\beta_{post} - \underbrace{\left(\min_{\delta} l'\delta_{post}, \text{ s.t. } \delta \in \Delta, \delta_{pre} = \beta_{pre} \right)}_{=:b^{min}(\beta_{pre};\Delta)}. \tag{10}$$

3.3 Inferential goal

Our goal is then to construct confidence sets that are valid for all parameter values θ in the identified set. That is, we wish to construct sets $\mathcal{C}_n$ satisfying

$$\inf_{\delta \in \Delta, \tau} \inf_{\theta \in \mathcal{S}(\Delta, \delta + \tau)} \mathbb{P}_{(\delta, \tau, \Sigma_n)} \left(\theta \in \mathcal{C}_n \right) \geqslant 1 - \alpha, \tag{11}$$

which is the finite-sample analog of the uniform asymptotic coverage criterion (2). Although in the main text we focus on this finite-sample coverage criterion, in Appendix F we show how our finite-sample coverage results translate to uniform asymptotic coverage results over a wide class of non-normal data-generating processes. We note that (11) requires proper coverage of the true parameter of interest $\theta = l'\tau_{post}$, rather than coverage of the full identified set (Imbens and Manski (2004)). Note also that in (11), we subscript the probability operator by (δ, τ, Σ_n) to make explicit that the distribution of $\hat{\beta}_n$ (and hence $\mathcal{C}_n$) depends on these parameters. We adopt this notation throughout the paper.

4 Conditional Confidence Sets

In this section, we describe our main procedure for constructing robust confidence sets. We focus on the case where the set of possible violations Δ takes the polyhedral form, $\Delta = \{\delta \, : \, A\delta \leqslant d\}$, which covers many leading examples. We first show that when the set Δ takes this form, the problem of conducting robust inference in difference-in-differences and event-study designs is equivalent to a moment inequality problem with linear nuisance parameters. The dimension of the nuisance parameters scales linearly with the number of post-treatment periods, and thus will frequently be

large (above 10) in practical applications. We next show that the conditioning approach of ARP, which exploits the linear structure of the problem, can be used to obtain confidence sets that satisfy the uniform coverage restriction (11) and remain computationally tractable in practice. In Section 5, we then derive novel results on the power of this procedure in our context.

4.1   Polyhedral forms for Δ

As mentioned, we focus on classes Δ that take a polyhedral form, meaning it can be expressed as a series of linear restrictions on δ.

Assumption 2 (Polyhedral shape restriction). *The class Δ takes the form*

$$\Delta = \{\delta \ : \ A\delta \leqslant d\}, \tag{12}$$

for some matrix A and vector d, where the matrix A has no all-zero rows.

While this assumption is not without loss of generality, all but one of the leading examples provided in Section 2.3 can be written in this form. Intuitively, the set Δ is a polyhedron if it is a convex set with flat sides. It is thus immediate from Figure 2 that with two dimensions, $\Delta^{SD}, \Delta^{SDPB}$, and Δ^{RMI} can all be written as polyhedra, and the geometric intuition from the two-period case extends easily to higher dimensions.[17] We also can see from Figure 2 that Δ^{RM} is not convex and thus not a polyhedron, although it can be expressed as the union of polyhedra. One can thus form a confidence set for Δ^{RM} by taking the union of the confidence sets for each of the polyhedra that compose it.

4.2   Representation as a moment inequality problem with linear nuisance parameters

Consider the problem of testing the null hypothesis, $H_0 : \theta = \bar{\theta}, \delta \in \Delta$ when $\Delta = \{\delta \ : \ A\delta \leqslant d\}$. In this section, we will show that testing H_0 is equivalent to testing a system of moment inequalities with linear nuisance parameters.

Our model implies $\mathbb{E}_{(\delta,\tau,\Sigma_n)}\left[\hat{\beta}_n - \tau\right] = \delta$, and hence $\delta \in \Delta$ if and only if $\mathbb{E}_{(\delta,\tau,\Sigma_n)}\left[A\hat{\beta}_n - A\tau\right] \leqslant d$. Define $Y_n = A\hat{\beta}_n - d$ and let $M_{post} = [0, I]'$ be the matrix such that $\tau = M_{post}\tau_{post}$. It is then immediate that the null hypothesis H_0 is equivalent to the composite null

$$H_0 : \exists \tau_{post} \in \mathbb{R}^{\bar{T}} \text{ s.t. } l'\tau_{post} = \bar{\theta} \text{ and } \mathbb{E}_{(\delta,\tau,\Sigma_n)}\left[Y_n - AM_{post}\tau_{post}\right] \leqslant 0. \tag{13}$$

---

[17]For example, recall that $|x| \leqslant M$ is equivalent to $x \leqslant M$ and $-x \leqslant M$. In the case with one pre-period and one post-period, we can thus write $\Delta^{SD}(M) = \{\delta \ : \ A^{SD}\delta \leqslant d^{SD}\}$ for $A^{SD} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ and $d^{SD} = \begin{pmatrix} M \\ M \end{pmatrix}$. This generalizes naturally when there are multiple pre-periods and multiple post-periods.

Here, $\tau_{post} \in \mathbb{R}^{\bar{T}}$ is a vector of nuisance parameters that must satisfy the linear constraint $l'\tau_{post} = \bar{\theta}$ under H_0'.

By applying a change of basis, we can further re-write H_0 as an equivalent composite null hypothesis with an unconstrained nuisance parameter. In particular, we can re-write the expression $AM_{post}\tau_{post}$ as $\tilde{A} \begin{pmatrix} \theta \\ \tilde{\tau} \end{pmatrix}$, where $\tilde{A}$ is the matrix that results from applying a suitable change of basis to the columns of AM_{post}, and $\tilde{\tau} \in \mathbb{R}^{\bar{T}-1}$.[18] We then see that H_0 is equivalent to

$$H_0 : \exists \tilde{\tau} \in \mathbb{R}^{\bar{T}-1} \text{ s.t. } \mathbb{E}\left[\tilde{Y}_n(\bar{\theta}) - \tilde{X}\tilde{\tau} \right] \leqslant 0, \tag{14}$$

where $\tilde{Y}(\theta) = Y_n - \tilde{A}_{(\cdot,1)}\bar{\theta}$ and $\tilde{X} = \tilde{A}_{(\cdot,-1)}$. Note that under the finite-sample normal model (5), $\tilde{Y}_n(\bar{\theta})$ is normally distributed with covariance matrix $\tilde{\Sigma}_n = A\Sigma_n A'$. We therefore have shown that testing the null hypothesis $H_0 : \theta = \bar{\theta}$ is equivalent to testing a set of moment inequalities with linear nuisance parameters.

4.3 Constructing conditional confidence sets

An important practical consideration for testing hypotheses of the form (14) is that the dimension of the nuisance parameter $\tilde{\tau} \in \mathbb{R}^{\bar{T}-1}$ may be large in practice. For instance, in Section 10 we consider a recent paper in which $\bar{T} = 23$. Moreover, 5 of the 12 recent event-study papers reviewed in Roth (2019) have $\bar{T} > 10$. This renders many moment inequality methods, particularly those which rely on test inversion over a grid for the full parameter vector, practically infeasible in this context. We now show how the conditional approach of ARP, which directly exploits the linear structure of the hypothesis (14), can be applied to obtain computationally tractable and powerful tests even when the number of post-periods $\bar{T}$ is large.[19]

To describe the conditional testing approach, suppose we wish to test (14) for some fixed $\bar{\theta}$. The conditional testing approach considers tests based on the test statistic

$$\hat{\eta} := \min_{\eta,\tilde{\tau}} \eta \text{ s.t. } \tilde{Y}_n(\bar{\theta}) - \tilde{X}\tilde{\tau} \leqslant \tilde{\sigma}_n \cdot \eta, \tag{15}$$

where $\tilde{\sigma}_n = \sqrt{\text{diag}(\tilde{\Sigma}_n)}$. This linear program selects the value of the nuisance parameters $\tilde{\tau} \in \mathbb{R}^{\bar{T}-1}$

---

[18]Specifically, let Γ be a square matrix with the vector l' in the first row and remaining rows chosen so that Γ has full rank. Define $\tilde{A} := AM_{post}\Gamma^{-1}$. Then $AM_{post}\tau = \tilde{A}\Gamma\tau_{post} = \tilde{A} \begin{pmatrix} \theta \\ \underbrace{\Gamma_{(-1,\cdot)}\tau_{post}}_{:=\tilde{\tau}} \end{pmatrix}$.

[19]Other moment inequality methods have been proposed for subvector inference, but typically do not exploit the linear structure of our setting — see, e.g, Chen, Christensen and Tamer (2018); Bugni, Canay and Shi (2017); Kaido et al. (2019); Chernozhukov, Newey and Santos (2015); Romano and Shaikh (2008). Gafarov (2019), Cho and Russell (2018), and Flynn (2019) also provide methods for subvector inference with linear moment inequalities, but in contrast to our approach require a linear independence constraint qualification (LICQ) assumption for size control.

that produces the most slack in the maximum moment, measured in standard deviation units. Standard duality results from linear programming (e.g. Schrijver (1986), Section 7.4) imply that if the value $\hat{\eta}$ obtained from the so-called primal linear program (15) is finite, then it is equal to the optimal value of the dual program,

$$\hat{\eta} = \max_{\gamma} \gamma' \tilde{Y}_n(\bar{\theta}) \text{ s.t. } \gamma' \tilde{X} = 0, \gamma' \tilde{\sigma}_n = 1, \gamma \geqslant 0. \tag{16}$$

If a vector γ_* is optimal in the dual problem above, then it is a vector of Lagrange multipliers for the primal problem. We will denote by $\hat{V}_n$ the set of optimal vertices of the dual program.[20]

To derive critical values for our test, we analyze the distribution of $\hat{\eta}$ conditional on the event that a vertex γ_* is optimal in the dual problem. Lemma 9 of ARP shows that conditional on the event $\gamma_* \in \hat{V}_n$ and a sufficient statistic S_n for the nuisance parameters, the test statistic $\hat{\eta}$ follows a truncated normal distribution. In particular,

$$\hat{\eta} \mid \{\gamma_* \in \hat{V}_n, S_n = s\} \sim \xi \mid \xi \in [v^{lo}, v^{up}], \tag{17}$$

where $\xi \sim \mathcal{N}\left(\gamma_*' \tilde{\mu}, \gamma_*' \tilde{\Sigma}_n \gamma_*\right)$, $\tilde{\mu} = \mathbb{E}\left[\tilde{Y}_n(\bar{\theta})\right]$, $S_n = (I - \frac{\tilde{\Sigma}_n \gamma_*}{\gamma_*' \tilde{\Sigma}_n \gamma_*} \gamma_*') \tilde{Y}_n(\bar{\theta})$, and v^{lo}, v^{up} are known functions of $\tilde{\Sigma}_n, s, \gamma_*$.[21] One can show that all quantiles of the conditional distribution of $\hat{\eta}$ in the previous display are increasing in $\gamma_*' \tilde{\mu}$,[22] and the null hypothesis (14) implies that $\gamma_*' \tilde{\mu} \leqslant 0$.

We therefore select the critical value for the conditional test to be the $1 - \alpha$ quantile of the truncated normal distribution $\xi | \xi \in [v^{lo}, v^{up}]$ under the worst-case assumption that $\gamma_*' \tilde{\mu} = 0$. To specify this formally, denote by $F_{\xi \mid \xi \in [v^{lo}, v^{up}]}(\cdot; \sigma^2)$ the CDF of $\xi \sim \mathcal{N}\left(0, \sigma^2\right)$ truncated to $[v^{lo}, v^{up}]$. Let $\psi_\alpha^C(\tilde{Y}_n(\bar{\theta}), \tilde{\Sigma}_n)$ denote an indicator for whether the conditional test rejects at the $1 - \alpha$ level. We define the conditional test such that

$$\psi_\alpha^C(\tilde{Y}_n(\bar{\theta}), \tilde{\Sigma}_n) = 1 \iff F_{\xi \mid \xi \in [v^{lo}, v^{up}]}(\hat{\eta}; \gamma_*' \tilde{\Sigma}_n \gamma_*) > 1 - \alpha. \tag{18}$$

It follows immediately from Proposition 6 in ARP that the conditional test controls size,

$$\sup_{\delta \in \Delta, \tau} \sup_{\theta \in \mathcal{S}(\Delta, \delta + \tau)} \mathbb{E}_{(\delta, \tau, \Sigma_n)} \left[\psi_\alpha^C(\tilde{Y}_n(\theta), \tilde{\Sigma}_n)\right] \leqslant \alpha. \tag{19}$$

A confidence set satisfying the uniform coverage criterion (11) can then be constructed by test

---

[20]In general, there may not be a unique solution to the dual program. However, Lemma 11 of ARP shows that conditional on any one vertex of the dual program's feasible set being optimal, every other vertex is optimal with either probability 0 or 1. It thus suffices to condition on the event that a vector $\gamma_* \in \hat{V}$.

[21]The cutoffs v^{lo} and v^{up} are the maximum and minimum of the set $\{x : x = \max_{\gamma \in F_n} \gamma'(s + \frac{\tilde{\Sigma}_n \gamma_*}{\gamma_*' \tilde{\Sigma}_n \gamma_*} x)\}$ when $\gamma_*' \tilde{\Sigma}_n \gamma_* \neq 0$, where F_n is the feasible set of the dual program (16). When $\gamma_*' \tilde{\Sigma}_n \gamma_* = 0$, we define $v^{lo} = -\infty$ and $v^{up} = \infty$, so the conditional test rejects if and only if $\hat{\eta} > 0$.

[22]This follows from the fact that the truncated normal distribution $\xi | \xi \in [v^{lo}, v^{up}]$ has the monotone likelihood ratio property in it is mean (see, e.g. Lemma A.1 in Lee, Sun, Sun and Taylor (2016)).

inversion for the scalar parameter θ. The conditional confidence set is given by

$$\mathcal{C}_{\alpha,n}^{C} := \{\bar{\theta} \,:\, \psi_{\alpha}^{C}(\tilde{Y}_n(\bar{\theta}), \tilde{\Sigma}_n) = 0\}. \tag{20}$$

Remark 1. For each value of $\bar{\theta}$, the test statistic $\hat{\eta}$ can be computed by solving the linear program (15). To form the confidence set $\mathcal{C}_{\alpha,n}^{C}$, one only needs to perform test inversion over a grid of values for the scalar parameter $\bar{\theta}$, and thus the problem remains highly tractable even when $\bar{T}$ is large. Moreover, the commonly-used dual simplex algorithm for linear programming returns a vertex to the dual solution (16), so an optimal dual vertex γ_* can be obtained from standard packages without further calculation. We provide R code for easy implementation of this procedure.

Remark 2. To gain intuition for the conditional test, consider the simple setting in which we have one post-period ($\bar{T} = 1$) and are interested in the treatment effect in the first period, $\theta = \tau_1$. In this case, there are no nuisance parameters, and the form of the conditional test simplifies substantially. The test statistic $\hat{\eta}$ is the maximum of the standardized moments, $\hat{\eta} = \max_j \tilde{Y}_{n,j}/\tilde{\sigma}_{n,j}$, where $\tilde{\sigma}_{n,j}$ is the standard deviation of $\tilde{Y}_{n,j}$. The conditional test rejects in this case if and only if $\frac{\Phi(\hat{\eta})-\Phi(v^{lo})}{1-\Phi(v^{lo})} > 1 - \alpha$. Moreover, if the moments $\tilde{Y}_n$ are uncorrelated with each other, then v^{lo} is the maximum of the non-binding standardized moments, $v^{lo} = \max_{j \neq \hat{j}} \tilde{Y}_{n,j}/\tilde{\sigma}_{n,j}$, where $\hat{j}$ denotes the location of the maximum standardized moment. ∎

5 Asymptotic Power of Conditional Confidence Sets

We now analyze the properties of the conditional confidence sets. For ease of exposition, we consider limits as $n \to \infty$ in the finite-sample normal model defined in Section 3.[23] In Appendix F, we show that these results hold uniformly over a wide class of non-normal data-generating processes.

We present two main results on the asymptotic power of the conditional confidence sets. We first show that the conditional test is pointwise consistent, meaning that any point outside of the identified set is rejected with probability going to one as $n \to \infty$. We next provide a condition under which the power of the conditional test converges to the power envelope in a $n^{-1/2}$-neighborhood of the identified set.

Both our consistency and local asymptotic power results are novel, and exploit additional structure in our problem not present in the somewhat more general setting considered in ARP. ARP consider testing null hypotheses of the form $H_0 : \exists \tau$ s.t. $\mathbb{E}[Y(\theta) - X\tau \mid X] \leq 0$, almost surely. Our setting, in which we are interested in testing the hypothesis (14), is thus a special case of the testing problem considered by ARP in which i) the variable X takes the degenerate distribution $X = \tilde{X}$, and ii) $Y(\theta) = \tilde{Y}(\theta)$ is linear in θ. Both of these features are important for the results obtained in this section. For instance, if i) fails and X is continuously distributed, then the tests proposed by

---

[23]The results in this section as $n \to \infty$ for $\hat{\beta}_n \sim \mathcal{N}\left(\beta, \frac{1}{n}\Sigma^*\right)$ are similar in spirit to the "small-σ" asymptotics in e.g., Kadane (1971); Moreira and Ridder (2019).

ARP will generally not be consistent, as they do not allow for the number of moments to grow with n. Likewise, our local asymptotic power results exploit the geometric structure imposed by features i) and ii).

5.1 Consistency

We first show that the conditional test is consistent, meaning that any fixed point outside of the identified set is rejected with probability approaching one as the sample size $n \to \infty$.

Proposition 5.1. *The conditional test is consistent. That is, for any $\delta_A \in \Delta$, $\tau_A \in \mathbb{R}^{\bar{T}}$, and $\theta^{out} \notin \mathcal{S}(\Delta, \delta_A + \tau_A)$,*

$$\lim_{n \to \infty} \mathbb{P}_{(\delta_A, \tau_A, \Sigma_n)} \left(\theta_{out} \notin \mathcal{C}_{\alpha,n}^C \right) = 1.$$

We present this result for fixed values of (δ_A, τ_A) for ease of exposition. Proposition F.3 in Appendix F shows that the conditional test is uniformly consistent for points at a fixed distance from the identified set over a wide class of data-generating processes.

5.2 Optimal Local Asymptotic Power

We next consider the local asymptotic power of the conditional test. We provide a condition under which the power of the conditional test converges to the power envelope in a $n^{-\frac{1}{2}}$-neighborhood of the identified set. Intuitively, this condition guarantees that the binding and non-binding moments are sufficiently well-separated at points close to the boundary of the identified set.

Assumption 3. *Let $\Delta = \{\delta : A\delta \leqslant d\}$ and fix $\delta_A \in \Delta$. Consider the optimization:*

$$b^{max}(\delta_{A,pre}) = \max_{\delta} l' \delta_{post} \ s.t. \ A\delta \leqslant d, \delta_{pre} = \delta_{A,pre},$$
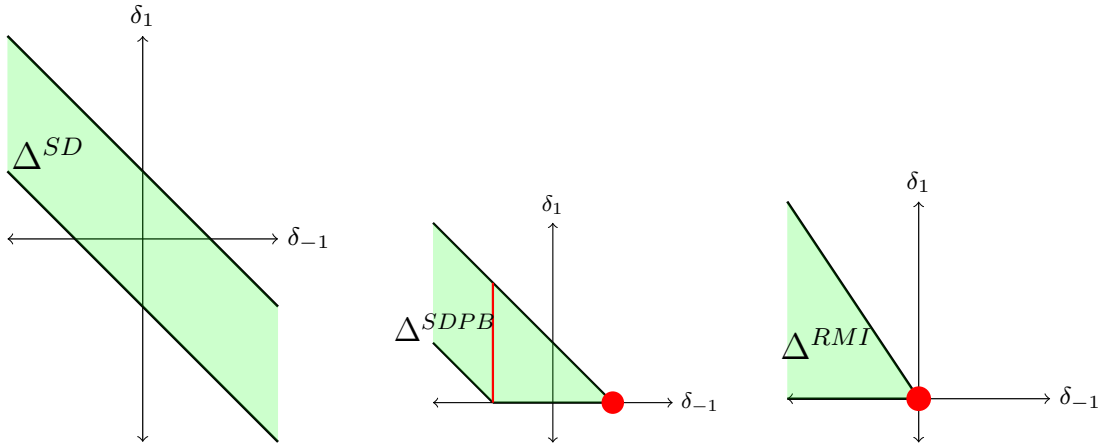
and assume it has a finite solution. For δ^* a maximizer to the above problem, let $B(\delta^*)$ index the set of binding inequality constraints, so that $A_{(B(\delta^*),\cdot)}\delta^* = d_{B(\delta^*)}$ and $A_{(-B(\delta^*),\cdot)}\delta^* - d_{-B(\delta^*)} = -\epsilon_{-B(\delta^*)} < 0$. Assume that there exists a maximizer δ^* to the problem above such that the rank of $A_{(B(\delta^*),post)}$ is equal to $|B(\delta^*)|$. Analogously, assume that there is a finite solution to the analogous problem that replaces $\max$ with $\min$, and that there is a minimizer δ^{**} such that $A_{(B(\delta^{**}),post)}$ has rank $|B(\delta^{**})|$.

Assumption 3 considers the problem of finding the differential trend $\delta \in \Delta$ that is consistent with the pre-trend identified from the data ($\delta_{A,pre}$) and causes $l' \hat{\beta}_{post}$ to be maximally biased for $\theta := l' \tau_{post}$. It requires that the "right" number of moments bind when we do this optimization.

Remark 3. Assumption 3 is closely related to, but slightly weaker than, linear independence constraint qualification (LICQ). LICQ has been used recently in the moment inequality settings of Gafarov (2019); Cho and Russell (2018); Flynn (2019) and Kaido and Santos (2014); see also Kaido et al. (2019), whose results connect constraint qualifications to other common assumptions in the moment inequality literature. We show in Appendix A that LICQ is equivalent to a modified version of Assumption 3 that replaces "there exists a maximizer δ^*" with "for every maximizer δ^*" (and analogously for the minimizer δ^{**}). Thus, LICQ is equivalent to Assumption 3 when the optimizations considered in Assumption 3 have unique solutions, but potentially weaker when there are multiple solutions. We note, however, that while many of the recent papers just mentioned require LICQ for asymptotic size control, we impose Assumption 3 only for our results on local asymptotic power. ∎

Remark 4. In the special case with one pre-period ($\underline{T} = 1$) and one post-period ($\bar{T} = 1$), Assumption 3 has a simple graphical interpretation. It is satisfied whenever Δ has non-empty interior and δ is not vertically aligned with a vertex. Figure 3 shows the areas at which Assumption 3 holds/fails for three of our ongoing examples. We see from the figure that the assumption holds everywhere for Δ^{SD} when $M > 0$, and Lebesgue almost everywhere for Δ^{SDPB} and Δ^{RMI} when $M > 0$ and $\bar{M} > 0$. We note, however, that the sets $\Delta^{SD}(0), \Delta^{SDPB}(0), \Delta^{RMI}(0)$ all have empty interior, and so Assumption 3 fails for these cases (in which θ is point-identified). More generally, one can show that Assumption 3 never holds if θ is point identified. ∎

Figure 3: Diagram of where Assumption 3 holds. The assumption holds (fails) for values of δ plotted in green (red).

Under Assumption 3, the local power of the conditional test converges to the power envelope as $n \to \infty$. Let $\mathcal{I}_\alpha(\Delta, \Sigma_n)$ denote the class of confidence sets that satisfy the finite sample coverage criterion in (11) at the $1-\alpha$ level. Then, the power of the conditional test converges to the optimum over $\mathcal{I}_\alpha(\Delta, \Sigma_n)$ in a $n^{-1/2}$-neighborhood of the identified set.

Proposition 5.2. *Fix $\delta_A \in \Delta, \tau_A$, and suppose Σ^* is positive definite. Let $\theta_A^{ub} = \sup_\theta \mathcal{S}(\Delta, \delta_A + \tau_A)$ be the upper bound of the identified set. Suppose Assumption 3 holds. Then, for any $x > 0$,*

$$\lim_{n\to\infty} \mathbb{P}_{(\delta_A, \tau_A, \Sigma_n)} \left((\theta_A^{ub} + \frac{1}{\sqrt{n}}x) \notin \mathcal{C}_{\alpha,n}^C \right) = \lim_{n\to\infty} \sup_{\mathcal{C}_{\alpha,n} \in \mathcal{I}_\alpha(\Delta, \Sigma_n)} \mathbb{P}_{(\delta_A, \tau_A, \Sigma_n)} \left((\theta_A^{ub} + \frac{1}{\sqrt{n}}x) \notin \mathcal{C}_{\alpha,n} \right)$$

$$= \Phi(c^* x - z_{1-\alpha}),$$

for a positive constant c^*.[24] The analogous result holds replacing $\theta_A^{ub} + \frac{1}{\sqrt{n}}x$ with $\theta_A^{lb} - \frac{1}{\sqrt{n}}x$, for θ_A^{lb} the lower bound of the identified set (although the constant c^* may differ).

Proposition F.4 in Appendix F shows that this result holds uniformly over a large class of distributions that satisfy a uniform version of Assumption 3, meaning that the non-binding moments in the problem b^{max} are uniformly bounded away from binding.
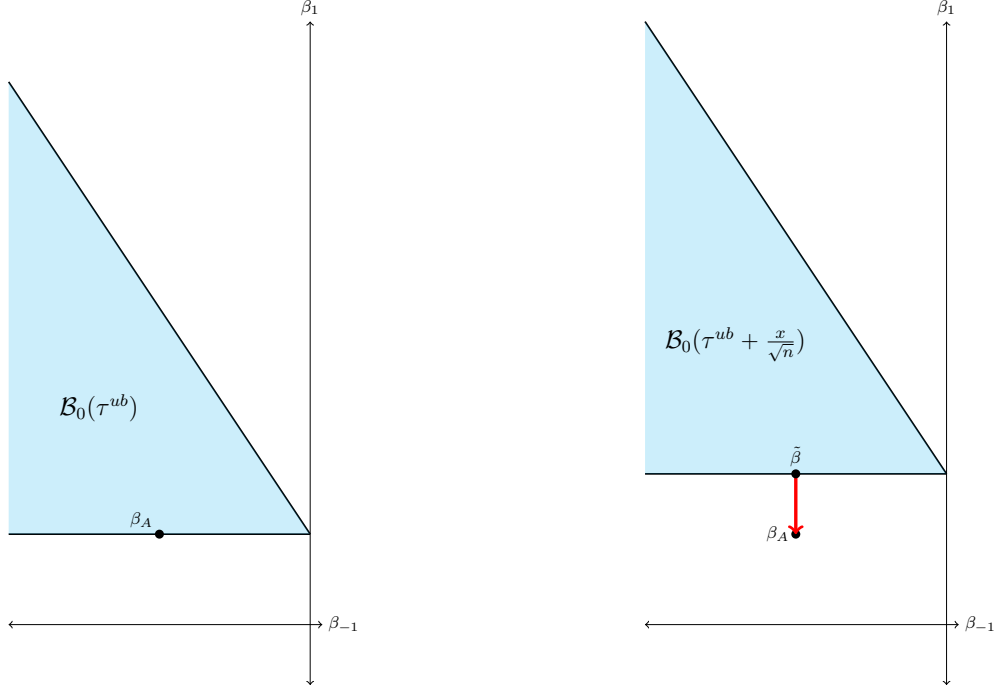
5.2.1 Intuition for local asymptotic optimality

We now sketch the intuition for why the local asymptotic power of the conditional test converges to the power envelope. For simplicity, we focus on the simple case with one pre-period and one post-period. Recall from Remark 4 that in this case, Assumption 3 implies that a single moment determines the upper bound of the identified set. The argument that the local asymptotic power of the conditional test converges to the power envelope then proceeds in two steps. First, we show that the optimal test converges to a t-test of the moment that determines the boundary of the identified set. Next, we show that the power of the conditional test converges to the power of this same t-test.

First, focus on the form of the optimal test. Consider the problem of testing the composite null hypothesis $H_0 : \tau_1 = \bar{\tau}_1, \delta \in \Delta$ against the point alternative $H_A : (\delta, \tau_1) = (\delta_A, \tau_A)$. Since by definition, $\beta = \delta + M_{post}\tau_1$, this is equivalent to testing $H_0 : \beta \in \mathcal{B}_0(\bar{\tau}_1) = \{\beta : \beta = \delta + M_{post}\bar{\tau}_1, \delta \in \Delta\}$ against $H_A : \beta = \delta_A + M_{post}\tau_A =: \beta_A$. It can be shown that $\mathcal{B}_0$ is convex, and so the Neyman-Pearson lemma implies that the most powerful test between H_0 and H_A is a t-test that rejects for large values of $(\beta_A - \tilde{\beta})'\Sigma_n^{-1}\hat{\beta}_n$, where $\tilde{\beta}$ is the closest point in $\mathcal{B}_0$ to β_A in the Mahalanobis distance using Σ_n. Moreover, if $\tilde{\beta}$ is in the interior of a side of $\mathcal{B}_0$ (i.e. not at a vertex), then this t-test is merely a one-sided t-test of the sample moment corresponding with that edge. Figure 4 depicts this testing problem for when $\Delta = \Delta^{RMI}(\bar{M})$. As shown in the figure, Assumption 3 implies that β_A falls in the interior of an edge of $\mathcal{B}_0(\tau^{ub})$. By continuity, for n sufficiently large, the closest point in $\mathcal{B}_0(\tau^{ub} + x/\sqrt{n})$ will also fall in the interior of that edge. Hence, the most powerful test against H_A will be a one-sided t-test of the moment that determines the upper bound of the identified set.

---

[24]In particular, letting $B = B(\delta^{**})$ as defined in Assumption 3, $c^* = -\bar{\gamma}_B' \tilde{A}_{(B,1)}/\sigma_B$, where $\sigma_B = \sqrt{\bar{\gamma}_B' A_{(B,\cdot)} \Sigma^* A'_{(B,\cdot)} \bar{\gamma}_B}$ and $\bar{\gamma}_B$ is a non-zero vector such that $\bar{\gamma}_B' \tilde{A}_{(B,-1)} = 0, \bar{\gamma}_B \geqslant 0$. The vector $\bar{\gamma}_B$ is unique up to scale.

Figure 4: Optimal testing problem for $\Delta = \Delta^{RMI}$

Note: This figure illustrates the optimal testing problem discussed in Section 5.2.1 when $\Delta = \Delta^{RMI}$. By construction, τ^{ub} is the value such that β_A falls on the lower envelope of $\mathcal{B}_0(\tau^{ub})$, as shown in the left panel. Assumption 3 guarantees that β_A is not at a vertex. The right panel depicts the most powerful test between $H_0 : \beta \in \mathcal{B}_0(\tau^{ub} + \frac{x}{\sqrt{n}})$ and $H_A : \beta = \beta_A$, which is a t-test between β_A and $\tilde{\beta}$, the closest point in $\mathcal{B}_0(\tau^{ub} + \frac{x}{\sqrt{n}})$ to β_A in Mahalanobis distance. For $\frac{x}{\sqrt{n}}$ sufficiently small, $\tilde{\beta}$ will also not be at a vertex. The optimal test is thus a t-test of the sample moment corresponding with the lower bound of $\mathcal{B}_0(\tau^{ub} + \frac{x}{\sqrt{n}})$. In this example, the most powerful test is t-test that rejects for large values of $\hat{\beta}_{n,1}$.

Next, consider the conditional test. In the case with one pre-period and post-period, the conditional test rejects if $\frac{\Phi(\hat{\eta}) - \Phi(v^{lo})}{1 - \Phi(v^{lo})} > 1 - \alpha$, where $\hat{\eta}$ is the t-statistic of the largest standardized moment, and conditional on the location of the largest moment (and a sufficient statistic), $\hat{\eta}$ is standard normal truncated to $[v^{lo}, \infty)$. Observe that if $v^{lo} = -\infty$, then the conditional test rejects if $\Phi(\hat{\eta}) > 1 - \alpha$, which corresponds with a t-statistic of the maximal moment. However, Assumption 3 guarantees that for values of $\bar{\tau}_1$ close to τ^{ub}, the mean of the moment that determines the upper bound of the identified set is larger than the mean of the other moments. Thus, since $\Sigma_n \to 0$ as $n \to \infty$, this moment will be maximal with probability going to 1. Intuitively, conditioning on the event that this moment is maximal will therefore provide no information as $n \to \infty$. Since v^{lo} represents the lower truncation from conditioning on the location of the maximal moment, $v^{lo} \to_p -\infty$. Hence, the power of the conditional test converges to the power of a one-sided t-test of the moment that determines the upper bound of the identified set.

Our optimal local asymptotic power results for the conditional test are thus closely related to the fact that as all but one of the moments becomes arbitrarily slack (in standard deviation

22

units), the conditional test converges in probability to a one-sided t-test of the binding moment (see Proposition 3 in ARP). To our knowledge, this feature is not shared by any other existing moment inequality procedure that controls size in the finite sample Gaussian model. Specifically, although relatively insensitive to the inclusion of slack moments, the procedures of Romano, Shaikh and Wolf (2014) and Andrews and Barwick (2012) are still affected by the inclusion of slack moments via the first-stage critical value and size-adjustment factor, respectively. The method of Cox and Shi (2019) is strongly insensitive to the inclusion of slack moments, but does not converge to a one-sided t-test as the other moments become slack.

5.3 Power in Finite Samples

While the conditional confidence sets have desirable asymptotic properties as the sample size $n \to \infty$, our results provide no guarantees on their performance in finite samples. Moreover, we anticipate that the conditional test may have low power in finite samples for cases in which Assumption 3 fails or is "close to failing," meaning that the binding and non-binding moments are not well-separated relative to the sampling variation. An important case where this will occur is when the parameter of interest θ is point-identified.

To develop intuition for why this may occur, we return to the simple case considered in Remark 2 in which there is only one post-period, the parameter of interest is the treatment effect in the first period, and the elements of $\tilde{Y}$ are uncorrelated. Recall that in this example, the conditional test rejects if and only if $\frac{\Phi(\hat{\eta}) - \Phi(v^{lo})}{1 - \Phi(v^{lo})} > 1 - \alpha$, where the test statistic $\hat{\eta}$ is the maximum standardized sample moment and v^{lo} is the value of the second-largest standardized sample moment. Note that if the difference in means between the two largest moments is small (relative to their variance), then we will have that $\hat{\eta} \approx v^{lo}$ with high probability, in which case the conditional test may not reject even if the test statistic $\hat{\eta}$ is large. This suggests that the conditional approach may have low power in finite-samples if there are multiple moments that are close to binding at the edge of the identified set. Recall that when Assumption 3 fails, there are multiple moments that are exactly binding at the identified set bound. We thus should expect the conditional approach to have poor power in finite samples in a neighborhood of points where Assumption 3 fails, where the size of the neighborhood depends on the sampling variation. This is a generic challenge for the conditional testing approach, and we show that it can lead to poor power in some cases in our simulation study in Section 8.

Owing to this limitation, we next turn our attention to fixed length confidence intervals (FL-CIs), which can offer finite-sample improvements in certain special cases of interest, but require stronger assumptions to obtain the asymptotic guarantees of the conditional approach. Looking ahead, we will then show how the conditional confidence sets can be hybridized with FLCIs or least-favorable tests to mitigate poor performance when the moments are not well-separated while retaining desirable asymptotic properties in a wider range of cases.

6 Finite-sample Improvements Using FLCIs

We now consider fixed length confidence intervals (FLCIs) based on affine estimators. Although the FLCIs require additional assumptions to obtain similar asymptotic performance to the conditional confidence sets, they provide finite-sample guarantees for certain special cases of interest.

6.1 Constructing FLCIs

Following Donoho (1994) and Armstrong and Kolesar (2018, 2019), we consider fixed length confidence intervals based on an affine estimator for θ,

$$\mathcal{C}_{\alpha,n}(a, v, \chi) = \left(a + v'\hat{\beta}_n\right) \pm \chi, \tag{21}$$

where a and χ are scalars and $v \in \mathbb{R}^{T+\bar{T}}$. We wish to minimize the confidence interval half-length χ subject to the constraint that $\mathcal{C}_{\alpha,n}(a, v, \chi)$ satisfies the coverage requirement (11). To do so, note that $a + v'\hat{\beta}_n \sim \mathcal{N}\left(a + v'\beta, \, v'\Sigma_n v\right)$, and hence $|a + v'\hat{\beta}_n - \theta| \sim |\mathcal{N}\left(b, \, v'\Sigma_n v\right)|$, where $b = a + v'\beta - \theta$ is the bias of the affine estimator $a + v'\hat{\beta}$ for θ. Observe further that $\theta \in \mathcal{C}_n(a, v, \chi)$ if and only if $|a + v'\hat{\beta}_n - \theta| \leqslant \chi$. For fixed values a and v, the smallest value of χ that satisfies (11) is therefore the $1 - \alpha$ quantile of the $|\mathcal{N}\left(\bar{b}, \, v'\Sigma_n v\right)|$ distribution, where $\bar{b}$ is the worst-case bias of the affine estimator,

$$\bar{b}(a, v) := \sup_{\delta \in \Delta, \tau_{post} \in \mathbb{R}^{\bar{T}}} |a + v'(\delta + M_{post}\tau_{post}) - l'\tau_{post}|. \tag{22}$$

Let $cv_\alpha(t)$ denote the $1 - \alpha$ quantile of the folded normal distribution $|\mathcal{N}\left(t, \, 1\right)|$.[25] Then, for fixed a and v, the smallest value of χ that satisfies the coverage requirement (11) is

$$\chi_n(a, v; \alpha) = \sigma_{v,n} \cdot cv_\alpha(\bar{b}(a, v)/\sigma_{v,n}), \tag{23}$$

where $\sigma_{v,n} := \sqrt{v'\Sigma_n v}$.

We can therefore construct the minimum length FLCI by choosing the values of a and v to minimize (23). Intuitively, this minimization optimally trades off bias and variance, since the half-length $\chi_n(a, v; \alpha)$ is increasing in both the worst-case bias $\bar{b}$ and the variance $\sigma_{v,n}^2$ (assuming $\alpha \in (0, 0.5]$). When Δ is convex, this minimization can be solved as a nested optimization problem, where both the inner and outer minimizations are convex (Low, 1995; Armstrong and Kolesar, 2018, 2019), and is thus simple to compute. We denote by $\mathcal{C}_{\alpha,n}^{FLCI}$ the $1 - \alpha$ level FLCI with the shortest length,

$$\mathcal{C}_{\alpha,n}^{FLCI} = \left(a_n + v_n'\hat{\beta}_n\right) \pm \chi_n,$$

---

[25] If $t = \infty$, we define $cv_\alpha = \infty$.

where $\chi_n := \inf_{a,v} \chi_n(a, v; \alpha)$ and a_n, v_n are the optimal values in the minimization.

Example: $\Delta^{SD}(M)$ Suppose $\underline{T} \geqslant 1$ and the parameter of interest is the treatment effect in the first post-period, $\theta = \tau_1$. For $\Delta^{SD}(M)$, the affine estimator used by the optimal FLCI takes the form

$$a + v'\hat{\beta}_n = \hat{\beta}_{n,1} - \sum_{s=-\underline{T}+1}^{0} w_s \left(\hat{\beta}_{n,s} - \hat{\beta}_{n,s-1} \right), \tag{24}$$

where the weights w_s are potentially negative and sum to one. Thus, the affine estimator used by the optimal FLCI takes the post-period event-study coefficient for period 1 and subtracts out a weighted sum of the estimated slopes between consecutive pre-periods. Intuitively, Δ^{SD} restricts the changes in the slope of the underlying trend across periods, not the slope itself. In order to have finite worst-case bias, an affine estimator must therefore subtract out an estimate of the slope of the trend between $t = 0$ and $t = 1$ using the observed slopes in the pre-period. The worst-case bias will be smaller if more weight is placed on pre-periods closer to the treatment date, since the trend has less time to change slope, but it may reduce variance to place more weight on earlier pre-periods. The weights w_s will then be chosen to optimally trade off these sources of bias and variance.

Example: $\Delta^{SDPB}(M)$ Suppose again that $\underline{T} \geqslant 1$ with $\theta = \tau_1$. In the case $\Delta = \Delta^{SDPB}(M)$, the optimal FLCI is unchanged from the form given in (24). This is the case because if $\delta \in \Delta^{SD}(M)$, then the vector that adds a vector of constant slope to δ, $\tilde{\delta} = \delta + c \cdot (-\underline{T}, ..., \bar{T})'$, lies in $\Delta^{SD}(M)$ as well. As discussed above, in order to have finite worst-case bias over $\Delta^{SD}(M)$, an affine estimator must take $\hat{\beta}_{n,1}$ and subtract out a weighted average of the pre-period slopes, and by analogous argument the same holds for $\Delta^{SDPB}(M)$. The bias of such an estimator will thus be the same for δ as it is for $\tilde{\delta}$. Hence, for any $\delta \in \Delta^{SD}(M)$, there exists $\tilde{\delta} \in \Delta^{SDPB}(M)$ that produces the same bias by choosing the constant c to be sufficiently large. The FLCI is thus unable to take advantage of the additional sign restrictions.

Example: $\Delta^{RMI}(\bar{M})$ Suppose that $\underline{T} \geqslant 1$ with $\theta = \tau_1$. If $\Delta = \Delta^{RMI}(\bar{M})$ and $\bar{M} > 0$, then all affine estimators for τ_1 have infinite worst-case bias, and the FLCI is thus the entire real line. To see why, note that since τ_{post} is unrestricted in (22), any affine estimator with finite bias must set $v_{post} = e_1$. Otherwise, the bias of the affine estimator could be made arbitrarily large by choosing the magnitude of τ_{post} to be large. But, if $v_{post} = e_1$, then the bias of the affine estimator $a + v'\hat{\beta}_n$ equals $|a + v'_{pre}\delta_{pre} + \delta_1|$ regardless of the value of τ. By the triangle inequality, $\max\{|a + v'_{pre}\delta_{pre} + \tilde{\delta}_1|, |a + v'_{pre}\delta_{pre} + \bar{\delta}_1|\} \geqslant \frac{1}{2}|\tilde{\delta}_1 - \bar{\delta}_1|$ for any two values $\tilde{\delta}_1$ and $\bar{\delta}_1$. However, as seen in Figure 2 for the two-period case, the range of values of δ_1 that are feasible given a pre-period trend δ_{pre} can be made arbitrarily large by setting δ_{pre} such that $|\delta_{-1}|$ is large. Hence $\bar{b}(a, v) = \infty$ for all choices of (a, v).
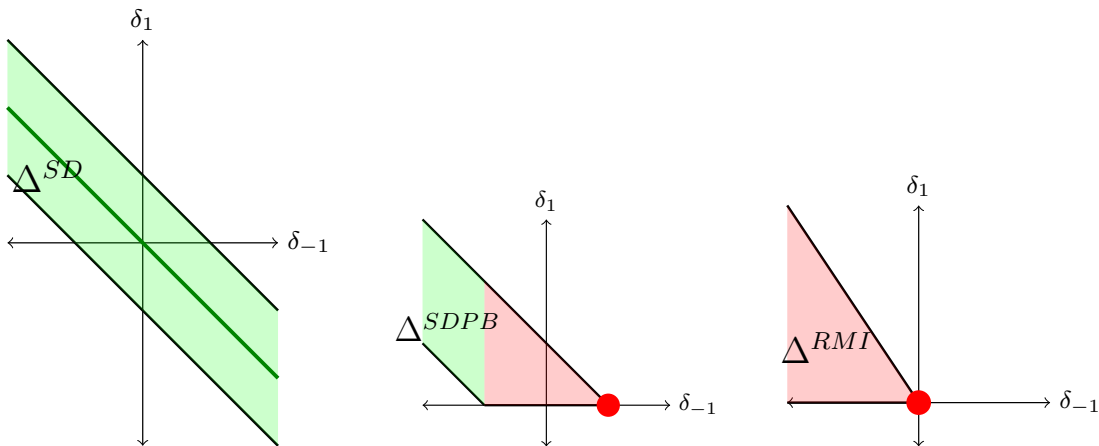
6.2 Consistency and Local Asymptotic Power

We now consider the performance of the optimal FLCIs as $n \to \infty$, as we did in Section 5 for the conditional confidence sets. As made clear by the example for Δ^{RMI} above, in which the FLCI always has infinite length, the FLCIs will not necessarily be consistent (nor optimal) without additional restrictions. We show that the FLCIs obtain similar asymptotic properties to the conditional confidence set if and only if the length of the identified set at δ equals its maximal length over all possible violations in the class Δ.

First, recall from the discussion in Section 3.2 that when Δ is convex, the identified set is an interval. Further, observe from (9) and (10) that the length of the identified set is $\theta^{ub} - \theta^{lb} = b^{max}(\beta_{pre}; \Delta) - b^{min}(\beta_{pre}; \Delta)$, which depends only on Δ and β_{pre}. We will therefore denote by $LID(\Delta; \beta_{pre})$ the length of the identified set. As we will show, the optimal FLCI shares the desirable asymptotic behavior of the ARP confidence sets when δ_{pre} is such that the length of the identified set is maximal and finite.

Assumption 4 (Identified set maximal length and finite). *Suppose δ_{pre} is such that $LID(\Delta, \delta_{pre}) = \sup_{\tilde{\delta}_{pre} \in \Delta_{pre}} LID(\Delta, \tilde{\delta}_{pre}) < \infty$.*

Remark 5. In our two-period examples, the length of the identified set corresponds with the height of Δ, and so Assumption 4 holds if and only if Δ achieves its maximal height at δ_{-1}. Figure 5 shows where this is the case for three of our ongoing examples. As can be seen, the assumption holds everywhere for Δ^{SD}, for values of δ where the sign restrictions do not bind for Δ^{SDPB}, and nowhere for Δ^{RMI}. The restrictiveness of Assumption 4 thus depends greatly on Δ. ∎

Figure 5: Diagram of where Assumptions 4 and 5 hold. The values of δ are colored red (neither holds), light green (Assumption 4 only), and dark green (both hold).

Our next result shows $\mathcal{C}_{\alpha,n}^{FLCI}$ is consistent if and only if Assumption 4 holds, provided that the identified set is not the equal to the real line (in which case any procedure that controls size is consistent).

Proposition 6.1. *Suppose that Δ is convex and $\alpha \in (0, .5]$. Fix $\delta_A \in \Delta$ and $\tau_A \in \mathbb{R}^{\bar{T}}$, and suppose $\mathcal{S}(\Delta, \delta_A + \tau_A) \neq \mathbb{R}$. Then Assumption 4 holds if and only if $\mathcal{C}_{\alpha,n}^{FLCI}$ is consistent, meaning for all $\theta_{out} \notin \mathcal{S}_{\theta}(\Delta, \delta_A + \tau_A)$,*

$$\lim_{n\to\infty} \mathbb{P}_{(\delta_A, \tau_A, \Sigma_n)} \left(\theta_{out} \in \mathcal{C}_{\alpha,n}^{FLCI} \right) = 0.$$

Moreover, if Assumption 4 holds in addition to the conditions in Proposition 5.2, then the FLCI has local asymptotic power approaching the power envelope.

Proposition 6.2. *Fix $\delta_A \in \Delta$, $\tau_A \in \mathbb{R}^{\bar{T}}$ and suppose Σ^* is positive definite. Let $\theta_A^{ub} = \sup_\theta \mathcal{S}(\Delta, \delta_A + \tau_A)$ be the upper bound of the identified set. Suppose that Assumption 3 holds and $\delta_{A,pre}$ satisfies Assumption 4. Then, for any $x > 0$ and $\alpha \in (0, 0.5]$,*

$$\lim_{n\to\infty} \mathbb{P}_{(\delta_A, \tau_A, \Sigma_n)} \left((\theta_A^{ub} + \frac{1}{\sqrt{n}}x) \notin \mathcal{C}_{\alpha,n}^{FLCI} \right) = \lim_{n\to\infty} \sup_{\mathcal{C}_{\alpha,n} \in \mathcal{I}_\alpha(\Delta, \Sigma_n)} \mathbb{P}_{(\delta_A, \tau_A, \Sigma_n)} \left((\theta_A^{ub} + \frac{1}{\sqrt{n}}x) \notin \mathcal{C}_{\alpha,n} \right).$$

The analogous result holds replacing $\theta_A^{ub} + \frac{1}{\sqrt{n}}x$ with $\theta_A^{lb} - \frac{1}{\sqrt{n}}x$, for θ_A^{lb} the lower bound of the identified set.

Thus, we see that $\mathcal{C}_{\alpha,n}^{FLCI}$ behaves similarly to $\mathcal{C}_{\alpha,n}^C$ as $n \to \infty$ when Assumption 4 holds, but it is otherwise inconsistent in the strong sense that points outside of the identified set are rejected with non-vanishing probability asymptotically.

6.3   Finite-sample near-optimality

While the FLCIs require stronger conditions to produce the same behavior asymptotically as the conditional confidence sets, an advantage of the FLCIs is that Armstrong and Kolesar (2018, 2019) establish finite-sample near-optimality results for particular cases of interest. The following result, which is an immediate consequence of results in Armstrong and Kolesar (2018, 2019), bounds the ratio of the expected length of the shortest possible confidence interval that controls size relative to the length of the optimal FLCI.

Assumption 5. *Assume that i) Δ is convex and centrosymmetric (i.e. $\delta \in \Delta$ implies $-\delta \in \Delta$), and ii) $\delta_A \in \Delta$ is such that $(\delta - \delta_A) \in \Delta$ for all $\delta \in \Delta$.*

Proposition 6.3. *Suppose δ_A and Δ satisfy Assumption 5. Then, for any $\tau_A \in \mathbb{R}^{\bar{T}}$, Σ^* positive definite, and $n > 0$,*

$$\frac{\inf_{\mathcal{C}_{\alpha,n} \in \mathcal{I}_\alpha(\Delta, \Sigma_n)} \mathbb{E}_{(\delta_A, \tau_A, \Sigma_n)} \left[\lambda(\mathcal{C}_{\alpha,n}) \right]}{2\chi_n} \geqslant (z_{1-\alpha}(1-\alpha) - \tilde{z}_\alpha \Phi(\tilde{z}_\alpha) + \phi(z_{1-\alpha}) - \phi(\tilde{z}_\alpha))/z_{1-\alpha/2},$$

where $\lambda(\cdot)$ denotes the length (Lebesgue measure) of a set and $\tilde{z}_\alpha = z_{1-\alpha} - z_{1-\alpha/2}$.

Part i) of Assumption 5 is satisfied for Δ^{SD} but not for our other ongoing examples. For example, Δ^{SDPB} and Δ^{RMI} are convex but not centrosymmetric, and Δ^{RM} is neither convex nor centrosymmetric. Part ii) of Assumption 5 is always satisfied when parallel trends holds ($\delta_A = 0$). It also holds whenever δ_A is a linear trend for the case of $\Delta^{SD}(M)$. We show in Lemma E.29 in the Appendix that the conditions of Proposition 6.3 imply that Assumption 4 holds. Thus, the FLCIs obtain finite sample near-optimality in a subset of the cases where they are consistent.

We note that Assumption 5 does not necessarily rule out point identification. For instance, it is satisfied when $\Delta = \Delta^{SD}(0)$, which restricts that the differential trend be linear. Recall that Assumption 3 fails when θ is point-identified, and so the conditional approach may have low power in such cases. We thus anticipate that FLCIs will outperform the conditional confidence sets in point-identified cases where Assumption 5 holds.

Finally, for $\alpha = 0.05$, the lower bound in Proposition 6.3 evaluates to 0.72, so the expected length of the shortest confidence possible interval that satisfies the coverage requirement (11) is at most 28% shorter than the length of the optimal FLCI.

6.4 Comparison to conditional confidence sets

The results above and in the previous section highlight tradeoffs between the conditional confidence sets and FLCIs. The conditional confidence sets are consistent and have optimal local asymptotic power under weaker conditions than the FLCIs, but may have low power in finite samples if the moments are not well-separated. By contrast, when the stronger conditions for consistency are satisfied, the FLCIs additionally have desirable finite-sample guarantees when Δ is convex and centrosymmetric. The balance of these tradeoffs depends on the choice of Δ, as can be seen from the examples discussed in Section 6.2. At one extreme, there are special cases — such as Δ^{SD} — in which Assumption 4 is non-restrictive and the conditions for finite sample-near optimality hold. The FLCIs are preferable in these special cases, as they have similar asymptotic properties to the conditional approach but also offer finite-sample guarantees. At the other extreme, there are cases — such as Δ^{RMI} — where the FLCIs are never consistent and thus are undesirable. Matters are somewhat less clear for intermediate cases between these two extremes. For Δ^{SDPB}, the FLCIs are consistent for some values of δ but not for others. One can also construct sets Δ for which the conditions for consistency and finite-sample near-optimality of the FLCIs hold for some values of $\delta \in \Delta$ but for other values of δ the FLCIs are inconsistent.[26]

To address these intermediate cases, we now turn our attention to hybridized conditional confidence sets, which mitigate the power issues of the conditional confidence sets when the moments are not well-separated while largely retaining their desirable asymptotic properties.

---

[26]For instance, consider the two-period case where $\Delta = \Delta^{SD}(M) \cap \{\delta : |\delta_1| \leqslant M\}$. Then Proposition 6.3 holds for $\delta = 0$, but Assumption 4 fails for any $\delta \in \Delta$ with $\delta_{-1} \neq 0$.

7   Conditional-FLCI Hybrid Confidence Sets

We showed in Section 5 that the conditional approach has generally desirable asymptotic properties, but may perform poorly in finite samples if the moments are not well-separated. In Section 6, we showed that FLCIs are an attractive alternative with finite sample guarantees for certain special cases of interest, but may be inconsistent outside of these special cases. It is therefore natural to ask whether it is possible to mitigate the poor performance of the conditional approach when the moments are not well-separated while retaining its desirable asymptotic properties.

To address this goal, we now propose a novel confidence set that hybridizes the conditional test with the optimal FLCI. This conditional-FLCI hybrid confidence set achieves similar desirable asymptotic properties as the conditional confidence set. Moreover, we find in simulations (discussed in the following section) that it leads to substantial improvements in the power of the conditional test in a variety of cases where the moments are not well-separated.

In the Appendix, we also consider a hybridization between conditional confidence sets and least-favorable critical values, which may be useful in cases, such as Δ^{RMI}, where we know that the FLCI will be uninformative. Because we only find moderate improvements of this conditional least-favorable confidence set relative to the conditional confidence set in simulations, we defer the details to Appendix C.

7.1   Constructing Conditional-FLCI Hybrid Confidence Sets

The conditional-FLCI hybrid confidence set is constructed by first testing whether a candidate parameter value lies within the optimal level-$(1 - \kappa)$ FLCI, and then applying a conditional test to all parameter values that lie within the optimal FLCI. In the second stage, we use a modified version of the conditional test that i) adjusts size to account for the first-stage test, and ii) conditions on the event that the first-stage test fails to reject.[27]

Formally, suppose that $0 < \kappa < \alpha$. Consider the level $(1 - \kappa)$ optimal FLCI, $\mathcal{C}_{\kappa,n}^{FLCI} = a_n + v_n' \hat{\beta}_n \pm \chi_n$. Lemma B.2 shows that the distribution of the test statistic $\hat{\eta}$ defined in (15) follows a truncated normal distribution conditional on the parameter value $\bar{\theta}$ falling within the level $(1 - \kappa)$ optimal FLCI. More concretely,

$$\hat{\eta} \,\big|\, \left\{ \gamma_* \in \hat{V}_n, S_n = s, \bar{\theta} \in \mathcal{C}_{\kappa,n}^{FLCI} \right\} \sim \xi \,\big|\, \xi \in [v_{C\text{-}FLCI}^{lo}, v_{C\text{-}FLCI}^{up}], \tag{25}$$

where $\xi \sim \mathcal{N}\left(\gamma_*' \tilde{\mu}, \, \gamma_*' \tilde{\Sigma}_n \gamma_* \right)$, $v_{C\text{-}FLCI}^{lo}, v_{C\text{-}FLCI}^{up}$ are defined in Lemma B.2, γ_* is a vector of Lagrange multipliers for the primal problem (16), and S_n is the sufficient statistic for the nuisance parameters as defined in Section 4. With this result, the construction of the second-stage of the conditional-FLCI hybrid test is analogous to the construction of the conditional test, except it uses the modified

---

[27]In simulations, we find that adjusting the conditioning event in the second stage yields better performance in terms of excess length than simply Bonferroni-adjusting the second-stage conditional test.

size $\tilde{\alpha} = \frac{\alpha - \kappa}{1 - \kappa}$, which accounts for the first-stage test. Formally, the conditional-FLCI hybrid test $\psi_{C\text{-}FLCI}$ is defined as

$$\psi_{\kappa,\alpha}^{C\text{-}FLCI}(\hat{\beta}_n, \bar{\theta}, \tilde{\Sigma}_n) = 1 \iff$$
$$\theta \notin \mathcal{C}_{\kappa,n}^{FLCI}, \text{ OR } F_{\xi \mid \xi \in [v_{C\text{-}FLCI}^{lo}, v_{C\text{-}FLCI}^{up}]}(\hat{\eta}; \gamma'_* \tilde{\Sigma}_n \gamma_*) > 1 - \tilde{\alpha},$$

where, as in Section 4, $F_{\xi \mid \xi \in [v_{C\text{-}FLCI}^{lo}, v_{C\text{-}FLCI}^{up}]}(\cdot; \gamma'_* \tilde{\Sigma}_n \gamma_*)$ denotes the CDF of a $\xi \sim \mathcal{N}\left(0, \gamma'_* \tilde{\Sigma}_n \gamma_*\right)$ truncated to $[v_{C\text{-}FLCI}^{lo}, v_{C\text{-}FLCI}^{up}]$.

Since the FLCI controls size, the first stage test rejects with probability at most κ under the null that $\theta = \bar{\theta}$. Likewise, by the same argument as for the conditional test, the second-stage test rejects with probability at most $\tilde{\alpha} = \frac{\alpha - \kappa}{1 - \kappa}$ conditional on $\theta \in \mathcal{C}_{\kappa,n}^{FLCI}$. Together these results imply that the conditional-FLCI hybrid test controls size,

$$\sup_{\delta \in \Delta, \tau} \sup_{\theta \in \mathcal{S}(\Delta, \delta + \tau)} \mathbb{E}_{(\delta, \tau, \Sigma_n)} \left[\psi_{\kappa,\alpha}^{C\text{-}FLCI}(\hat{\beta}_n, \bar{\theta}, \tilde{\Sigma}_n) \right] \leqslant \alpha. \tag{26}$$

Therefore, we can construct a conditional-FLCI hybrid confidence set for the parameter θ that satisfies (11) by inverting the conditional-FLCI test,

$$\mathcal{C}_{\kappa,\alpha,n}^{C\text{-}FLCI} = \{\theta : \psi_{\kappa,\alpha}^{C\text{-}FLCI}(\hat{\beta}_n, \bar{\theta}, \tilde{\Sigma}_n) = 0\}. \tag{27}$$

7.2 Asymptotic power of the conditional-FLCI hybrid confidence sets

We now show that the conditional-FLCI hybrid confidence sets have asymptotic properties similar to those of the conditional test. First, the conditional-FLCI hybrid is consistent.

Proposition 7.1. *The conditional-FLCI hybrid test is consistent. That is, for any $\delta_A \in \Delta, \tau_A \in \mathbb{R}^{\bar{T}}$, $\theta_{out} \notin \mathcal{S}(\Delta, \delta_A + \tau_A)$, $\alpha \in (0, .5]$, and $\kappa \in (0, \alpha)$,*

$$\lim_{n \to \infty} \mathbb{P}_{(\delta_A, \tau_A, \Sigma_n)} \left(\theta_{out} \notin \mathcal{C}_{\kappa,\alpha,n}^{C\text{-}FLCI} \right) = 1.$$

Second, when Assumption 3 holds, the local asymptotic power of the conditional-FLCI hybrid is at least as good as the power of the optimal level $(1 - \tilde{\alpha})$ confidence set. Thus, when the first-stage κ is small, the power of the conditional-FLCI hybrid is near the power envelope.

Proposition 7.2. *Fix $\delta_A \in \Delta, \tau_A$, and Σ^* positive definite. Suppose Assumption 3 holds. Suppose $\alpha \in (0, .5]$, $\kappa \in (0, \alpha)$, and let $\tilde{\alpha} = \frac{\alpha - \kappa}{1 - \kappa}$. Then,*

$$\liminf_{n \to \infty} \mathbb{P}_{(\delta_A, \tau_A, \Sigma_n)} \left((\theta_A^{ub} + \frac{1}{\sqrt{n}} x) \notin \mathcal{C}_{\kappa,\alpha,n}^{C\text{-}FLCI} \right) \geqslant \lim_{n \to \infty} \sup_{\mathcal{C}_{\alpha,n} \in \mathcal{I}_{\tilde{\alpha}}(\Delta, \Sigma_n)} \mathbb{P}_{(\delta_A, \tau_A, \Sigma_n)} \left((\theta_A^{ub} + \frac{1}{\sqrt{n}} x) \notin \mathcal{C}_{\alpha,n} \right).$$

The analogous result holds replacing $\theta_A^{ub} + \frac{1}{\sqrt{n}}x$ with $\theta_A^{lb} - \frac{1}{\sqrt{n}}x$, for θ_A^{lb} the lower bound of the identified set (although the constant c^* may differ).

8 Simulation study

In this section, we present a series of simulations calibrated to 12 recent papers illustrating the relative performance of the confidence sets discussed above and comparing them to an optimal benchmark.

8.1 Simulation Design

Our simulations are calibrated using the estimated covariance matrix from the 12 papers surveyed in Roth (2019). Roth (2019) systematically reviewed the set of papers containing an event-study plot published in the *American Economic Review, AEJ: Applied Economics* and *AEJ: Economic Policy* between 2014 and mid-2018, and analyzed 12 (out of a total of 70) for which replication data was available and other criteria (such as reporting standard errors) were met.[28] For any given paper in the survey, we will denote by $\hat{\Sigma}$ the estimated variance-covariance matrix from the event-study in the paper, calculated using the clustering scheme specified by the authors. We then simulate event-study coefficients $\hat{\beta}_s$ from a normal model under the assumption of parallel trends ($\delta = 0$) and zero treatment effects ($\tau_{post} = 0$), $\hat{\beta}_s \sim \mathcal{N}\left(0, \hat{\Sigma}\right)$. The choice of zero treatment effect is only a normalization, since all of the methods considered are equivariant to shifts in τ_{post}. In simulation s, we then construct confidence intervals using the pair $(\hat{\beta}_s, \hat{\Sigma})$ for each of our proposed procedures. In all of our simulations, we use nominal 95% confidence intervals.[29]

We conduct simulations in this finite-sample normal model with known covariance for both practical and substantive reasons. We do this because we expect the normal approximation to be a reasonable approximation to finite-sample behavior in this setting, and because there are tractable benchmarks for the optimal performance of confidence sets in this context. To examine the quality of the normal approximation, in Appendix H.2 we report results from a set of simulations calibrated to the empirical distribution of the first paper in our survey, in which the covariance matrix is estimated from the data. There, we find that the finite-sample performance of the procedures we study is indeed well-approximated by that in the normal model.

For a given choice of Δ, we compute the identified set $\mathcal{S}(\Delta, 0)$ and calculate the expected excess length for each of the proposed confidence sets. The excess length of a confidence set $\mathcal{C}(\hat{\beta})$ is the length of the part of the confidence set that falls outside of the identified set, defined as $EL(\mathcal{C}; \hat{\beta}) = \lambda(\mathcal{C}(\hat{\beta}) \backslash \mathcal{S}(\Delta, 0))$. We benchmark the expected excess length of our proposed procedures relative to the optimal bound over confidence sets that satisfy the uniform coverage requirement (11).

---

[28]See Section 4.1 of Roth (2019) for a detailed discussion of the sample selection criteria for this survey of published event studies.

[29]Following Romano et al. (2014) and ARP, we use $\kappa = \alpha/10$ for our hybrid procedures.

A formula for this optimal bound is provided in Appendix D, and follows from results in Armstrong and Kolesar (2018), which use the observation that the confidence set that minimizes expected (excess) length inverts most powerful (Neyman-Pearson) tests for each candidate parameter value. For each paper, we conduct 1000 simulations and compute the optimal bound and the average excess length for the conditional confidence set, FLCI, and conditional-FLCI hybrid confidence set. We then define the excess length efficiency for each procedure to be the ratio of the optimal bound to the simulated expected excess length. In the main text, we report the efficiency ratios for the median paper in the survey. Figures in Appendix H show the maximum and minimum over the papers in the survey.

We consider three choices of Δ to highlight the performance of the confidence sets across a range of conditions. Table 1 summarizes which of our theoretical results hold for each of the simulation designs. We first consider sets of the form $\Delta = \Delta^{SD}(M)$. In this simulation design, the conditions hold for all three procedures to be consistent, and they all have (near-)optimal local asymptotic power provided $M > 0$. The conditions also hold for the FLCIs to have near-optimal expected length in finite sample. Next, we consider sets of the form $\Delta = \Delta^{SDPB}(M)$. Under this simulation design, the conditional (hybrid) confidence sets are again consistent, and are asymptotically (near-) optimal when $M > 0$. However, the conditions needed for the FLCI to be consistent do not hold when $M > 0$. Finally, we consider sets of the form $\Delta = \Delta^{SDI}(M)$, which combines Δ^{SD} with the restriction that δ be increasing. In this final specification, only the conditions for the consistency of the conditional and hybrid confidence sets hold. None of the criteria needed for asymptotic or finite-sample optimality are satisfied, and the FLCIs are inconsistent when $M > 0$.

	Δ^{SD}	Δ^{SDPB}	Δ^{SDI}
Conditional / Hybrid			
Consistent	✓	✓	✓
Asymptotically (near-)optimal	✓	✓	✗
FLCI			
Consistent	✓	✗	✗
Asymptotically (near-)optimal	✓	✗	✗
Finite-sample near-optimal	✓	✗	✗

Table 1: Summary of expected properties for each simulation design

Finally, for each choice of Δ, the units of the parameter M are standardized to equal the standard error of the first post-period event study coefficient (σ_1). We show results for a variety of choices of M/σ_1. All of the procedures considered and the optimal bounds are invariant up to scale in the sense that the confidence set using $(M, \frac{1}{n}\Sigma^*)$ is $\frac{1}{n}$ times the confidence set using (nM, Σ^*). Thus, our simulation results on excess length as M/σ_1 grows large are isomorphic to our asymptotic results presented earlier in which $n \to \infty$ for $\Sigma_n = \frac{1}{n}\Sigma^*$ and $M > 0$ fixed. They also, however, have a finite-sample interpretation, illustrating how our results change as we allow the set of underlying trends to be more non-linear, holding Σ^* constant. In the main text, the parameter of interest in

all simulations is the causal effect in the first post-period ($\theta = \tau_1$).

In Appendix H, we present several extensions to the main simulation exercise discussed here. First, we repeat the simulation study, setting the parameter of interest to be the average causal effect across all post-periods ($\theta = \bar{\tau}_{post}$), with qualitatively similar results. Second, we present results for the conditional-least favorable hybrid discussed briefly in Section 7 and detailed in Appendix C. Finally, we present simulation results for $\Delta = \Delta^{RMI}(\bar{M})$. This choice of Δ illustrates the performance of our proposed confidence sets in a case in which the FLCI is always infinite length, but the conditional and hybrid approaches remain consistent.
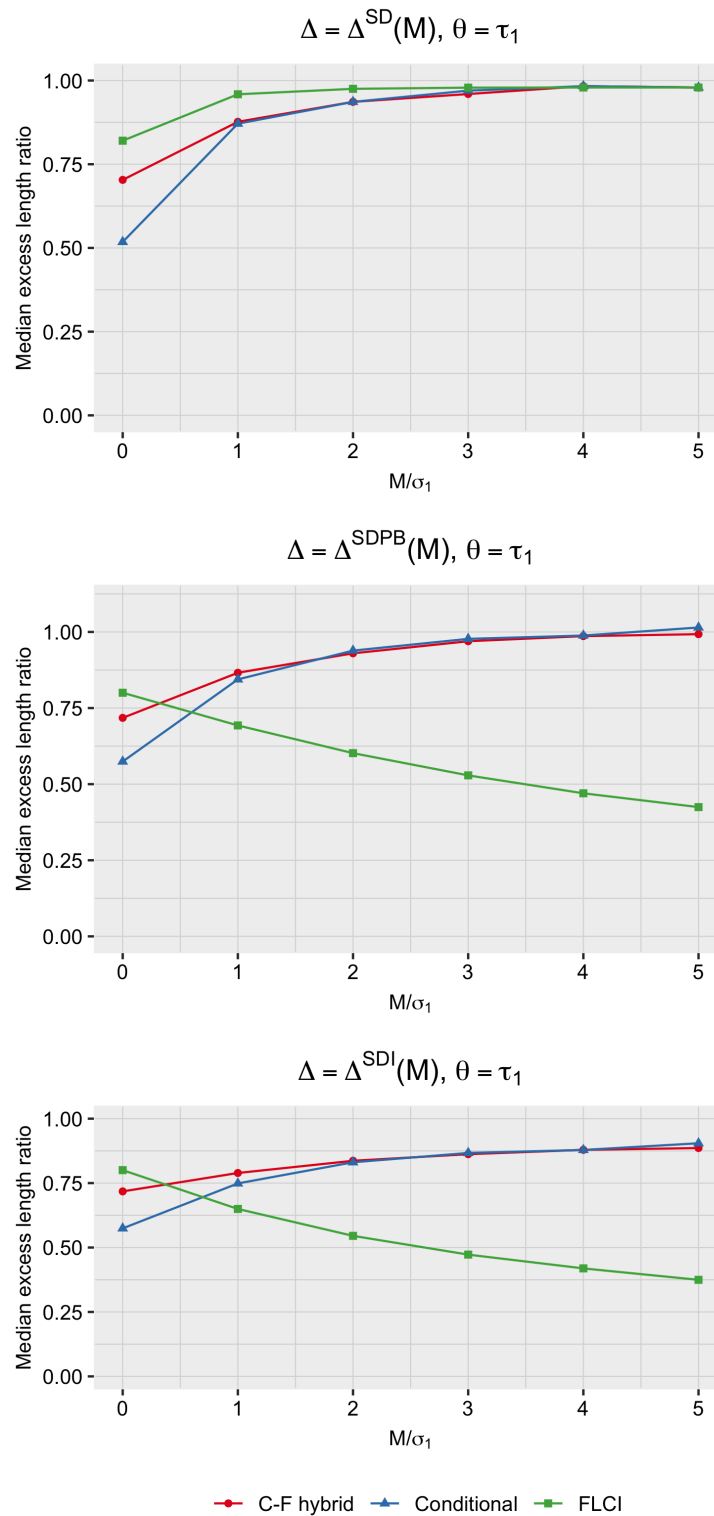
8.2 Simulation Results

Results for $\Delta^{SD}(M)$: The top panel of Figure 6 plots the efficiency ratio for each of our procedures as a function of M/σ_1 when $\Delta = \Delta^{SD}$. As expected, all of the procedures perform well as M/σ_1 grows large, with the efficiency ratios approaching 1. This illustrates our asymptotic (near-) optimality results for all the procedures considered in this simulation design. The FLCIs also perform quite well for smaller values of M/σ_1, including the point-identified case where $M = 0$, illustrating the finite-sample near-optimality results for the FLCIs when Assumption 5 holds. Although the conditional confidence sets have efficiency approaching the optimal bound for M/σ_1 large, their efficiency when $M/\sigma_1 = 0$ is only about 50%. This reflects the fact that when $M = 0$, the parameter is point-identified and so Assumption 3 fails, leading to low power for the conditional test. We see that the conditional-FLCI hybrid substantially improves efficiency for small values of M/σ_1, while still retaining near-optimal performance as M/σ_1 grows large.

Results for $\Delta^{SDPB}(M)$: The middle panel of Figure 6 plots the efficiency ratio for each of our procedures as a function of M/σ_1 when $\Delta = \Delta^{SDPB}$. Again, the efficiency ratios for the conditional and hybridized confidence sets are (near-)optimal as M/σ_1 grows large, highlighting our asymptotic (near-)optimality results for these procedures in this simulation design. However, the efficiency ratios for the FLCIs steadily decrease as M/σ_1 increases, which reflects the fact that the FLCIs are not consistent in this simulation design when $M > 0$. We again see that the conditional-FLCI hybrid improves efficiency when M/σ_1 is small, while retaining near-optimal performance as M/σ_1 grows large.

Results for Δ^{SDI} The bottom panel of Figure 6 plots the efficiency ratio for each of our procedures as a function of M/σ_1 when $\Delta = \Delta^{SDI}$. As discussed above and summarized in Table 1, the conditions for asymptotic (near-)optimality do not hold for any of our procedures in this simulation design. Nonetheless, we see that the conditional and hybrid procedures still perform quite well for large values of M/σ_1, with efficiency approaching about 90%. We find this evidence encouraging as it shows that they may perform well asymptotically in cases where Assumption 3 fails. We again see that the efficiency of the FLCIs degrades as M/σ_1 grows, reflecting the fact that the FLCIs are

Figure 6: Simulation results: Median efficiency ratios for proposed procedures.

inconsistent under this simulation design when $M > 0$. Finally, we note that as in the previous simulations, the conditional-FLCI hybrid improves efficiency when M/σ_1 is small, while retaining similar performance to the conditional approach as M/σ_1 grows large.

8.3 Key takeaways from the simulation study

Summarizing our results from the simulation study, we find that the conditional confidence sets perform well across a range of applications but can have low power when the moments are not well-separated. In contrast, the FLCIs perform well in terms of excess length when the needed conditions for consistency and finite-sample optimality are satisfied (e.g. $\Delta^{SD}(M)$) but perform quite poorly elsewhere (e.g. $\Delta^{SDPB}(M)$ and $\Delta^{SDI}(M)$). Across each choice of Δ considered, the conditional-FLCI hybrid confidence set performs well in terms of excess length for all values of M/σ_1.

Based on these results, our recommended approach depends on the properties of the set Δ. In special cases, such as Δ^{SD}, in which the assumptions for consistency of the FLCIs are non-restrictive and the finite-sample optimality conditions are met, we recommend the FLCIs. Outside of these special cases, we recommend the use of a hybrid approach using the conditional confidence sets. Our simulations suggest that the FLCI-hybrid works quite well for cases that combine Δ^{SD} with sign or shape restrictions, and we thus recommend its use for such cases. For cases such as Δ^{RMI}, where we know that the FLCIs will be uninformative, we recommend the conditional-least favorable hybrid approach discussed in Appendix C.

9 Practical Guidance

This section provides practical guidance on how the methods developed in this paper can be used to assess the robustness of conclusions in difference-in-differences and event-study designs. The R package `HonestDiD` provides functions for easy and fast implementation of our recommended approach. As an overview, we recommend that researchers conduct the following steps to assess the robustness of their conclusions in difference-in-differences and related designs.

1) Estimate an "event-study"-type specification that produces a vector of estimates $\hat{\beta}$, consisting of "pre-period" coefficients $\hat{\beta}_{pre}$ and "post-period" coefficients $\hat{\beta}_{post}$, where the post-period coefficients have a causal interpretation under a suitable parallel trends assumption.

2) Perform a sensitivity analysis where inference is conducted under different assumptions about the set of possible violations of parallel trends Δ. For particular null hypotheses of interest, report the "breakdown" point at which the null can no longer be rejected (i.e. the least restrictive Δ (among some class of restrictions) at which the null is no longer rejected).

3) Provide economic benchmarks for evaluating the different choices of Δ. This can be done using information about potential confounding factors or using information from pre-treatment periods or placebo groups.

We provide additional details on each of the steps in the subsequent subsections. We also provide two illustrations of this approach in our applications in Section 10.

9.1   When to Use Our Methods

The methods in this paper can be applied in most practical settings in which researchers use an "event-study plot" to evaluate pre-existing trends. Specifically, to use our methods, we require one to have an estimator $\hat{\beta}_n$ with an asymptotic normal limit, $\sqrt{n}(\hat{\beta}_n - \beta) \to_d \mathcal{N}(0, \Sigma^*)$. Our methods are applicable when β satisfies the decomposition

$$\beta = \underbrace{\begin{pmatrix} 0 \\ \tau_{post} \end{pmatrix}}_{\textbf{causal effect}} + \underbrace{\begin{pmatrix} \delta_{pre} \\ \delta_{post} \end{pmatrix}}_{\textbf{bias from trend}} ,$$

where τ_{post} is a vector of causal parameters of interest, and one is willing to place restrictions on the relationship between the pre-period difference in trends δ_{pre} and the post-period bias δ_{post}. We now show how this framework can can accommodate two considerations that commonly arise in practice: i) staggered treatment timing, and ii) anticipatory effects.

Consider first the case with staggered treatment timing. A recent literature has shown that when there is staggered treatment timing and heterogeneous treatment effects across cohorts, the coefficients from standard two-way fixed effect models may not be causally interpretable. Specifically, even under a strong parallel trends assumption, the probability limit of the pre-period coefficients β_{pre} may be non-zero and the probability limit of the post-period coefficients β_{post} may not correspond with a convex weighted averages of cohort-specific treatment effects at a particular lag (Abraham and Sun, 2018). To address these issues, Abraham and Sun (2018) and Callaway and Sant'Anna (2019) provide alternative strategies for estimating weighted averages of cohort-specific treatment effects at a fixed lag (or lead) relative to treatment, which yield consistent estimates under a suitable parallel trends assumption.[30] The estimates for different leads and lags can be aggregated to form a vector of coefficients $\hat{\beta}_n$ resembling a traditional "event-study plot." Moreover, these estimates are asympotically normal under mild regularity conditions, and so our recommended sensitivity analysis can be then be applied to gauge sensitivity to violations of the needed parallel trends assumption. In settings with staggered treatment timing where the researcher is not willing to assume homogeneous treatment effects, we therefore recommend that researchers first use the

---

[30]Specifically, the generalized parallel trends assumption requires that the untreated potential outcomes for treated cohorts on average be parallel to the untreated outcomes for the never-treated group. See, e.g., Assumption 2 of Callaway and Sant'Anna (2019).

methods of Abraham and Sun (2018) or Callaway and Sant'Anna (2019) for estimation and then apply our results to conduct sensitivity analysis.

Next, consider the case where there may be changes in behavior in anticipation of the policy of interest. In such cases, β_{pre} may reflect the anticipatory effects of the policy rather than differences in untreated potential outcomes between the treated and untreated groups (Malani and Reif, 2015). This presents a challenge to our approach, which intuitively relies on placing restrictions on the relationship between pre-treatment and post-treatment differences in untreated potential outcomes. There is, however, a simple solution if one is willing to impose that anticipatory effects can only occur in a fixed window prior to the policy change — e.g., because policy is announced one year before it takes effect. In this case, one can re-normalize the definition of the "pre-treatment" period to be the period prior to when anticipatory effects can occur, in which case β_{pre} is determined only based on untreated potential outcomes.

9.2 Sensitivity Analysis

To conduct sensitivity analysis, we recommend that researchers report confidence sets under different assumptions about the set of possible violations of parallel trends Δ. This allows the reader to evaluate what assumptions need to be imposed in order to obtain informative inference.

For instance, in many cases a reasonable baseline choice for Δ may be a set of the form $\Delta^{SD}(M)$, which recall from Section 2.3 relaxes the assumption of linear differences in trends by imposing that the slope of the differential trend can change by no more than M in consecutive periods. By reporting robust confidence sets for different values of M, the researcher can then evaluate the extent to which their conclusions change as we allow for the possibility of greater non-linearities in the underlying trend. If there is a particular null hypothesis of interest (e.g. the treatment effect is zero), the researcher can report the "breakdown" value of M at which the null hypothesis can no longer be rejected.[31] The researcher could then also assess how these conclusions change if they impose further sign or shape restrictions motivated by economic knowledge — e.g. the bias is weakly positive, or the trend is monotone.

9.3 Evaluating a choice of Δ

When conducting sensitivity analysis over possible choices of Δ, researchers may wonder how to evaluate whether a particular choice of Δ is reasonable. For instance, suppose a researcher is considering sets of the form $\Delta^{SD}(M)$, which relaxes the assumption of linear differences in trends by allowing the slope of the differential trend to change by M across periods. How, then, should she evaluate what value of possible non-linearity M is reasonable? For ease of exposition, we focus

---

[31]Formally, one can show that if $\mathcal{C}(M)$ satisfies (11) for $\Delta^{SD}(M)$, then the sample breakdown value, $\hat{M}_{BE} := \inf\{M : \theta_0 \in \mathcal{C}(M)\}$, is a valid $(1-\alpha)$-level lower bound for the population breakdown value, $M_{BE} := \inf\{M : \theta_0 \in \mathcal{S}(\Delta^{SD}(M), \beta)\}$. That is, $\mathbb{P}\left(\hat{M}_{BE} \leqslant M_{BE}\right) \geqslant 1 - \alpha$.

on how to evaluate M in this case, although analogous remarks apply to conducting sensitivity for other forms of Δ as well.

We stress that nothing in the data itself can place an upper bound on the parameter M. This is because M places limits on the possible non-linearity in the counterfactual difference in trends δ_{post}, which is not identified from the data. Researchers therefore must use domain-specific knowledge to evaluate whether a particular choice of M is plausible in a given empirical setting. To help researchers do so, we discuss two approaches that benchmark possible values of M using knowledge of potential confounds and using untreated groups or periods. These benchmarking exercises may be useful in calibrating the restrictiveness of a particular set of assumptions about the counterfactual difference in trends.

Benchmarking using knowledge of potential confounds. One way to calibrate M is to consider the types of mechanisms which may produce a violation of parallel trends, and to benchmark M using knowledge of the likely magnitudes of those mechanisms. We provide an example of this type of reasoning in our application to Lovenheim and Willen (2019) in Section 10. In that setting, there are concerns that differential changes in education quality across states may produce non-parallel trends in employment outcomes, and we benchmark the magnitude of such changes using estimates of the effects of teacher value-added on employment outcomes from Chetty, Friedman and Rockoff (2014). This suggestion echoes Kahn-Lang and Lang (2018), who advocate for the use of domain-specific knowledge in justifying the parallel trends assumption. ∎

Benchmarking using untreated groups or periods. In some settings, researchers may have a set of untreated (placebo) groups who they think are "at least as dissimilar" as the groups used in their main analysis. In this case, it may be sensible to form a confidence interval for the largest change in slope between periods among the placebo groups, and use the upper bound of that interval as a benchmark for M.[32] This allows researchers to make statements of the form, "we can reject the null hypothesis unless we allow for changes in slope more than X times as large as the upper bound for the largest change observed for the placebo groups." Along similar lines, one could also construct an upper bound on the largest change in slope in the pre-period for the groups used in the main event-study of interest. One could then benchmark M in terms of multiples of the

---

[32] A simple way of forming such an interval is as follows: Let $\hat{\beta}^{placebo}$ be the vector of estimated event-study coefficients for the placebo groups, and $\delta^{placebo}$ its expectation. We wish to form a lower one-sided CI for $M^{placebo} := \max_j (A^{SD} \delta^{placebo})_j$, where A^{SD} is the matrix that creates second differences. Let $ub_j = (A^{SD}\hat{\beta}^{placebo})_j + \sigma_j z_{1-\alpha}$ be the upper bound of a standard one-sided confidence interval for the mean of $(A^{SD}\hat{\beta}^{placebo})_j$, where σ_j is the standard deviation of $(A^{SD}\hat{\beta}^{placebo})_j$. Define $ub := \max_j ub_j$. Then $(-\infty, ub]$ is a level $(1-\alpha)$ CI for $M^{placebo}$. Proper coverage of this CI follows from results by Berger and Hsu (1996) on the "union-intersection" approach, since $H_0 : M^{placebo} \geqslant m$ is equivalent to $H_0 : \mathbb{E}\left[A^{SD}\hat{\beta}^{placebo}\right] \in \bigcup_j \{\mu : \mu_j \geqslant m\}$.

largest value observed in the pre-period.[33] ∎

We note that while an upper bound on the parameter M may not be learned from the data, the observed pre-period data can possibly provide an informative *lower bound* on possible values of M. We now discuss how such a bound can be constructed, and how it should (and should not) be used.

Lower bounding M. Recall that M is an upper limit on how much the slope of δ can change between periods. Thus, the largest second difference in δ_{pre} must be no larger than M in absolute value. Since δ_{pre} is the expected value of the pre-period coefficients $\hat{\beta}_{pre}$, we can test this hypothesis using the pre-period coefficients. In particular, we can reject a value of M if we reject the null hypothesis $H_0 : \mathbb{E}\left[A^{SD}\hat{\beta}_{pre}\right] \leqslant M$, where A^{SD} is the matrix that produces the positive and negative second differences of $\hat{\beta}_{pre}$. This is again a test of a system of moment inequalities, and can be tested using the conditional test described in Section 4.[34]

This exercise is a sensible approach to test whether a given value of M is rejected by the data, and researchers should be skeptical of confidence sets relying on small values of M that are strongly rejected by the data. We caution, however, that this exercise provides only a *lower bound* on the possible value of M. The fact that the pre-period second differences in δ are small does not on its own place a bound on how large the second differences in the post-period could have been under the counterfactual. Thus, if a small value of M is not rejected by the data, we should not necessarily conclude that it is reasonable. We specifically caution against concluding that $M = 0$ (i.e. linear violations of parallel trends) is reasonable simply because it cannot be rejected by the data. We further note that relying on such tests raises standard concerns associated with pre-testing.

9.4 Choice of Confidence Sets and Computation

As discussed in detail in Section 8.3, we recommend that researchers select either the optimal FLCI or a hybridized conditional confidence set based upon the properties of their specified choice of Δ. For special cases (e.g. $\Delta = \Delta^{SD}(M)$) where the conditions for the consistency of the FLCIs are non-restrictive and the conditions for finite-sample near-optimality under parallel trends hold, we recommend that the researchers use the optimal FLCI. Otherwise, we recommend that researchers use a hybrid confidence set that hybridizes the conditional approach with the optimal FLCI or with

---

[33]We view the exercise of upper-bounding the largest change in the pre-period slope (M_{pre}) as a way of casting the assumption that $\Delta = \Delta^{SD}(M)$ in easily-interpretable units. If one instead wants to use this as the basis of an identifying assumption (i.e. assuming that the post-period change in slope is no larger than the largest pre-period change), then a tractable method for obtaining valid inference over the implied Δ is to first form a level $1 - \alpha/2$ upper bound on M_{pre}, say $\hat{M}_{pre}^{ub}$, and then estimate a robust $1 - \alpha/2$ level confidence set using $\Delta^{SD}(\hat{M}_{pre}^{ub})$.

[34]More generally, the null $H_0 : \delta \in \Delta = \{\delta : A\delta \leqslant d\}$ is equivalent to $H_0 : \exists \tau_{post}$ s.t. $\mathbb{E}\left[A(\hat{\beta} - M_{post}\tau_{post}) - d\right] \leqslant 0$, which can be tested using the conditional test, or the conditional-LF test described in the Appendix.

a least-favorable test. These confidence sets are all quick and straightforward to compute. For example, each of the sensitivity analysis plots in Section 10 take less than 10 minutes to produce on a 2014 Macbook Pro. Our R package `HonestDiD` by default chooses our recommended method depending on the specified choice of Δ, but allows the user to override these defaults.

10 Empirical Applications

We now apply our methods to two recently published papers. In doing so, we illustrate how empirical researchers can use our methods to conduct formal sensitivity analyses in difference-in-differences and event-study designs.

10.1 The effect of duty-to-bargain laws on long-run student outcomes

We first apply our results to the analysis of the effects of teacher collective bargaining on long-run student outcomes in Lovenheim and Willen (2019, henceforth LW).

LW study the impact of state-level public sector duty-to-bargain (DTB) laws, which mandated that school districts bargain in good faith with teachers' unions. The majority of these laws were passed in the 1960s-1980s. LW use data from the American Community Survey (ACS) from 2005-2012 to examine the impacts of these laws on the adult labor market outcomes of people who were students around the time that these laws were passed. Their identification strategy compares individuals across different states and different birth cohorts to exploit the differential timing of the passage of DTB laws across states. In particular, the authors estimate the following regression specification separately for men and women,
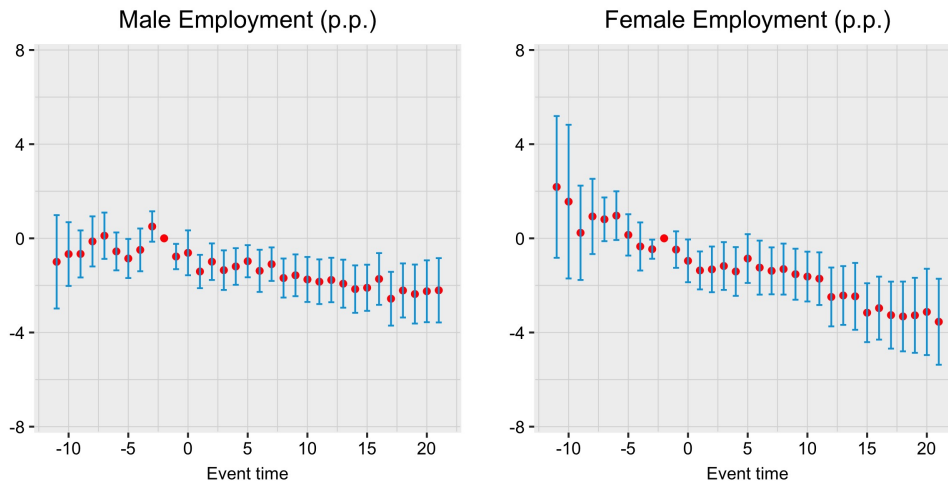
$$Y_{sct} = \sum_{r=-11}^{21} D_{scr}\beta_r + X'_{sct}\gamma + \lambda_{ct} + \phi_s + \epsilon_{sct}. \tag{28}$$

Y_{sct} is an average outcome for the cohort of students born in state s in cohort c in ACS calendar year t. D_{scr} is an indicator for whether state s passed a DTB law r years before cohort c turned age 18.[35] Thus, the coefficients $\{\hat{\beta}_r\}$ estimate the dynamic treatment effect (or placebo effect) r years after DTB passage. The remaining terms in the regression are time-varying controls, birth-cohort-by-ACS-year fixed effects, and state fixed effects. Standard errors are clustered at the state level, and we normalize the coefficient β_{-2} to 0.[36]

---

[35]$D_{sc,-11}$ is set to 1 if state s passed a law 11 years or more after cohort c turned 18. Likewise, $D_{sc,21}$ is set to 1 if state s passed a law 21 or more years before cohort c turned 18.

[36]LW normalize event time -1 to 0, but discuss how cohorts at event time -1 may have been partially treated, since LW impute the year that a student starts school with error. Since our robust confidence sets assume that there is no causal effect in the pre-period ($\tau_{pre} = 0$), we instead treat event-time -2 as the

Figure 7: Event-study coefficients $\{\beta_r\}$ for employment, estimated using the event-study specification in (28).
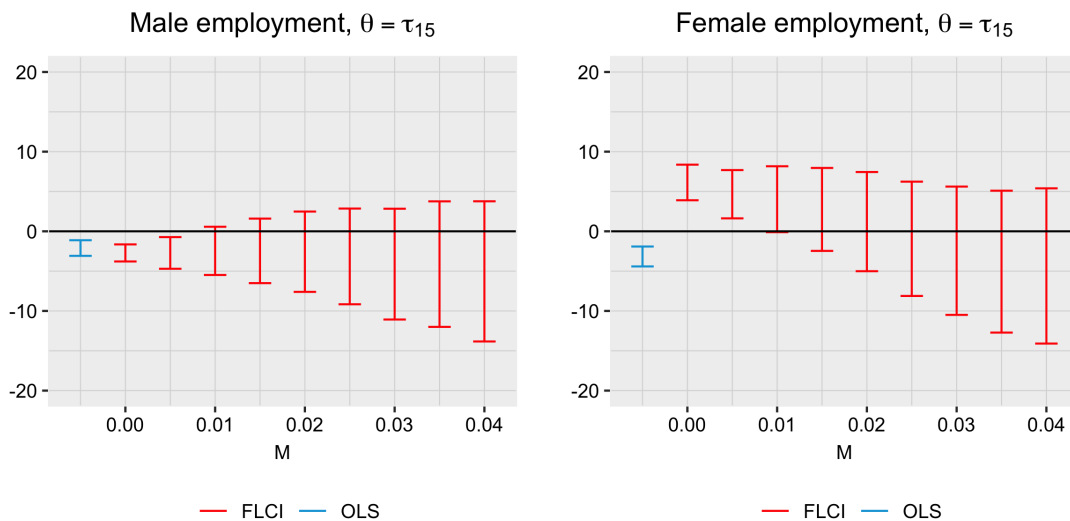

For illustration, we focus on the results where the outcome is employment. Figure 7 plots the event-study coefficients $\{\hat{\beta}_r\}$ estimated from the specification in (28). In the event-study for men (left panel), the pre-period coefficients are relatively close to zero, whereas the longer-run post-period coefficients are negative. Following standard practice, LW interpret the $\hat{\beta}_r$ as estimates of the causal effect of DTB laws on hours worked for men. By contrast, the results for women (right panel) suggest a downward-sloping pre-existing trend. LW therefore focus on the results for men, and "urge caution in lending a causal interpretation" to the results for women owing to the pre-existing trend. This example therefore allows us to show both how our methodology can be used to gauge the robustness of results when we do not detect an underlying trend, as well as to construct confidence intervals for the treatment effect of interest when parallel trends is clearly violated.

Figure 8 reports the results of a sensitivity analysis wherein we construct robust confidence sets under varying assumptions on the class of possible violations of parallel trends. To fix ideas, we conduct sensitivity analysis for the treatment effect on employment for the cohort 15 years after the passage of a DTB law, as in Table 2 of LW. In blue, we plot the original OLS confidence intervals for $\hat{\beta}_{15}$ from specification (28). In red, we plot robust confidence sets when $\Delta = \Delta^{SD}(M)$ for different values of M. We construct the confidence sets using FLCIs, following our recommendation for the case when $\Delta = \Delta^{SD}$. Turning first to the results for men in Figure 8, we see that the OLS point estimates are negative and the confidence intervals rule out zero. Our conclusions are fairly similar when we allow for the possibility of linear violations of parallel trends ($M = 0$). However, as we begin to allow for non-linear violations of parallel trends, the confidence sets become less informative. For instance, when $M = 0.01$, the upper bound of the confidence set is close to 0, whereas for $M = 0.04$ the upper bound is about 4 p.p. Turning next to the results for women in

---

reference period in our analysis.

Figure 8, we see that the original OLS estimates are negative and the confidence interval rules out 0. When we allow for linear violations of parallel trends ($M = 0$), however, the picture changes substantially owing to the pre-existing downward trend that is visible in Figure 7. Indeed, for $M = 0$ the robust confidence set contains only positive values. Intuitively, this is because $\hat{\beta}_{15}$ lies above the linear extrapolation of the downward trend in the pre-period. Once again, as M grows larger, the robust confidence sets become wider. Interestingly, we again include 0 within the robust confidence set for M around 0.01.
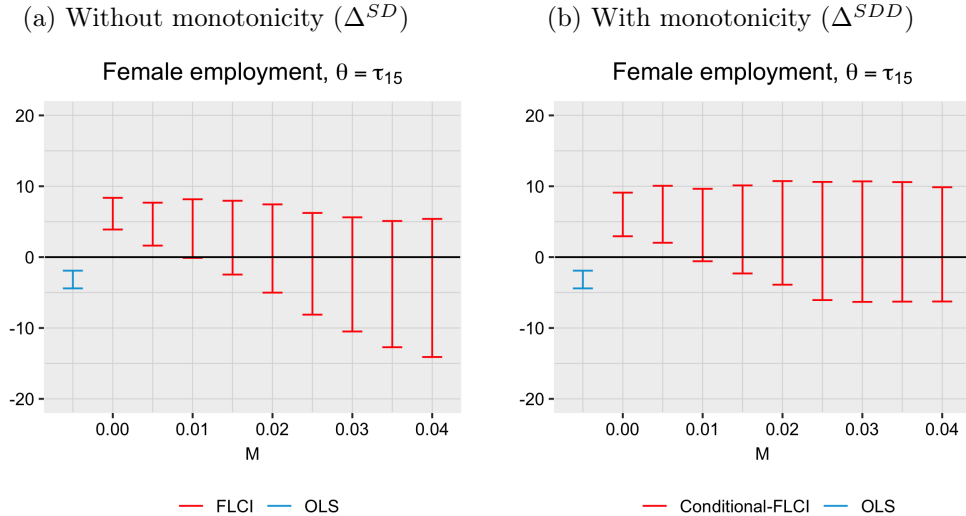
<div align="center">

Figure 8: Sensitivity analysis for $\theta = \tau_{15}$ using $\Delta = \Delta^{SD}(M)$

</div>

We next show how we can incorporate context-specific knowledge into the restrictions that we use in our sensitivity analysis. LW attribute the downward-sloping pre-trend in employment for women to "secular trends" affecting female labor supply. It seems reasonable to assume that such trends would have continued absent treatment, which motivates the shape restriction that δ is monotonically decreasing. In Figure 9, we show how the sensitivity analysis changes when we incorporate this shape restriction, comparing the results discussed above using $\Delta = \Delta^{SD}(M)$ to those using $\Delta = \Delta^{SDD} := \Delta^{SD}(M) \cap \{\delta : \delta_t \leqslant \delta_{t-1}, \forall t\}$. The confidence sets imposing the sign restrictions are constructed using the conditional-FLCI hybrid described in Section 7, following our recommendation for sets of the form Δ^{SDD}. With the added monotonicity restrictions, the lower bound of the robust confidence set never falls substantially below the lower bound of the OLS confidence interval, even when M is large. This is intuitive, since the restriction that δ is decreasing implies that the OLS coefficient is weakly downward biased. We thus see that imposing shape restrictions motivated by context-specific knowledge can be informative in obtaining tighter bounds.

The results above highlight that our confidence in the degree to which DTB laws affect employment depends on the parameter M, which represents the extent to which the slope of the difference

Figure 9: Sensitivity analysis with and without imposing monotonicity

(a) Without monotonicity (Δ^{SD}) (b) With monotonicity (Δ^{SDD})

in trends can change from period to period. Following the guidance provided in Section 9.3, we now discuss two ways for evaluating the reasonableness of various choices of M.

We begin with an evaluation of what values of M are rejected by the data using information from the pre-period. For men, the pre-period data does not provide an informative lower bound on M; we cannot reject $M = 0$ (i.e. linear trends) at conventional levels. For women, we obtain a p-value of 0.03 for the hypothesis that $M = 0$. The data thus provide some evidence against linear trends for women, although not overwhelmingly so. The lowest value of M which we do not reject at the 5% level for women is 0.28. The robust confidence set using Δ^{SD} with $M = 0.28$ is $[-58, 38]$, which includes quite large effects in either dimension. (We do not plot this confidence interval in Figure 8 so as to maintain a reasonable scale on the plot.) The robust confidence set with $\Delta = \Delta^{SDD}$ and $M = 0.28$ is $[-4.4, 43]$. Thus, again we obtain a somewhat informative lower bound when imposing sign restrictions motivated by context-specific knowledge, even at larger values of M.

Finally, we benchmark M using context-specific knowledge of possible mechanisms that could produce non-parallel trends. One concern in this setting is that education quality may have evolved differently in states that passed DTB laws even if the DTB laws had not been passed. This would lead to a violation of parallel trends if education quality directly impacts employment. Chetty et al. (2014) estimate that a 1 standard deviation increase in teacher value-added corresponds with a 0.4 percentage point increase in adult employment. We can use this estimate to calibrate M under the thought experiment in which violations of parallel trends are driven by different evolutions of teacher quality. (We can alternatively think of this as measuring other confounds in units of "teacher-quality equivalents.") For instance, a value of $M = 0.01$ would correspond with allowing for changes in the differential slope of teacher quality of about one-fourtieth of a standard deviation.
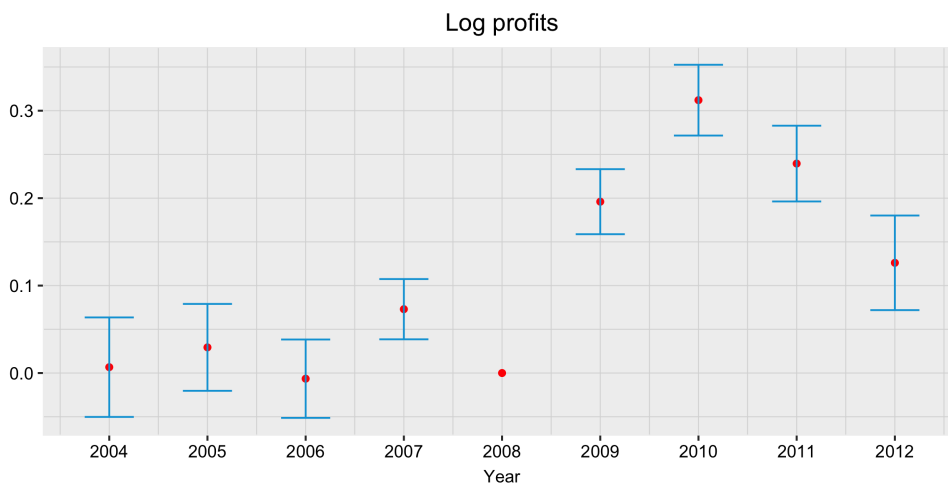
Since the robust confidence sets for both men and women begin to include 0 for M roughly equal to 0.01, the strength with which we can rule out a null effect depends on our assessment of the plausibility of such non-linearities.

10.2 Estimating the incidence of a value-added tax cut

Our second empirical application is to Benzarti and Carloni (2019, henceforth, BC), who study the incidence of a decrease in the value-added tax (VAT) on restaurants in France. France reduced its VAT on sit-down restaurants from 19.6 to 5.5 percent in July of 2009. BC analyze the impact of this change using a dynamic difference-in-differences design that compares restaurants to a control group of other market services firms, who were not impacted by the VAT change. Their regression specification is

$$Y_{irt} = \sum_{s=2004}^{2012} \beta_s \times 1[t = s] \times D_{ir} + \phi_i + \lambda_t + \epsilon_{irt}, \tag{29}$$

Figure 10: Event-study coefficients $\{\beta_s\}$ for log profits, estimated using the event-study specification in (29).
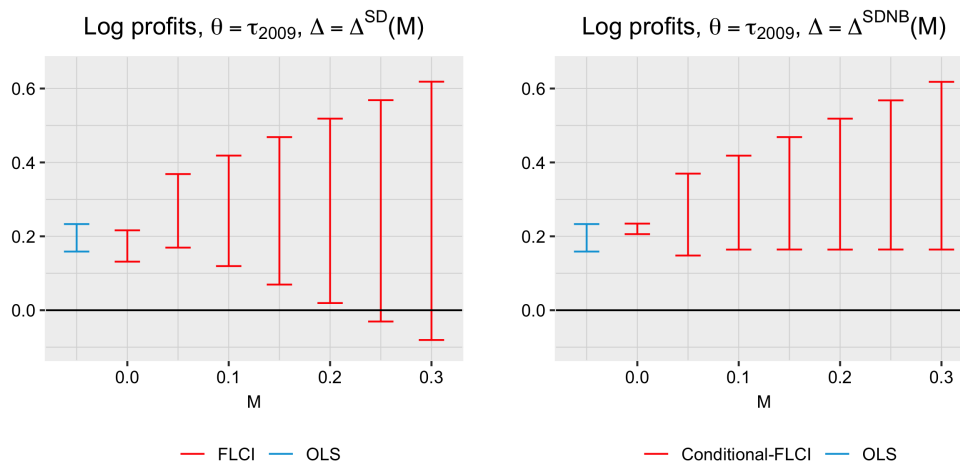

where Y_{irt} is an outcome of interest for firm i in region (i.e. département) r in year t; D_{ir} is an indicator for whether firm i in region r is a restaurant; and ϕ_i and λ_t are firm and year fixed effects. The coefficient β_{2008} is normalized to 0. Standard errors are clustered at the regional level. BZ are interested in the incidence of the VAT cut on firm owners, consumers, and employees, and they therefore estimate specification (29) using profits, prices, wages, and employment as the outcome. Their main finding is that there is a large effect on restaurant profits, and so we focus on this outcome. Figure 10 shows the estimated coefficients from specification (29) when the outcome is

the log of (before-tax) restaurant profits.[37] The figure shows relatively flat pre-trends, although there is a significant coefficient for 2007, and a seemingly large jump in 2009, the first treated year.

The left panel of Figure 10 shows a sensitivity analysis for the treatment effect in 2009 using $\Delta = \Delta^{SD}(M)$. As expected, the robust confidence sets become wider as we allow M to increase. Nonetheless, the confidence sets contain only positive values unless M exceeds 0.2. We can thus reject a null effect on profits in 2009 if we are willing to restrict the slope of the differential trend to change by no more than 20 log points between periods. To provide additional context for the magnitudes of M, we consider benchmarks using changes in slope in the pre-period, as discussed in Section 9.3. Using a 5% level test, we obtain a lower bound on M of 0.1 using the pre-period data. Intuitively, we reject small values of M owing to the spike in 2007 in the pre-period. Using a 5% test, we also obtain an upper bound on the largest change in slope in the pre-period of 0.2. Thus, we can reject a null effect in 2009 unless we are willing to allow for changes in slope of the underlying trend that are somewhat larger than the upper bound for the largest change in slope in the pre-period.

Figure 11: Sensitivity analysis for $\theta = \tau_{2009}$ using $\Delta = \Delta^{SD}(M)$ and $\Delta = \Delta^{SDNB}(M)$.

Using context-specific knowledge, BC argue that their estimates likely understate the effect of the VAT cut on profits. In particular, the VAT cut occurred at the same time that a payroll subsidy for restaurants was terminated, and the authors additionally suspect that the VAT cut may have reduced incentives to under-report costs to the tax authority. Since these mechanisms would have a negative effect on profits, BC therefore write that "a conservative interpretation of our results is that we are estimating a lower bound on the effect of the VAT cut on profits" (pg. 40). We can make this argument precise in our framework by imposing that the bias of the post-period event-study coefficients is negative. The right panel of Figure 11 imposes the additional constraint that the sign

---

[37]BZ drop observations with zero profits from their specification. They report finding similar results using the inverse hyperbolic sine.

of the bias be negative — that is, we set $\Delta = \Delta^{SDNB} := \Delta^{SD}(M) \cap \{\delta : \delta_{post} \leqslant 0\}$. With the added constraint on the sign of the bias, the robust confidence sets rule out effects on profits smaller than 15 log points for all values of M. The conclusion that the VAT cut had a substantial positive impact on profits is thus highly robust if we are willing to incorporate these sign restrictions.

Our results so far have focused on the robustness of the results for 2009, the first year after treatment. The results are substantially more sensitive if we consider the treatment effect for 2012, the least year in the sample. For instance, the FLCI using Δ^{SD} with $M = 0.1$, the lower bound on M from the pre-period, is $[-0.7, 1.5]$, which includes quite large effects in both directions. The reason for this is that secular differences in trends have more time to accumulate when we consider periods farther after treatment. Hence, in general, confidence sets for later post-periods will tend to grow faster as we increase the allowed degree of non-linearity in the underlying trends.[38]

11   Conclusion

This paper considers the problem of conducting inference in difference-in-differences and related designs that is robust to violations of the parallel trends assumption. Our approach formalizes the intuition underlying pre-trends tests, which is that the pre-trends are informative about the post-treatment differences in trends that would have occurred absent treatment. We consider a variety of possible restrictions on the class of possible differences in trend that make this intuition precise, and introduce powerful inference procedures that are uniformly valid so long as the difference in trends satisfies these restrictions. In practice, we recommend that researchers use our proposed confidence sets to conduct formal sensitivity analyses, in which they report confidence sets for the causal effect of interest under a variety of possible restrictions on the underlying trends. This exercise makes it transparent what assumptions are needed in order to obtain informative inference and assess whether those assumptions are plausible in a given setting.

We finish by discussing several potential directions for future research. First, while we have focused on robustness to violations of the parallel trends assumption, similar tools could be developed for alternative methods for inference in panel settings that rely on different identifying assumptions. Examples include the synthetic control method (Abadie, Diamond and Hainmueller, 2010), which relies on a "convex hull" assumption, and the non-linear difference-in-differences estimator of Athey and Imbens (2006). Second, throughout this paper, we adopt a super-population perspective on inference in difference-in-differences and event-study designs, in which the units are sampled randomly from some larger super-population. Athey and Imbens (2018) consider an alternative setting in which the observed units are viewed as a fixed, finite population and the critical identifying

---

[38]We note that this conclusion holds when Δ allows for smooth secular difference in trends that can grow over time (as when $\Delta = \Delta^{SD}(M)$). One could in principle specify Δ such that the bounds on the magnitude of the bias are tighter in periods farther from treatment. This might be sensible in certain cases — for instance, if one suspects that policy is passed in response to temporary, short-run shocks that revert to the mean over longer horizons.

assumption is that the timing of treatment is completely random. An interesting open question is how to conduct inference that is robust to possible non-random treatment timing in this finite-population setting. Finally, pre-trends testing is but one example of the wide use of placebo testing in quasi-experimental research designs. Similar methods could potentially be extended to conduct robust inference in other types of quasi-experimental research designs.

References

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 2005, *72* (1), 1–19.

__ , **Alexis Diamond, and Jens Hainmueller**, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, June 2010, *105* (490), 493–505.

__ **and Guido W. Imbens**, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 2006, *74* (1), 235–267.

Abraham, Sarah and Liyang Sun, "Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects," SSRN Scholarly Paper ID 3158747, Social Science Research Network, Rochester, NY April 2018.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

Andrews, Donald W. K. and Panle Jia Barwick, "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 2012, *80* (6), 2805–2826.

Andrews, Donald, Xu Cheng, and Patrik Guggenberger, "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," *Journal of Econometrics*, Forthcoming.

Andrews, Isaiah, Jonathan Roth, and Ariel Pakes, "Inference for Linear Conditional Moment Inequalities," Working Paper 2019.

__ , **Matthew Gentzkow, and Jesse Shapiro**, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 2017, *132* (4), 1553–1592.

__ , __ , **and** __ , "On the Informativeness of Descriptive Statistics for Structural Estimates," Technical Report 2019.

Angrist, Joshua D. and Jorn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press, 2009.

Arellano, Manuel, "Computing robust standard errors for within-groups estimators," *Oxford Bulletin of Economics and Statistics*, 1987, *49* (4), 431–434.

Armstrong, Timothy and Michal Kolesar, "Optimal Inference in a Class of Regression Models," *Econometrica*, 2018, *86*, 655–683.

_ **and** _ , "Sensitivity Analysis using Approximate Moment Condition Models," Technical Report 2019.

Ashenfelter, Orley, "Estimating the Effect of Training Programs on Earnings," *The Review of Economics and Statistics*, 1978, *60* (1), 47–57.

Athey, Susan and Guido Imbens, "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *arXiv:1808.05293 [cs, econ, math, stat]*, August 2018.

_ **and Guido W. Imbens**, "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 2006, *74* (2), 431–497.

Balke, Alexander and Judea Pearl, "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 1997, *92* (439), 1171–1176.

Benzarti, Youssef and Dorian Carloni, "Who Really Benefits from Consumption Tax Cuts? Evidence from a Large VAT Reform in France," *American Economic Journal: Economic Policy*, February 2019, *11* (1), 38–63.

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

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-In-Differences Estimates?," *The Quarterly Journal of Economics*, February 2004, *119* (1), 249–275.

Bhattacharya, Jay, Azeem M. Shaikh, and Edward Vytlacil, "Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization," *American Economic Review*, May 2008, *98* (2), 351–356.

Bhuller, Manudeep, Tarjei Havnes, Edwin Leuven, and Magne Mogstad, "Broadband Internet: An Information Superhighway to Sex Crime?," *The Review of Economics and Statistics*, 2013, *80*, 1237—-1266.

Bilinski, Alyssa and Laura A. Hatfield, "Seeking Evidence of Absence: Reconsidering Tests of Model Assumptions," *arXiv:1805.03273 [stat]*, May 2018.

Bonhomme, Stephanne and Martin Weidner, "Minimizing Sensitivity to Model Misspecification," Technical Report 2018.

Borusyak, Kirill and Xavier Jaravel, "Revisiting Event Study Designs," SSRN Scholarly Paper ID 2826228, Social Science Research Network, Rochester, NY August 2016.

Bugni, Federico A., Ivan A. Canay, and Xiaoxia Shi, "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 2017, *8* (1), 1–38.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," SSRN Scholarly Paper ID 3148250, Social Science Research Network, Rochester, NY March 2019.

Canay, Ivan and Azeem Shaikh, "Practical and Theoretical Advances in Inference for Partially Identified Models," in Bo Honoré, Ariel Pakes, Monika Piazzesi, and Larry Samuelson, eds., *Advances in Economics and Econometrics*, 2017, pp. 271–306.

Chen, Xiaohong, Timothy M. Christensen, and Elie Tamer, "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 2018, *86* (6), 1965–2018.

Chernozhukov, Victor, Whitney K. Newey, and Andres Santos, "Constrained Conditional Moment Restriction Models," *arXiv:1509.06311 [math, stat]*, September 2015. arXiv: 1509.06311.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, September 2014, *104* (9), 2633–2679.

Cho, JoonHwan and Thomas M. Russell, "Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments," October 2018.

Christensen, Timothy and Benjamin Connault, "Counterfactual Sensitivity and Robustness," *arXiv:1904.00989 [econ]*, April 2019. arXiv: 1904.00989.

Conley, Timothy, Christian Hansen, and Peter Rossi, "Plausibly Exogenous," *The Review of Economics and Statistics*, 2012, *94* (1), 260–272.

Conley, Timothy G. and Christopher R. Taber, "Inference with "Difference in Differences" with a Small Number of Policy Changes," *The Review of Economics and Statistics*, July 2010, *93* (1), 113–125.

Cox, Gregory and Xiaoxia Shi, "A Simple Uniformly Valid Test for Inequalities," *arXiv:1907.06317 [econ, math, stat]*, July 2019. arXiv: 1907.06317.

Cunningham, Scott, *Causal Inference: The Mixtape*, Unpublished manuscript, 2018.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Fuzzy Differences-in-Differences," *The Review of Economic Studies*, April 2018, *85* (2), 999–1028.

_ **and** _ , "Two-way fixed effects estimators with heterogeneous treatment effects," *arXiv:1803.08807 [econ]*, March 2018. arXiv: 1803.08807.

Ding, Peng and Tyler J. VanderWeele, "Sensitivity Analysis without Assumptions," *Epidemiology*, 2016, *27* (3), 368–377.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo, "The Economic Consequences of Hospital Admissions," 2018, *108* (2), 308–352.

Donald, Stephen G. and Kevin Lang, "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, 2007, *89* (2), 221–233.

Donoho, David L., "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 1994, *22* (1), 238–270.

Ferman, Bruno and Cristine Pinto, "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," *The Review of Economics and Statistics*, July 2018, *101* (3), 452–467.

Flynn, Zach, "Inference Based on Continuous Linear Inequalities via Semi-Infinite Programming," SSRN Scholarly Paper ID 3390788, Social Science Research Network, Rochester, NY May 2019.

Frandsen, Brigham R, "The Effects of Collective Bargaining Rights on Public Employee Compensation: Evidence from Teachers, Firefighters, and Police," *ILR Review*, 2016, *69* (1), 84–112.

Freyaldenhoven, Simon, Christian Hansen, and Jesse Shapiro, "Pre-event Trends in the Panel Event-study Design," *American Economic Review*, 2019, *109* (9), 3307–3338.

Gafarov, Bulat, "Inference in high-dimensional set-identified affine models," March 2019.

Ginther, Donna K., "Alternative Estimates of the Effect of Schooling on Earnings," *The Review of Economics and Statistics*, February 2000, *82* (1), 103–116.

González, Libertad, "Nonparametric bounds on the returns to language skills," *Journal of Applied Econometrics*, 2005, *20* (6), 771–795.

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," Working Paper 25018, National Bureau of Economic Research September 2018.

﹘ , "Public Insurance and Mortality: Evidence from Medicaid Implementation," *Journal of Public Economics*, 2018, *126* (1), 216–262.

Greenstone, Michael and Rema Hanna, "Environmental Regulations, Air and Water Pollution, and Infant Mortality in India," *American Economic Review*, October 2014, *104* (10), 3038–3072.

Gunsilius, Florian, "Bounds in continuous instrumental variable models," Technical Report 2019.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 1998, *66* (5), 1017–1098.

Heckman, James J. and Edward J. Vytlacil, "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the National Academy of Sciences*, April 1999, *96* (8), 4730.

Ho, Kate and Adam Rosen, "Partial Identification in Applied Research: Benefits and Challenges," in Bo Honoré, Ariel Pakes, Monika Piazzesi, and Larry Samuelson, eds., *Advances in Economics and Econometrics*, 2017, pp. 307–359.

Horowitz, Joel L. and Charles F. Manski, "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 1995, *63* (2), 281–302.

Hudson, Sally, Peter Hull, and Jack Liebersohn, "Interpreting Instrumented Difference-in-Differences," Working Paper 2017.

Hull, Peter, "Estimating Treatment Effects in Movers Designs," Technical Report 2018.

Imai, Kosuke and In Song Kim, "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data," 2019.

Imbens, Guido, "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 2003, *93* (2), 126–132.

___ **and Charles Manski**, "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 2004, *72* (6), 1845–1857.

Kadane, Joseph B., "Comparison of k-Class Estimators When the Disturbances Are Small," *Econometrica*, 1971, *39* (5), 723–737.

Kahn-Lang, Ariella and Kevin Lang, "The Promise and Pitfalls of Differences-in-Differences: Reflections on '16 and Pregnant' and Other Applications," Working Paper 24857, National Bureau of Economic Research July 2018.

Kaido, Hiroaki and Andres Santos, "Asymptotically Efficient Estimation of Models Defined by Convex Moment Inequalities," *Econometrica*, 2014, *82* (1), 387–413.

___ **, Francesca Molinari, and Jörg Stoye**, "Constraint Qualifications in Partial Identification," *arXiv:1908.09103 [econ]*, August 2019. arXiv: 1908.09103.

Keele, Luke J., Dylan S. Small, Jesse Y. Hsu, and Colin B. Fogarty, "Patterns of Effects and Sensitivity Analysis for Differences-in-Differences," Technical Report 2019.

Kim, Wooyoung, Koohyun Kwon, Soonwoo Kwon, and Sokbae Lee, "The identification power of smoothness assumptions in models with counterfactual outcomes," *Quantitative Economics*, 2018, *9* (2), 617–642.

Kline, Patrick and Andres Santos, "Sensitivity to missing data assumptions: Theory and an evaluation of the U.S. wage structure," *Quantitative Economics*, 2013, *4* (2), 231–267.

Kolesar, Michal and Christoph Rothe, "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review*, 2018, *108* (8), 2277–2304.

Kreider, Brent, John V. Pepper, Craig Gundersen, and Dean Jolliffe, "Identifying the Effects of SNAP (Food Stamps) on Child Health Outcomes When Participation Is Endogenous and Misreported," *Journal of the American Statistical Association*, September 2012, *107* (499), 958–975.

Kropko, Jonathan and Robert Kubinec, "Why the Two-Way Fixed Effects Model Is Difficult to Interpret, and What to Do About It," SSRN Scholarly Paper ID 3062619, Social Science Research Network, Rochester, NY May 2018.

Lechner, Michael, "The Estimation of Causal Effects by Difference-in-Difference Methods," *Foundations and Trends in Econometrics*, 2011, *4*, 165–224.

Lee, David, "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 2009, *76* (3), 1071—-1102.

Lee, Jason D., Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor, "Exact Post-Selection Inference, with Application to the Lasso," *The Annals of Statistics*, June 2016, *44* (3), 907–927.

Lee, Jin Young and Gary Solon, "The Fragility of Estimated Effects of Unilateral Divorce Laws on Divorce Rates," *The B.E. Journal of Economic Analysis & Policy*, 2011, *11* (1).

Li, Ker-Chau, "Honest Confidence Regions for Nonparametric Regression," *The Annals of Statistics*, 1989, *17* (3), 1001–1008.

Lovenheim, Michael F. and Alexander Willen, "The Long-Run Effects of Teacher Collective Bargaining," *American Economic Journal: Economic Policy*, 2019, *11* (3), 292–324.

Low, Mark G., "Bias-Variance Tradeoffs in Functional Estimation Problems," *The Annals of Statistics*, 1995, *23* (3), 824–835.

MacKinnon, James G. and Matthew D. Webb, "Randomization Inference For Difference-in-differences With Few Treated Clusters," Technical Report 1355, Economics Department, Queen's University January 2019.

Malani, Anup and Julian Reif, "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform," *Journal of Public Economics*, April 2015, *124*, 1–17.

Manski, Charles, *Partial Identification of Probability Distributions*, Springer, 2003.

Manski, Charles F., "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 1990, *80* (2), 319–323.

_ **and Daniel S. Nagin**, "Bounding Disagreements about Treatment Effects: A Case Study of Sentencing and Recidivism," *Sociological Methodology*, August 1998, *28* (1), 99–137.

_ **and John V. Pepper**, "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 2000, *68* (4), 997–1010.

_ **and** _ , "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 2018, *100* (2), 232–244.

Masten, Matthew A. and Alexandre Poirier, "Inference on Breakdown Frontiers," May 2019.

Meer, Jonathan and Jeremy West, "Effects of the Minimum Wage on Employment Dynamics," *Journal of Human Resources*, 2016, *51* (2), 500–522.

Molinari, Francesca, "Econometrics with Partial Identification," Technical Report 2019.

Moreira, Marcelo J. and Geert Ridder, "Efficiency Loss of Asymptotically Efficient Tests in an Instrumental Variables Regression," SSRN Scholarly Paper ID 3348716, Social Science Research Network, Rochester, NY March 2019.

Moulton, Brent R., "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit," *The Review of Economics and Statistics*, 1990, *72* (2), 334–338.

Mukhin, Yaroslav, "Sensitivity of Regular Estimators," Technical Report, arXiv Working Paper: arXiv:1805.08883 2018.

Nevo, Aviv and Adam M. Rosen, "Identification with Imperfect Instruments," *The Review of Economics and Statistics*, 2012, *94* (3), 659—-671.

Oster, Emily, "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 2019, *37* (2), 187–204.

Romano, Joseph P. and Azeem M. Shaikh, "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, September 2008, *138* (9), 2786–2807.

_ , _ , **and Michael Wolf**, "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 2014, *82* (5), 1979–2002.

Rosenbaum, Paul, "Sensitivity Analysis in Observational Studies," in B. S. Everitt and D. C. Howell, eds., *Encyclopedia of Statistics in Behavioral Science*, 2005.

_ **and Donald Rubin**, "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society. Series B*, 1983, *45* (2), 212–218.

Roth, Jonathan, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *Working paper*, 2019.

Sant'Anna, Pedro H. C. and Jun B. Zhao, "Doubly Robust Difference-in-Differences Estimators," *arXiv:1812.01723 [econ]*, September 2019. arXiv: 1812.01723.

Schrijver, Alexander, *Theory of Linear and Integer Programming*, Wiley-Interscience, 1986.

Sun, Yu and Karen X Yan, "Inference on Difference-in-Differences average treatment effects: A fixed-b approach," *Journal of Econometrics*, 2019, *211* (2), 560–588.

Słoczyński, Tymon, "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands," 2018.

Tamer, Elie, "Partial Identification in Econometrics," *Annual Review of Economics*, 2010, *2*, 167–195.

Wolfers, Justin, "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results," *American Economic Review*, 2006, *96*, 1802–1820.