Review of Economic Studies (2023) **90**, 2555–2591 doi:10.1093/restud/rdad018

Advance access publication 15 February 2023

A More Credible Approach to Parallel Trends

ASHESH RAMBACHAN

Microsoft Research New England and MIT

and

JONATHAN ROTH

Brown University

First version received July 2021*; Editorial decision October* 2022*; Accepted January* 2023 (Eds.)

This paper proposes tools for robust inference in difference-in-differences and event-study designs where the parallel trends assumption may be violated. Instead of requiring that parallel trends holds exactly, we impose restrictions on how different the post-treatment violations of parallel trends can be from the pre-treatment differences in trends ("pre-trends"). The causal parameter of interest is partially identified under these restrictions. We introduce two approaches that guarantee uniformly valid inference under the imposed restrictions, and we derive novel results showing that they have desirable power properties in our context. We illustrate how economic knowledge can inform the restrictions on the possible violations of parallel trends in two economic applications. We also highlight how our approach can be used to conduct sensitivity analyses showing what causal conclusions can be drawn under various restrictions on the possible violations of the parallel trends assumption.

Key words: Difference-in-differences, Event-study, Parallel trends, Robust inference, Sensitivity analysis, Partial identification

JEL codes: C1

1. INTRODUCTION

Researchers using difference-in-differences (DiD) and related methods are often unsure about the validity of the parallel trends assumption needed for point identification of the causal parameter of interest. It has therefore become common practice to assess the plausibility of the parallel trends assumption by testing for pre-treatment differences in trends ("pre-trends"). Although pre-trends tests are intuitive, recent research has shown that they may suffer from low power (Freyaldenhoven *et al.*, 2019; Bilinski and Hatfield, 2020; Kahn-Lang and Lang, 2020; Roth, 2022), and that conditioning the analysis on passing pre-trends tests introduces statistical issues related to pre-testing (Roth, 2022). How then should researchers proceed when they are unsure about the validity of the parallel trends assumption?

This paper proposes methods for robust inference and sensitivity analysis in empirical settings where the parallel trends assumption may not hold. Building on work by Manski and Pepper (2018), we show that the causal parameter of interest can be (partially) identified under

The editor in charge of this paper was Francesca Molinari.

a large class of restrictions that impose that the post-treatment violations of parallel trends cannot be "too different" from the pre-trends. We then introduce methods that yield uniformly valid inference for the treatment effect under the imposed restrictions. Intuitively, our inference methods account for both statistical uncertainty (we can only noisily estimate the true pre-trend) as well as "identification uncertainty" (even if the true pre-trend were known, we may not know exactly how to extrapolate it). Our approach thus formalizes the intuition motivating pre-trends testing while avoiding the statistical issues described above.

More concretely, we consider a setting in which the researcher estimates a vector of "event-study" coefficients $\hat{\beta} = (\hat{\beta}'_{pre}, \hat{\beta}'_{post})' \in \mathbb{R}^{\underline{T}+\bar{T}}$, where $\hat{\beta}_{pre}$ and $\hat{\beta}_{post}$ respectively correspond with estimates for $\underline{T}$ pre-treatment periods and $\bar{T}$ post-treatment periods. We assume that $\hat{\beta}$ is consistent for the reduced-form parameter β, which can be decomposed as

$$\beta = \underbrace{\begin{pmatrix} 0 \\ \tau_{post} \end{pmatrix}}_{=:\tau} + \underbrace{\begin{pmatrix} \delta_{pre} \\ \delta_{post} \end{pmatrix}}_{=:\delta}, \tag{1}$$

where τ is a causal parameter of interest that is assumed to be 0 in the pre-treatment period and δ is a bias from a difference in trends. For instance, in the canonical (non-staggered) DiD framework, $\hat{\beta}$ may be the coefficients from an "event-study regression" specification, τ the vector of period-specific average treatment effects on the treated (ATT) for some policy of interest, and δ the difference in trends of untreated potential outcomes between the treated and comparison groups. As we discuss in Section 2, this framework also applies to more complicated empirical settings, such as those with staggered treatment timing (*e.g.* Callaway and Sant'Anna, 2020; Sun and Abraham, 2020). The usual parallel trends assumption used to point identify τ_{post} is that $\delta_{post} = 0$, and researchers frequently assess the plausibility of this assumption by testing the null hypothesis $\delta_{pre} = 0$ (a "pre-trends" test).

Instead of imposing that the parallel trends assumption holds exactly, we place restrictions on the possible post-treatment differences in trends δ_{post} given the point identified pre-trends δ_{pre}. Such restrictions formalize the intuition motivating pre-trends tests, namely that pre-trends are informative about counterfactual post-treatment differences in trends. Formally, we assume that $\delta \in \Delta$ for some researcher-specified set Δ, and show that the causal parameter τ_{post} is partially identified under such restrictions.

Restrictions of this form can be used to formalize a wide variety of intuitions about possible violations of the parallel trends assumption that are commonly expressed in applied work. For example, as discussed in Manski and Pepper (2018), researchers may be willing to assume that the confounding factors that create post-treatment violations of parallel trends are similar in magnitude to those in the pre-treatment period. This intuition can be formalized by specifying a Δ that bounds the maximal post-treatment violation of parallel trends by a parameter $\bar{M}$ times the maximal pre-treatment violation of parallel trends. In other contexts, researchers are concerned about violations of parallel trends from secular trends that are assumed to evolve smoothly over time. This intuition can be formalized by bounding the extent to which the slope of the violation of parallel trends can change over time. We adopt a flexible framework that allows researchers to capture these intuitions, as well as many other restrictions that are implied by context-specific knowledge about possible confounding factors.

We then introduce methods to conduct uniformly valid inference on a scalar parameter of the form $\theta = l'\tau_{post}$ under the restriction $\delta \in \Delta$. As emphasized in the recent literature on pre-trends testing, the pre-treatment coefficients $\hat{\beta}_{pre}$ are often imprecise estimates of δ_{pre}. It is therefore important to introduce inference methods that account for the statistical uncertainty in the estimation of the event-study coefficients. We introduce two main inference approaches, with different desirable properties depending on the shape of Δ.

We first introduce a general inference approach that can accommodate a large class of choices for Δ. This approach is based on the observation that conducting inference on θ can be cast as the problem of testing a system of moment inequalities, allowing us to leverage a large econometrics literature on moment inequality testing (Canay and Shaikh, 2017 provide a recent review). The moments have a potentially large number of nuisance parameters that enter linearly, and we therefore consider an implementation of this approach based on the conditional and hybrid approaches of Andrews *et al.* (forthcoming, henceforth ARP), who considered moment inequalities with this structure. Uniform size control for these tests follows nearly immediately from results in ARP.

We then prove that the tests proposed by ARP have some desirable power properties in our context. First, the conditional and hybrid tests are consistent, in the sense that they have power approaching 1 against fixed alternatives outside of the identified set. Second, we prove that the conditional test has optimal local asymptotic power under a linear independence constraint qualification (LICQ) assumption. As described in Kaido *et al.* (2021), LICQ and related constraint qualifications have been used widely in the partial identification literature, and are often imposed to ensure size control. By contrast, we show that the ARP conditional test is asymptotically valid even when LICQ fails, but has optimal local asymptotic power when LICQ is satisfied. Intuitively, this result implies that the conditional test will perform well when the binding and non-binding moments are "far apart" relative to the sampling variation in the data. We provide several intuitive examples to illustrate when this result will and will not be applicable. Our result also implies that the ARP hybrid test will have near-optimal local asymptotic power under LICQ. These power results are new, and exploit additional structure in our context not contained in ARP.

Our second approach to inference is based on fixed length confidence intervals (FLCIs) (Donoho, 1994). FLCIs have desirable finite-sample guarantees for particular Δs of interest. In particular, results from Armstrong and Kolesár (2018, 2020b) imply that when Δ is convex and centrosymmetric, FLCIs have near-optimal expected length in the finite-sample normal model. These results are applicable for one of our leading examples, Δ^{SD}, which restricts the smoothness of the difference in trends. In Monte Carlo simulations, we find that the use of such FLCIs can lead to substantial power gains over the conditional/hybrid approaches for Δ^{SD} when the length of the identified set is short relative to the sampling variation in the data. This is intuitive since the asymptotic power guarantees for the conditional/hybrid approaches are in the asymptotic regime where sampling uncertainty is small relative to the length of the identified set, in contrast to the finite-sample guarantees for FLCIs. On the other hand, FLCIs are applicable for a much smaller class of Δs: indeed, we show that for many other choices of Δ, they will be inconsistent in the strong sense that power against fixed points outside the identified set need not converge to one asymptotically.

Based on our theoretical results and Monte Carlo simulations, we recommend the ARP hybrid approach for general forms of Δ, but prefer the FLCI approach in special cases (such as for Δ^{SD}) where the conditions for consistency and finite-sample near-optimality are met.

We recommend that applied researchers use our methods to construct robust confidence sets under economically-motivated restrictions on how the pre-trends relate to the post-treatment violations of parallel trends. Our tools can also be used to to conduct sensitivity analyses in which the researcher reports confidence sets under varying restrictions on the possible differences in trends. For example, if the researcher suspects that the confounding factors in the post-treatment periods are similar in magnitude to those in the pre-treatment periods, then it may be reasonable to impose that the post-treatment violations of parallel trends are no larger than the maximum pre-treatment violation of parallel trends. As a sensitivity analysis, the researcher might also report confidence sets that allow the maximum post-treatment violation of parallel trends to be up to $\bar{M}$ times larger than the maximum pre-treatment violation for different values of $\bar{M}$. The proposed sensitivity analysis thus follows the "top-down" approach described by Tamer (2010), in which the researcher reports what can be learned under a sequence of progressively weaker assumptions. Performing such sensitivity analyses makes clear what must be assumed about the possible differences in trends in order to draw specific causal conclusions, in line with the partial identification paradigm initiated by Charles Manski and co-authors in a variety of contexts (Manski, 1989, 1990, 2003, 2008). We provide the HonestDiD R and Stata packages to implement our methods.[1] We illustrate our recommended approach with applications to two recently published papers, in which we show how the choice of the restrictions Δ can be tailored to the economic context.

1.1. *Related literature*

The approach in this paper builds on the foundational partial identification analysis for DiD in Manski and Pepper (2018). Manski and Pepper consider identification under researcher-specified bounds on the magnitude of δ_{post} (what they call "bounded DiD variation"), and calibrate these bounds using the maximal pre-treatment violation of parallel trends in their empirical application on the effects of right-to-carry gun laws.[2] One of our leading classes of restrictions, Δ^{RM}, formalizes this calibration approach by bounding the magnitude of post-treatment violations of parallel trends by $\bar{M}$ times the maximal pre-treatment violation. Our framework also allows for many other intuitive restrictions—such as bounds on how far δ can deviate from linearity—and it can be applied to a variety of DiD estimators, including recent proposals for settings with staggered treatment timing. Most importantly, while Manski and Pepper (2018) provide a framework for identification, we provide inference methods to construct uniformly valid confidence sets for the treatment effect of interest. This allows applied researchers to account for statistical uncertainty in their analyses, which can be important since event-study coefficients are often imprecisely estimated in practice.

Several other recent papers consider various relaxations of the parallel trends assumption. Keele *et al.* (2019) develop techniques for testing the sensitivity of DiD designs to violations of the parallel trends assumption, but they do not incorporate information from the observed pre-trends in their sensitivity analysis. Empirical researchers commonly adjust for the extrapolation of a linear trend from the pre-treatment periods when there are concerns about violations of the parallel trends assumption, which is valid if the difference in trends is exactly linear (*e.g.*Bhuller *et al.*, 2013; Dobkin *et al.*, 2018; Goodman-Bacon, 2018, 2021). Our methods nest this approach as a special case, but allow for valid inference under less restrictive assumptions about the class of possible differences in trends (such as when δ is only approximately linear). Freyaldenhoven *et al.* (2019) propose a method that allows for violations of the parallel trends assumption but requires an additional covariate that is affected by the same confounding factors as the outcome but not by the treatment of interest. Ye *et al.* (2020) consider partial identification of treatment effects when there exist two control groups whose outcomes have a bracketing relationship with

1. The latest version of the R and Stata packages are respectively available at https://github.com/asheshrambachan/HonestDiD and https://github.com/mcaceresb/stata-honestdid/.

2. Manski and Pepper (2018) also consider "bounded time" and "bounded state" restrictions that bound how much the mean of $Y(0)$ can differ either across treatment groups or within-groups over time. Such restrictions could also be incorporated into our framework by augmenting the vector $\hat{\beta}$ to include group-specific sample averages.

the outcome of the treated group. Leavitt (2020) proposes an empirical Bayes approach calibrated to pre-treatment differences in trends, and Bilinski and Hatfield (2020) and Dette and Schumann (2020) propose non-inferiority approaches based on pre-tests for the magnitude of the pre-treatment violations of parallel trends.

Our methods address several concerns related to current empirical practice in DiD and event-study designs. First, pre-trends tests may be underpowered against meaningful violations of parallel trends, potentially leading to severe undercoverage of conventional confidence intervals (Freyaldenhoven *et al.*, 2019; Bilinski and Hatfield, 2020; Kahn-Lang and Lang, 2020; Roth, 2022). Second, statistical distortions from pre-trends tests may further undermine the performance of conventional inference procedures (Roth, 2022). Third, parametric approaches to controlling for pre-existing trends may be sensitive to functional form assumptions (Wolfers, 2006; Lee and Solon, 2011). We address these issues by providing tools for inference that do not rely on an exact parallel trends assumption, incorporate statistical uncertainty about the estimated event-study coefficients, and make clear the mapping between the researcher's assumptions about the potential differences in trends and the strength of their causal conclusions.

Our work complements a growing literature on the causal interpretation of event-study coefficients in two-way fixed effects models in the presence of staggered treatment timing and heterogeneous treatment effects (Borusyak and Jaravel, 2016; Callaway and Sant'Anna, 2020; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2020; Athey and Imbens, 2021; Goodman-Bacon, 2021). A key finding is that regression coefficients from conventional approaches may not produce convex weighted averages of treatment effects even if parallel trends holds. Several alternative estimators have been proposed that consistently estimate interpretable causal estimands under a suitable parallel trends assumption. Our methodology can be used in conjunction with these alternative estimators to assess their sensitivity to violations of the corresponding parallel trends assumption; see Section 2.1 for additional details.

More broadly, our work contributes to a larger econometric literature that uses partial identification to provide empirical researchers with tractable tools to conduct inference under assumptions that may be more credible in empirical practice; see, for example, Manski (2003, 2007, 2013), Tamer (2010), Ho and Rosen (2017), and Molinari (2020) for reviews.

2. MODEL SET-UP

2.1. *Event-study coefficients*

We suppose that the researcher has estimated a vector of "event-study coefficients" $\hat{\beta}_n \in \mathbb{R}^{\underline{T}+\bar{T}}$, which can be partitioned into vectors of coefficients 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 $\hat{\beta}_{n,post} \in \mathbb{R}^{\bar{T}}$. Event-study estimates of this form arise from non-staggered DiD as well as a variety of related estimators, as we illustrate with several examples

Example 1 (Non-staggered DiD)**.** Consider the canonical DiD setting in which we have a balanced panel of units from period $t = -\underline{T}, \dots, \bar{T}$, and units with $D_i = 1$ receive a treatment beginning in period $t = 1$, while units with $D_i = 0$ never receive the treatment. It is common to report DiD estimates of the form

$$\hat{\beta}_s = (\bar{Y}_{s1} - \bar{Y}_{s0}) - (\bar{Y}_{01} - \bar{Y}_{00}),$$

where $\bar{Y}_{sd}$ is the sample mean of the outcome for units with $D_i = d$ in period $t = s$. Intuitively, $\hat{\beta}_s$ compares the change in the mean outcome between period 0 and period s for the treated and comparison units. In this setting, the estimates $\hat{\beta}_s$ are numerically equivalent to the OLS coefficients from the regression

$$Y_{it} = \lambda_i + \phi_t + \sum_{s \neq 0} \beta_s \times \mathbb{1}[t = s] \times D_i + \epsilon_{it}. \tag{2}$$

In this case, $\hat{\beta}_{post}$ collects the estimated coefficients corresponding with treated periods, $(\hat{\beta}_1, \ldots, \hat{\beta}_{\bar{T}})$, while $\hat{\beta}_{pre}$ collects the estimated coefficients corresponding with periods before treatment $(\hat{\beta}_{-\underline{T}}, \ldots \hat{\beta}_{-1})$.

Example 2 (Staggered DiD)**.** Event-study coefficients can also be obtained from more complicated DiD procedures. For example, in settings with staggered treatment timing, Callaway and Sant'Anna (2020) propose event-study estimates of the form

$$\hat{\beta}_r = \sum_g w_g \widehat{ATT}(g, g + r),$$

where $\widehat{ATT}(g, t)$ is a DiD estimate that compares the evolution of the outcome for units first treated at period g to units first-treated after period t between time periods $g - 1$ and t, and the w_g are weights that sum to one (*e.g.* proportional to sample size). In this case, $\hat{\beta}_{post}$ collects the values of $\hat{\beta}_r$ for $r \geq 0$ (*i.e.* estimates where one of the groups is treated), and $\hat{\beta}_{pre}$ collects the values of $\hat{\beta}_r$ for values of $r < 0$. Several other related procedures have been proposed for constructing event-study coefficients in contexts with staggered treatment timing; see de Chaisemartin and D'Haultfœuille (2021) and Roth *et al.* (2022) for reviews.

Example 3 (Other related estimators)**.** Other examples of estimators that can be used to produce event-studies coefficients of the form considered here include the GMM procedure proposed by Freyaldenhoven *et al.* (2019), instrumental variables event-studies (Hudson *et al.*, 2017), as well estimators that flexibly control for differences in covariates between treated and comparison groups (*e.g.* Heckman *et al.*, 1998; Abadie, 2005; Sant'Anna and Zhao, 2020).

2.2. *Causal decomposition*

Under mild regularity conditions, all of the estimators described above will be asymptotically normally distributed, satisfying $\sqrt{n}(\hat{\beta}_n - \beta) \to \mathcal{N}(0, \Sigma^*)$ for some parameter vector β. We assume the parameter vector β satisfies the following causal decomposition.

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

$$\beta = \underbrace{\begin{pmatrix} \tau_{pre} \\ \tau_{post} \end{pmatrix}}_{=:\tau} + \underbrace{\begin{pmatrix} \delta_{pre} \\ \delta_{post} \end{pmatrix}}_{=:\delta} \quad \textit{with } \tau_{pre} = 0. \tag{3}$$

The first term, τ, represents the treatment effects of interest. We assume 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 comparison groups that would have occurred absent treatment. The parallel trends assumption imposes that $\delta_{post} = 0$, and therefore $\beta_{post} = \tau_{post}$ under parallel trends.[3]

3. Although our focus is on generalized DiD settings, our results apply to any setting where the researcher has an asymptotically normally distributed estimator $\hat{\beta}$ of a reduced form parameter β that satisfies (3). For example, $\hat{\beta}$ could

2.2.1. Example: non-staggered DiD (continued). Suppose the observed outcome satisfies $Y_{it} = D_i Y_{it}(1) + (1 - D_i)Y_{it}(0)$, where $Y_{it}(1)$ and $Y_{it}(0)$ are, respectively, the potential outcomes when unit i is ultimately treated/not treated. Assume further that there is no anticipation of treatment, so that $Y_{it}(1) = Y_{it}(0)$ for all $t < 1$. Then, for any s, under mild regularity conditions $\hat{\beta}_s$ will be consistent for

$$\beta_s = \tau_{ATT,s} + \underbrace{\mathbb{E}[Y_{is}(0) - Y_{i0}(0) \mid D_i = 1] - \mathbb{E}[Y_{is}(0) - Y_{i0}(0) \mid D_i = 0]}_{\text{differential trend} =: \delta_s},$$

where $\tau_{ATT,s} = \mathbb{E}[Y_{is}(1) - Y_{is}(0) \mid D_i = 1]$ is the average treatment effect on the treated in period s, and δ_s is the difference in trends in potential outcomes between period 0 and period s.[4] Since the no-anticipation assumption implies that $\tau_{ATT,s} = 0$ for $s < 0$, this yields the decomposition (3).

2.2.2. Example: staggered DiD (continued). Likewise, in the staggered DiD context, define $Y_{it}(g)$ to be the potential outcome for unit i in period t if they are first treated at period g and $Y_{it}(\infty)$ to be the never-treated potential outcome. Then $\hat{\beta}_r$ will be consistent for the parameter $\beta_r = \tau_r + \delta_r$, where $\tau_r = \sum_g w_g ATT(g, g+r)$ and $ATT(g, g+r) = \mathbb{E}\left[Y_{i,g+r}(g) - Y_{i,g+r}(\infty) \mid G_i = g\right]$ is the ATT in period $g+r$ for units first treated at period g. Likewise, $\delta_r = \sum w_g \delta_{g,g+r}$, where $\delta_{g,g+r} = \mathbb{E}\left[Y_{i,g+r}(\infty) - Y_{i,g-1}(\infty) \mid G_i = g\right] - \mathbb{E}\left[Y_{i,g+r}(\infty) - Y_{i,g-1}(\infty) \mid G_i > g+r\right]$ is the difference in trends in never-treated potential outcomes between units first treated at period g and units first treated after period $g+r$. Under a no-anticipation assumption, $\tau_r = 0$ for $r < 0$, which again yields the decomposition (3).

2.2.3. Example: other related estimators (continued). We can decompose β as in (3) for other estimators as well. For example, for event-study IVs (with non-staggered timing), τ_{post} is a vector containing the local average treatment effect for each period, and δ represents the asymptotic bias of the IV estimator (*e.g.* from violations of independence or exclusion) for each period. For methods that flexibly control for covariate differences between treated and comparison groups, τ_{post} is again a vector of ATTs, and δ_{post} represents a weighted average (across covariates) of the violation of the conditional parallel trends assumption.

2.3. *Target parameter and identification*

We suppose the target parameter is a linear combination of the post-treatment causal effects, $\theta := l'\tau_{post}$ for some known $\bar{T}$-vector l. For example, θ equals the tth period causal effect τ_t when the vector l equals the tth standard basis vector. Similarly, θ equals the average causal effect across all post-treatment periods when $l = (1/\bar{T}, \ldots, 1/\bar{T})'$.

We obtain partial identification of θ by assuming that δ lies in a researcher-specified set of possible differences in trends $\Delta \subseteq \mathbb{R}^{\underline{T}+\bar{T}}$. This nests the usual parallel trends assumption as the special case with $\Delta = \{\delta : \delta_{post} = 0\}$. Since $\delta_{pre} = \beta_{pre}$ is identified, the assumption that

correspond with a vector of first-differences or triple-differences estimators, and δ the violation of the corresponding identifying assumption.

4. We focus on the ATT as the target parameter, as in most of the DiD literature. If one is interested in the population-wide average treatment effect (ATE), one could obtain bounds on the ATE under restrictions on treatment effect heterogeneity, or other assumptions that allow one to bound the treatment effects for untreated units; see Manski and Pepper (2013) for an insightful discussion.

$\delta = (\delta_{pre}', \delta_{post}')' \in \Delta$ restricts the possible values of δ_{post} given the identified value of the pre-treatment difference in trends δ_{pre}.

It is natural to place restrictions on the relationship between δ_{pre} and δ_{post}, since applied researchers frequently test the null hypothesis $\delta_{pre} = 0$ in order to assess the plausibility of the assumption that $\delta_{post} = 0$. Our identification framework, which generalizes the partial identification framework in Manski and Pepper (2018), thus helps formalize the intuition motivating pre-trends testing.

Under the assumption that $\delta \in \Delta \neq \{\delta : \delta_{post} = 0\}$, the parameter θ will typically be set-identified. The *identified set* is the set of values for θ that are consistent with a given value of β under the restriction $\delta \in \Delta$,

$$\mathcal{S}(\beta, \Delta) := \left\{ \theta : \exists \delta \in \Delta, \tau_{post} \in \mathbb{R}^{\bar{T}} \text{ s.t. } l'\tau_{post} = \theta, \beta = \delta + \begin{pmatrix} 0 \\ \tau_{post} \end{pmatrix} \right\}. \tag{4}$$

When Δ is a closed and convex set, the identified set has a simple characterization.[5]

Lemma 2.1. *If Δ is closed and convex, then $\mathcal{S}(\beta, \Delta)$ is an interval in $\mathbb{R}$, $\mathcal{S}(\beta, \Delta) = [\theta^{lb}(\beta, \Delta), \theta^{ub}(\beta, \Delta)]$, where*

$$\theta^{lb}(\beta, \Delta) := l'\beta_{post} - \underbrace{\left(\max_{\delta} l'\delta_{post}, \ s.t. \ \delta \in \Delta, \delta_{pre} = \beta_{pre} \right)}_{=: b^{max}(\beta_{pre}, \Delta)}, \tag{5}$$

$$\theta^{ub}(\beta, \Delta) := l'\beta_{post} - \underbrace{\left(\min_{\delta} l'\delta_{post}, \ s.t. \ \delta \in \Delta, \delta_{pre} = \beta_{pre} \right)}_{=: b^{min}(\beta_{pre}, \Delta)}. \tag{6}$$

Proof. Re-arranging terms in (4), the identified set can be equivalently written as $\mathcal{S}(\beta, \Delta) = \{\theta : \exists \delta \in \Delta \text{ s.t. } \delta_{pre} = \beta_{pre}, \theta = l'\beta_{post} - l'\delta_{post}\}$. The result is then immediate. □

2.3.1. Example: non-staggered DiD (continued). In the three-period DiD model ($\underline{T} = \bar{T} = 1$), the ATT in period 1 is point identified if we assume that the counterfactual post-treatment difference in trends δ_1 is exactly zero (parallel trends). Instead, we assume $\delta = (\delta_{-1}, \delta_1)' \in \Delta$ for some set Δ. When Δ is closed and convex, the identified set for the ATT in period 1 is $[\beta_1 - b^{max}, \beta_1 - b^{min}]$, where $b^{max} = \max_{\delta} \delta_1$ s.t $(\delta_{-1}, \delta_1)' \in \Delta$ is the maximum possible bias of $\hat{\beta}_1$ given $\delta_{-1} = \beta_{-1}$ and b^{min} is defined analogously.

Additionally, it is immediate from the definition of the identified set in (4) that if Δ is the finite union of sets, $\Delta = \bigcup_{k=1}^{K} \Delta_k$, then the identified set is the union of the identified sets for its subcomponents,

$$\mathcal{S}(\beta, \Delta) = \bigcup_{k=1}^{K} \mathcal{S}(\beta, \Delta_k). \tag{7}$$

This fact will be useful, since several empirically relevant choices of Δ can be written as the finite union of convex sets, as we will see below.

5. Our focus is on inference on θ. If one were instead interested in estimating the endpoints of the identified, a natural estimator when Δ is closed and convex would be $[\theta^{lb}(\hat{\beta}, \Delta), \theta^{ub}(\hat{\beta}, \Delta)]$, which is the sample analogue to the bounds derived in Lemma 2.1; see Supplementary Material, Appendix C for additional discussion.

2.4. *Possible choices of* Δ

The class of possible differences in trends Δ must be specified by the researcher, and the choice of Δ will depend on the economic context. We highlight several possible choices of Δ that may be reasonable in empirical applications and formalize intuitive arguments that are commonly made by applied researchers regarding possible violations of parallel trends. Throughout our discussion, we write $\delta_{pre} = (\delta_{-\underline{T}}, \dots \delta_{-1})'$ and $\delta_{post} = (\delta_1, \dots \delta_{\bar{T}})'$, with δ_0 normalized to zero. This aligns the notation with Example 1, where δ corresponds to the difference in trends between treated and comparison groups, and δ_0 is normalized to zero.

2.4.1. Bounding relative magnitudes. In empirical applications, researchers may be willing to assume that the confounding factors which produce non-parallel trends in the post-treatment periods are not too much larger in magnitude than the confounding factors in the pre-treatment periods. In their empirical application to right-to-carry gun laws, Manski and Pepper (2018) operationalize this intuition by calibrating bounds on $|\delta_1|$ to the largest violations of parallel trends in the pre-treatment period (see their Table 3).[6] Such a restriction can be formalized in our framework by imposing that $\delta \in \Delta^{RM}(\bar{M})$ for $\bar{M} \geq 0$, where

$$\Delta^{RM}(\bar{M}) = \{\delta : \forall\, t \geq 0,\ |\delta_{t+1} - \delta_t| \leq \bar{M} \cdot \max_{s<0} |\delta_{s+1} - \delta_s|\}.$$

$\Delta^{RM}(\bar{M})$ bounds the maximum post-treatment violation of parallel trends between consecutive periods by $\bar{M}$ times the maximum pre-treatment violation of parallel trends. We use the abbreviation RM for "relative magnitudes." The choice $\Delta^{RM}(\bar{M})$ may be reasonable if the researcher suspects that possible violations of parallel trends are driven by confounding economic shocks that are of a similar magnitude to confounding economics shocks in the pre-period. When the number of pre-treatment and post-treatment periods is similar, a natural benchmark may be $\bar{M} = 1$, which bounds the worst-case post-treatment difference in trends by the equivalent maximum in the pre-treatment period.[7]

2.4.2. Example: non-staggered DiD (continued). In the three-period DiD model ($\underline{T} = \bar{T} = 1$), assuming $\delta \in \Delta^{RM}(\bar{M}) = \{(\delta_{-1}, \delta_1)' : |\delta_1| \leq \bar{M}|\delta_{-1}|\}$ bounds the magnitude of δ_1 based on the magnitude of δ_{-1}. The larger the magnitude of the pre-treatment violation in parallel trends, $|\delta_{-1}|$, the wider the range of possible post-treatment violations of parallel trends. ▲

2.4.3. Smoothness restrictions. In other empirical settings, researchers may be worried about confounding from secular trends (*e.g.* long-run changes in labour supply) that they suspect evolve smoothly over time. In such settings, it is common for empirical researchers to control for a linear group-specific time trend.[8] This approach is valid if the difference in trends is linear,

6. In their application, Manski and Pepper (2018) observe the outcome for the entire population of interest, and thus their observed pre-treatment data corresponds with δ_{pre} rather than $\hat{\beta}_{pre}$.

7. In settings where later pre-periods are thought to be more informative, the researcher could use a different value of $\bar{M}$ for different pre-periods, *e.g.* imposing that $|\delta_{t+1} - \delta_t| \leq \max\{|\delta_0 - \delta_{-1}|, 2 \cdot |\delta_{-1} - \delta_{-2}|\}$. Likewise, although we've considered bounds on changes across consecutive periods, the researcher could also impose bounds on changes across multiple periods, *e.g.* $|\delta_{t+2} - \delta_t| \leq \bar{M} \cdot \max_{s<-1} |\delta_{s+2} - \delta_s|$.

8. Specifically, researchers often augment specification (2) with group-specific linear trends, an approach 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, 2018, 2021).

i.e. $\Delta = \{\delta\colon \delta_t = \gamma \cdot t, \gamma \in \mathbb{R}\}$. There are often concerns, however, that the linear specification is not exactly correct (Wolfers, 2006; Lee and Solon, 2011). A natural relaxation is therefore to impose only that the differential trends evolve smoothly over time by bounding the extent to which its slope may change across consecutive periods. Such a restriction can be formalized in our framework by imposing that $\delta \in \Delta^{\mathrm{SD}}(M)$ for $M \geq 0$, where

$$\Delta^{\mathrm{SD}}(M) := \{\delta : |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leq M, \ \forall\, t\}. \tag{8}$$

The parameter $M \geq 0$ governs the amount by which the slope of δ can change between consecutive periods, and thus bounds the discrete analogue of the second derivative. We use the abbreviation SD for "second differences" or "second derivative."[9] In the special case where $M = 0$, $\Delta^{\mathrm{SD}}(0)$ requires that the difference in trends be exactly linear, which corresponds with the assumption underlying the parametric linear specification common in applied work.

2.4.4. Example: non-staggered DiD (continued). In the three-period DiD model, assuming the differential trend is exactly linear is equivalent to assuming $\Delta = \{\delta : \delta_1 = -\delta_{-1}\}$. Assuming $\delta \in \Delta^{\mathrm{SD}}(M)$ requires only that the linear extrapolation be *approximately* correct, $\delta_1 \in [-\delta_{-1} - M, -\delta_{-1} + M]$.

2.4.5. Combining smoothness and relative magnitudes bounds. In some contexts, researchers may be willing to assume that the difference in trends evolves relatively smoothly over time but may be unsure about the smoothness bound $M \geq 0$ introduced above. In such cases, it may be reasonable to assume that the possible non-linearities in the post-treatment difference in trends are bounded by the observed non-linearities in the pre-treatment difference in trends. This can be formalized with the restriction

$$\Delta^{SDRM}(\bar{M}) = \{\delta : \forall\, t \geq 0,\ |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leq \bar{M} \cdot \max_{s<0} |(\delta_{s+1} - \delta_s) - (\delta_s - \delta_{s-1})|\},$$

which bounds the maximum deviation from a linear trend in the post-treatment period by $\bar{M} \geq 0$ times the equivalent maximum in the pre-treatment period. The set $\Delta^{SDRM}(\bar{M})$ is thus similar to $\Delta^{\mathrm{SD}}(M)$ introduced above, except it allows the magnitude of the possible non-linearity to explicitly depend on the observed pre-trends.

2.4.6. Sign and monotonicity restrictions. Context-specific knowledge may sometimes also suggest sign or monotonicity restrictions on the differential trend. For instance, if the policy of interest occurs at the same time as a confounding policy change that we expect to have a positive effect on the outcome, we might restrict the post-treatment bias to be positive, $\delta \in \Delta^{PB} := \{\delta : \delta_t \geq 0\ \forall\ t \geq 0\}$. Likewise, there may be secular pre-existing trends that we expect would have continued following the treatment date.[10] We may then wish to impose that the differential trend be increasing, $\delta \in \Delta^I := \{\delta : \delta_t \geq \delta_{t-1}\ \forall\ t\}$, or monotone with unknown sign, $\delta \in \Delta^{Mon} := \Delta^I \cup (-\Delta^I)$. Sign and monotonicity restrictions may be combined with the

9. Restrictions on the second derivative of the conditional expectation function or density have been used in regression discontinuity settings (Kolesár and Rothe, 2018; Frandsen, 2016; Noack and Rothe, 2020). Smoothness restrictions are also used to obtain partial identification in *e.g.* Manski (1997); Kim *et al.* (2018).

10. Monotone violations of parallel trends are often discussed in applied work. For example, Lovenheim and Willen (2019a) 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.

previously discussed restrictions, such as $\Delta^{SDPB}(M) := \Delta^{\mathrm{SD}}(M) \cap \Delta^{PB}$, $\Delta^{SDI}(M) := \Delta^{\mathrm{SD}}(M) \cap \Delta^{I}$, and $\Delta^{RMI}(\bar{M}) := \Delta^{RM}(\bar{M}) \cap \Delta^{I}$.

2.4.7. Polyhedral restrictions. Although the restrictions described above will be sensible in many empirical contexts, researchers will often have context-specific knowledge that motivates alternative restrictions than what we introduced above. To accommodate such cases, we consider the broad class of Δs that can be written as polyhedra (sets defined by linear inequalities), or the finite union of polyhedra.

Definition 1 (Polyhedral restriction)**.** The class Δ is polyhedral if it takes the form $\Delta = \{\delta : A\delta \leq d\}$ for some known matrix A and vector d.

All of the examples described above can be written either as polyhedral restrictions or finite unions of such restrictions. For instance, $\Delta^{\mathrm{SD}}(M)$ and $\Delta^{SDPB}(M)$ can be written directly as polyhedra.[11] Likewise, $\Delta^{RM}(\bar{M})$ and $\Delta^{SDRM}(\bar{M})$ can be written as the finite union of polyhedra, where each polyhedron corresponds with a different location for the maximum pre-treatment violation.[12]

The class of (finite unions of) polyhedra is quite broad, and allows for a variety of other restrictions that may be relevant in empirical work. For example, researchers studying labour market training and related programs may be concerned about Ashenfelter's dip (Ashenfelter, 1978), in which earnings for the treated group trend downwards (relative to control) before treatment and upwards afterwards. In this type of setting, researchers might naturally use a polyhedral Δ to impose (i) restrictions on the signs of the pre-treatment and post-treatment biases, as well as (ii) restrictions on the magnitude of the rebound effect relative to the pre-treatment shock.

2.5. *Inferential goal*

As discussed above, the event study coefficients $\hat{\beta}_n$ will satisfy $\sqrt{n}(\hat{\beta}_n - \beta) \to_d \mathcal{N}(0, \Sigma^*)$ for a wide variety of commonly used estimators. This suggests the finite-sample normal approximation

$$\hat{\beta}_n \approx_d \mathcal{N}(\beta, \Sigma_n), \tag{9}$$

where $\approx_d$ denotes approximate equality in distribution and $\Sigma_n = \Sigma^*/n$. We will construct confidence sets that are uniformly valid for all parameter values θ in the identified set when the approximation in (9) holds exactly with Σ_n known. That is, we construct confidence sets $\mathcal{C}_n(\hat{\beta}_n, \Sigma_n)$ satisfying

$$\inf_{\delta \in \Delta, \tau} \inf_{\theta \in \mathcal{S}(\delta+\tau, \Delta)} \mathbb{P}_{\hat{\beta}_n \sim \mathcal{N}(\delta+\tau, \Sigma_n)} \left(\theta \in \mathcal{C}_n(\hat{\beta}_n, \Sigma_n)\right) \geq 1 - \alpha. \tag{10}$$

In Section 3.3, we will show that finite-sample size control in the normal model in the sense of (10) translates to uniform asymptotic size control over a large class of data-generating processes when Σ_n is replaced by a consistent estimate $\hat{\Sigma}_n$. That is, we will show that the constructed confidence sets further satisfy

11. In our ongoing three-period DiD example, $\Delta^{\mathrm{SD}}(M) = \{\delta : A^{\mathrm{SD}}\delta \leq d^{\mathrm{SD}}\}$ for $A^{\mathrm{SD}} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$ and $d^{\mathrm{SD}} = (M, M)'$. This generalizes naturally when there are multiple pre-periods and multiple post-periods.

12. For example, define the polyhedra $\Delta^{RM}_{s,+}(\bar{M}) = \{\delta : \forall\, t \geq 0, |\delta_{t+1} - \delta_t| \leq \bar{M}(\delta_{s+1} - \delta_s)\}$ and $\Delta^{RM}_{s,-} = \{\delta : \forall\, t \geq 0,\ |\delta_{t+1} - \delta_t| \leq -\bar{M}(\delta_{s+1} - \delta_s)\}$. Then $\Delta^{RM}(\bar{M}) = \bigcup_{s<0}(\Delta^{RM}_{s,+}(\bar{M}) \cup \Delta^{RM}_{s,-}(\bar{M}))$.

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\mathcal{S}(\delta_P+\tau_P,\Delta)} \mathbb{P}_P\left(\theta \in \mathcal{C}_n(\hat{\beta}_n, \hat{\Sigma}_n)\right) \geq 1-\alpha. \tag{11}$$

for a large class of distributions $\mathcal{P}$ such that $\delta_P \in \Delta$ for all $P \in \mathcal{P}$.

We will focus on constructing confidence sets for the case where Δ is a polyhedron. For the case where Δ is the finite union of polyhedra, a valid confidence set can be constructed by taking the union of the confidence sets for each of its components.

Lemma 2.2. *Suppose that for each $k = 1, \ldots, K$, the confidence set $\mathcal{C}_{n,k}(\hat{\beta}_n, \Sigma_n)$ satisfies (10) with $\Delta = \Delta_k$. Then the confidence set $\mathcal{C}_n(\hat{\beta}_n, \Sigma_n) = \bigcup_{k=1}^{K} \mathcal{C}_{n,k}(\hat{\beta}_n, \Sigma_n)$ satisfies (10) with $\Delta = \bigcup_{k=1}^{K} \Delta_k$.*

In the next two sections, we introduce two approaches to obtain confidence sets satisfying (10), with different desirable properties depending on the form of Δ. The first approach, based on moment inequalities, accommodates a wide range of restrictions Δ and has some desirable asymptotic power guarantees. The second approach, based on FLCIs, can potentially offer finite-sample power improvements for certain special classes of Δ of interest, such as $\Delta^{\mathrm{SD}}(M)$.

3. INFERENCE USING MOMENT INEQUALITIES

In this section, we introduce a general approach for inference that has good asymptotic properties over a large class of possible restrictions Δ. We show that inference on the partially identified parameter $\theta = l'\tau_{post}$ in this setting is equivalent to testing a system of moment inequalities with a potentially large number of nuisance parameters that enter the moments linearly. We consider an implementation based on the conditional approach developed in ARP, which allows us to obtain computationally tractable confidence sets with desirable power properties for many parameter configurations.

3.1. *Representation as a moment inequality problem with linear nuisance parameters*

Consider the problem of conducting inference on $\theta = l'\tau_{post}$ when Δ takes the polyhedral form $\Delta = \{\delta : A\delta \leq d\}$. We will develop tests that control size under the null hypothesis $H_0 : \theta = \bar{\theta}, \delta \in \Delta$ when the normal approximation (9) holds exactly with known variance matrix Σ_n. In Section 3.3, we will provide conditions under which size control in the finite sample normal model translates to uniform asymptotic size control over a large class of data-generating processes.

As a first step, we show that testing H_0 is equivalent to testing a system of moment inequalities with linear nuisance parameters in the normal model. Observe that if $\hat{\beta}_n \sim \mathcal{N}(\beta, \Sigma_n)$ for β satisfying (3), then $\mathbb{E}_{\hat{\beta}_n\sim\mathcal{N}(\beta,\Sigma_n)}[\hat{\beta}_n - \tau] = \delta$. It follows that $\delta \in \Delta = \{\delta : A\delta \leq d\}$ if and only if $\mathbb{E}_{\hat{\beta}_n\sim\mathcal{N}(\beta,\Sigma_n)}[A\hat{\beta}_n - A\tau] \leq d$. Defining $Y_n = A\hat{\beta}_n - d$ and $L_{post} = [0, I]'$ to be the matrix such that $\tau = L_{post}\tau_{post}$, it is immediate that the null hypothesis H_0 is equivalent to the composite null

$$H_0 : \exists \tau_{post} \in \mathbb{R}^{\bar{T}} \text{ s.t. } l'\tau_{post} = \bar{\theta} \quad \text{and} \quad \mathbb{E}_{\hat{\beta}_n\sim\mathcal{N}(\beta,\Sigma_n)}\left[Y_n - AL_{post}\tau_{post}\right] \leq 0. \tag{12}$$

Testing the null hypothesis H_0 is therefore equivalent to testing that the moment inequalities $\mathbb{E}_{\hat{\beta}_n\sim\mathcal{N}(\beta,\Sigma_n)}[Y_n - AL_{post}\tau_{post}] \leq 0$ hold for some value of τ_{post} satisfying $l'\tau_{post} = \theta$.

For the purposes of developing tests, it will be useful to re-cast this null hypothesis in terms of moments involving an unrestricted nuisance parameter $\tilde{\tau}$ of dimension $\bar{T} - 1$. By applying a change of basis to the matrix AL_{post}, we can re-write the expression $AL_{post}\tau_{post}$ as $\tilde{A}(\theta, \tilde{\tau}')'$ for $\tilde{\tau} \in \mathbb{R}^{\bar{T}-1}$.[13] The null H_0 is then equivalent to

$$H_0 : \exists \tilde{\tau} \in \mathbb{R}^{\bar{T}-1} \text{ s.t. } \mathbb{E}\left[\tilde{Y}_n(\bar{\theta}) - \tilde{X}\tilde{\tau}\right] \leq 0, \tag{13}$$

where $\tilde{Y}(\bar{\theta}) = Y_n - \tilde{A}_{(\cdot,1)}\bar{\theta}$ and $\tilde{X} = \tilde{A}_{(\cdot,-1)}$.[14] Since $\tilde{Y}_n(\bar{\theta})$ is normally distributed with covariance matrix $\tilde{\Sigma}_n = A\Sigma_n A'$ under the finite-sample normal model, testing $H_0 : \theta = \bar{\theta}, \delta \in \Delta$ is equivalent to testing a system of moment inequalities with linear nuisance parameters.

The testing problem (13) is a special case of the problem studied in ARP, which focuses on testing null hypotheses of the form $H_0 : \exists \tau \text{ s.t. } \mathbb{E}[Y(\theta) - X\tau \mid X] \leq 0$ (a.s.). Our setting is a special case of this framework in which: i) the variable X takes the degenerate distribution $X = \tilde{X}$, and ii) $Y(\theta) = \tilde{Y}(\theta)$ is linear in θ. This additional structure will play an important role in the development of our asymptotic power results below.

3.2. *Constructing conditional and hybrid confidence sets*

We consider tests for the system of moment inequalities described above using the conditional and hybrid methods proposed by ARP. This is for both computational and efficiency reasons. From the computational perspective, a practical challenge to testing the hypothesis (13) in our setting is that the dimension of the nuisance parameter $\tilde{\tau}$ is $\bar{T} - 1$, and thus will be large if there are many post-treatment periods. For example, 5 of the 12 recent event-study papers reviewed in Roth (2022) have $\bar{T} > 10$. This renders many moment inequality methods, especially those which rely on test inversion over a grid for the full parameter vector, computationally infeasible. To tractably deal with the nuisance parameter, we consider tests based on the conditional and hybrid approaches of ARP, which directly exploit the linear structure of the hypothesis (13) to deliver computationally tractable tests even when the number of post-treatment periods $\bar{T}$ is large.[15] From the perspective of power, we will show that the tests proposed by ARP have (near-)optimal local asymptotic power in our setting when an LICQ condition is satisfied.

We briefly sketch the construction of the conditional testing approach in our setting, and refer the reader to ARP for full details. These tests are implemented in the HonestDiD R and Stata packages that accompany the paper.

13. 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 $\tilde{A} := AL_{post}\Gamma^{-1}$. Then $AL_{post}\tau_{post} = \tilde{A}\Gamma\tau_{post} = \tilde{A}\left(\begin{array}{c}\theta \\ \underbrace{\Gamma_{(-1,\cdot)}\tau_{post}}_{:=\tilde{\tau}}\end{array}\right)$. If $\bar{T} = 1$, then $\tilde{\tau}$ is 0-dimensional and should be interpreted as 0.

14. We use the notation $V_{(\cdot,1)}$ to denote the first column of a matrix V, and $V_{(\cdot,-1)}$ to denote the matrix containing all but the first column of V.

15. Other moment inequality methods have been proposed for subvector inference, but typically do not exploit the linear structure of our setting—see, e.g. Romano and Shaikh (2008), Chernozhukov *et al.* (2015), Bugni *et al.* (2017) and Chen *et al.* (2018) and Kaido *et al.* (2019). Cho and Russell (2019), Gafarov (2019), 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. More recently, Cox and Shi (2022) introduced new tests for the linear moment inequality setting in ARP; see Section 3.5 below for further discussion.

3.2.1. Conditional confidence sets. Suppose we wish to test (13) for some fixed $\bar{\theta}$. The conditional testing approach considers tests based on the profiled test statistic

$$\hat{\eta} := \min_{\eta,\tilde{\tau}} \eta \text{ s.t. } \tilde{Y}_n(\bar{\theta}) - \tilde{X}\tilde{\tau} \leq \tilde{\sigma}_n \cdot \eta, \tag{14}$$

where $\tilde{\sigma}_n = \sqrt{\text{diag}(\tilde{\Sigma}_n)}$. This linear program selects the value of the nuisance parameters $\tilde{\tau} \in \mathbb{R}^{\bar{T}-1}$ that minimizes the maximum studentized moment. Duality results from linear programming (*e.g.* Schrijver (1986), Section 7.4) imply that the value $\hat{\eta}$ obtained from the primal program (14) equals the optimal value of the dual program[16]

$$\hat{\eta} = \max_{\gamma} \gamma' \tilde{Y}_n(\bar{\theta}) \text{ s.t. } \gamma' \tilde{X} = 0, \gamma' \tilde{\sigma}_n = 1, \quad \gamma \geq 0. \tag{15}$$

If a vector γ_* is optimal in the dual problem above, then it is a vector of Lagrange multipliers for the primal problem. Standard results in linear programming imply that the optimum is always obtained at one of the finite set of vertices, $V(\Sigma_n)$ (also known as the set of basic feasible solutions). We denote by $\hat{V}_n \subset V(\Sigma_n)$ the set of optimal vertices of the dual program.[17]

To construct critical values, ARP use the fact that the distribution of $\hat{\eta}$ has a truncated normal distribution conditional on the event that γ_* is optimal in the dual problem. Specifically,

$$\hat{\eta} \,|\, \{\gamma_* \in \hat{V}_n, S_n = s\} \sim \xi \,|\, \xi \in [\upsilon^{lo}, \upsilon^{up}],$$

where $\xi \sim \mathcal{N}(\gamma_*' \tilde{\mu}(\bar{\theta}), \gamma_*' \tilde{\Sigma}_n \gamma_*)$, $\tilde{\mu}(\bar{\theta}) = \mathbb{E}[\tilde{Y}_n(\bar{\theta})]$, $S_n = (I - (\tilde{\Sigma}_n \gamma_* / \gamma_*' \tilde{\Sigma}_n \gamma_*) \gamma_*') \tilde{Y}_n(\bar{\theta})$, and $\upsilon^{lo}, \upsilon^{up}$ are known functions of $\tilde{\Sigma}_n, s, \gamma_*$ (see Lemma 1 in ARP).[18] Intuitively, the distribution of $\hat{\eta}$ depends on the vector $\tilde{\mu}(\bar{\theta})$, and so to eliminate the dependence on the components of $\tilde{\mu}(\bar{\theta})$ other than $\gamma' \tilde{\mu}(\bar{\theta})$, we condition on S_n, which is a sufficient statistic for the components of $\tilde{\mu}(\bar{\theta})$ that are orthogonal to $\gamma_*' \tilde{\mu}(\bar{\theta})$.

ARP show that all quantiles of the conditional distribution of $\hat{\eta}$ in the previous display are increasing in $\gamma_*' \tilde{\mu}(\bar{\theta})$. Moreover, the null hypothesis (13) implies $\gamma_*' \tilde{\mu}(\bar{\theta}) \leq 0$. To see why this is the case, note that the definition of the dual problem (15) implies that $\gamma_* \geq 0$ and $\gamma_*' \tilde{X} = 0$, whereas the null hypothesis implies that there exists $\tilde{\tau}$ such that $\tilde{\mu}(\bar{\theta}) - \tilde{X}\tilde{\tau} \leq 0$. It follows that $\gamma_*' \tilde{\mu}(\bar{\theta}) = \gamma_*' (\tilde{\mu}(\bar{\theta}) - \tilde{X}\tilde{\tau}) \leq 0$ under the null. The ARP conditional test therefore uses the critical value $\max\{0, c_{C,\alpha}\}$, where $c_{C,\alpha}$ is the $1-\alpha$ quantile of the truncated normal distribution $\xi \,|\, \xi \in [\upsilon^{lo}, \upsilon^{up}]$ under the worst-case assumption that $\gamma_*' \tilde{\mu}(\bar{\theta}) = 0$.[19] We denote by

16. Technically, the duality results require that $\hat{\eta}$ be finite. However, one can show that $\hat{\eta}$ is finite with probability 1, unless the span of $\tilde{X}$ contains a vector with all negative entries, in which case the identified set for θ is the real line. We therefore trivially define our test never to reject if $\hat{\eta} = -\infty$.

17. In general, there may not be a unique solution to the dual program. ARP show that in the context of the finite sample normal, 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. In the finite sample normal model, it thus suffices to condition on the event that a vector $\gamma_* \in \hat{V}$. Our conditions for asymptotic validity of the conditional test below, however, ensure that the optimal vertex will be unique w.p.a. 1.

18. The cut-offs υ^{lo} and υ^{up} are the maximum and minimum of the set $\{x : x = \max_{\gamma \in F_n} \gamma'(s + (\tilde{\Sigma}_n \gamma_* / \gamma_*' \tilde{\Sigma}_n \gamma_*) x)\}$ when $\gamma_*' \tilde{\Sigma}_n \gamma_* \neq 0$, where F_n is the feasible set of the dual program (15). When $\gamma_*' \tilde{\Sigma}_n \gamma_* = 0$, we define $\upsilon^{lo} = -\infty$ and $\upsilon^{up} = \infty$, so the conditional test rejects if and only if $\hat{\eta} > 0$.

19. As noted in ARP, the truncation at 0 is not necessary for the conditional test to control size in the finite sample normal model, but it simplifies asymptotic arguments. It also prevents the test from rejecting when all moments are satisfied in sample.

$\psi_\alpha^C(\hat{\beta}_n, A, d, \bar{\theta}, \Sigma_n)$ an indicator for whether the conditional test rejects the null that $\theta = \bar{\theta}$ for $\Delta = \{\delta : A\delta \leq d\}$.

We can then form a confidence set for θ by test inversion, $\mathcal{C}^C_{\alpha,n}(\hat{\beta}_n, \Sigma_n) := \{\bar{\theta} : \psi_\alpha^C(\hat{\beta}_n, A, d, \bar{\theta}, \Sigma_n) = 0\}$. The construction of the conditional test implies that $\mathbb{E}_{\hat{\beta}_n \sim \mathcal{N}(\delta+\tau,\, \Sigma_n)}[\psi_\alpha^C(\hat{\beta}_n, A, d, l'\tau_{post}, \Sigma_n)] \leq \alpha$ for any $\delta \in \Delta$. It therefore follows that $\mathcal{C}^C_{\alpha,n}(\hat{\beta}_n, \Sigma_n)$ satisfies the finite-sample coverage requirement (10). In Section 3.3 below, we show that coverage in the normal model translates to uniform asymptotic coverage over a large class of DGPs.

Example 4. An instructive example is when $\bar{T} = 1$ (so that there are no nuisance parameters), and $\tilde{\Sigma}_n = I$. Then $\hat{\eta} = \max_j \tilde{Y}_{n,j}$ is the maximum component of $\tilde{Y}_n$, $v^{lo} = \max_{j \neq \hat{j}} \tilde{Y}_{n,j}$ is the second-largest element of $\tilde{Y}_n$ (where $\hat{j}$ denotes the index of the max), and $v^{up} = \infty$. Thus, the conditional test rejects when $\hat{\eta}$ exceeds the $1 - \alpha$ quantile of the standard normal distribution truncated to $[v^{lo}, \infty)$. Intuitively, this means that the conditional test will tend to reject when the maximum sample moment is far enough away from the second-largest sample moment. Two special cases are worth special consideration. First, consider the case where in population one moment is violated and the remaining moments are very slack, *e.g.* $\tilde{\mu}_1 > 0$ while $\tilde{\mu}_j \ll 0$ for $j \neq 1$. Then with high probability $\hat{\eta}$ will equal $\tilde{Y}_{n,1}$ and v^{lo} will be very negative. Thus, the conditional test will behave similarly to a one-sided t-test using $\tilde{Y}_{n,1}$, which can be shown to be the most powerful test in the finite-sample normal model in this example. On the other hand, if $\mu_1 \approx \mu_2 > 0$, then the maximum and second-largest sample moments (*i.e.* $\hat{\eta}$ and v^{lo}) will be close together with high probability, so the conditional test may not reject with substantial probability even if both μ_1 and μ_2 are large, and thus the conditional test may have poor power. To improve power in these settings where the binding and non-binding moments are close together (relative to sampling variation), ARP introduce a "hybrid" test, which we describe next.

3.2.2. Hybrid confidence sets. ARP propose a "hybrid" test that combines the conditioning approach above with a test based on the "least-favourable" assumption that $\tilde{\mu}(\bar{\theta}) = 0$. In particular, ARP show that the distribution of $\hat{\eta}$ under the null is bounded above (in the sense of first-order stochastic dominance) by the distribution of $\hat{\eta}$ when $\tilde{\mu}(\bar{\theta}) = 0$ (see Section 3.2 of ARP). One can therefore construct a size-κ least-favourable (LF) test in the finite-sample normal model that rejects whenever $\hat{\eta}$ exceeds the $1 - \kappa$ quantile of $\max_{\gamma \in V(\Sigma)} \gamma'\xi$, where $\xi \sim \mathcal{N}(0,\ \tilde{\Sigma}_n)$. This critical value, which we will denote by $c_{LF,\kappa}$, can easily be calculated by simulation. For $0 < \kappa < \alpha$, the ARP conditional-LF hybrid test is defined to reject if a first-stage, size-κ LF test rejects. If this first-stage test does not reject, then in the second stage the hybrid test conducts a modified version of the size-$((\alpha - \kappa)/(1 - \kappa))$ conditional test that also conditions on the event that the first-stage LF test did not reject. In particular, by similar logic as for the conditional test, we have that

$$\hat{\eta} \,|\, \{\gamma_* \in \hat{V}_n, S_n = s,\ \hat{\eta} \leq c_{LF,\kappa}\} \sim \xi \,|\, \xi \in [v^{lo}, v^{up}_H],$$

where $v^{up}_H = \min\{v^{lo}, c_{LF,\kappa}\}$ (see Section 3.4 of ARP). The second-stage of the hybrid test rejects if $\hat{\eta}$ exceeds the critical value for the size-$((\alpha - \kappa)/(1 - \kappa))$ conditional test that uses v^{up}_H instead of v^{up}. We will denote by $\psi^{C\text{-}LF}_{\kappa,\alpha}(\hat{\beta}_n, A, d, \bar{\theta}, \Sigma_n)$ an indicator for whether the hybrid test rejects at a particular value $\bar{\theta}$, and denote by $\mathcal{C}^{C\text{-}LF}_{\kappa,\alpha,n}(\hat{\beta}_n, \Sigma_n)$ the confidence set that collects the values of $\bar{\theta}$ for which the hybrid test does not reject. As with the conditional test, by construction the hybrid confidence set satisfies that coverage criterion (10) in the finite-sample normal model. In our implementation below, we use $\kappa = \alpha/10$, following ARP.

3.3. *Uniform asymptotic size control*

We now provide conditions under which size control in the finite sample normal model translates to uniform asymptotic size control over a large class of data-generating processes $\mathcal{P}$ under which $\hat{\beta}_n$ is asymptotically normally distributed and Σ_n is replaced with a consistent estimator $\hat{\Sigma}_n$. In particular, we provide sufficient conditions on $\hat{\beta}_n$, $\hat{\Sigma}_n$, and Δ such that the higher-level conditions for size control in ARP are satisfied (Proposition 2 in ARP).

Throughout this section, we fix $\Delta = \{A\delta \leq d\}$ for some A with all non-zero rows, and assume that Δ is non-empty. We consider a class of data-generating processes, indexed by $P \in \mathcal{P}$, under which $\sqrt{n}(\hat{\beta}_n - \beta_P)$ is asymptotically normal, where β_P satisfies the causal decomposition in (3), *i.e.* $\beta_P = \delta_P + L_{post}\tau_{P,post}$ for $\delta_P \in \Delta$ and $\tau_{P,post} \in \mathbb{R}^{\bar{T}}$. The parameter of interest is $\theta_P := l'\tau_{P,post}$, for some fixed $l \neq 0$. Our first assumption imposes uniform asymptotic normality of $\hat{\beta}_n$.

Assumption 2. *Let BL_1 denote the set of Lipschitz functions which are bounded by* 1 *in absolute value and have Lipschitz constant bounded by* 1. *We assume*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{f\in BL_1} \left| \mathbb{E}_P\left[f(\sqrt{n}(\hat{\beta}_n - \beta_P))\right] - \mathbb{E}\left[f(\xi_P)\right]\right| = 0,$$

where $\xi_P \sim \mathcal{N}(0,\ \Sigma_P)$, and $\beta_P = \delta_P + L_{post}\tau_{P,post}$ for $\delta_P \in \Delta$ and $\tau_{P,post} \in \mathbb{R}^{\bar{T}}$.

Convergence in distribution is equivalent to convergence in bounded Lipschitz metric (see Theorem 1.12.4 in van der Vaart and Wellner, 1996), so Assumption 2 formalizes the notion of uniform convergence in distribution of $\sqrt{n}(\hat{\beta}_n - \beta_P)$ to a $\mathcal{N}(0,\ \Sigma_P)$ variable under P.

Our next two assumptions require that the eigenvalues of the asymptotic variance of $\hat{\beta}_n$ be bounded above and away from zero, and that there exists a uniformly consistent estimator for the variance of $\hat{\beta}_n$.

Assumption 3. *Let* **S** *denote the set of matrices with eigenvalues bounded below by $\underline{\lambda} > 0$ and above by $\bar{\lambda} \geq \underline{\lambda}$. For all $P \in \mathcal{P}$, $\Sigma_P \in$* **S**.

Assumption 4. *We have an estimator $\hat{\Sigma}_n$ that is uniformly consistent for Σ_P,*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\|\hat{\Sigma}_n - \Sigma_P\| > \epsilon\right) = 0,$$

for all $\epsilon > 0$.

Finally, we impose some regularity conditions on the matrix A.

Assumption 5. *At least one of the following holds.*

(A) *For $k_1 + k_2 = \dim(\delta)$, the matrix A can be written as TQ, where Q has full row-rank and $T = \begin{pmatrix} I_{k_1} & 0 \\ -I_{k_1} & 0 \\ 0 & I_{k_2} \end{pmatrix}$. (We allow for the case where one of k_1 or k_2 is 0, in which case the zero-dimensional blocks can be ignored).*

(B) *Let $\bar{\gamma}_1, \ldots, \bar{\gamma}_K$ be the elements of $V(I)$. Then for all k, either $\bar{\gamma}_k' A = 0$ or $\inf_{a\geq 0} \inf_{j\neq k} \|(\bar{\gamma}_k - a\bar{\gamma}_j)'A\| > 0$.*

Part (A) of Assumption 5 imposes that the only source of degeneracy in the rows of A is matching inequalities of opposite signs. This is the case for many restrictions of interest, such as $\Delta^{SD}(M)$ and the polyhedra that form $\Delta^{RM}(\bar{M})$. Part (B) provides an alternative, higher-level condition that ensures that for distinct vertices $\bar{\gamma}_k, \bar{\gamma}_j$, the random variables $\bar{\gamma}_j'\tilde{Y}_n$ and $\bar{\gamma}_k'\tilde{Y}_n$

are not perfectly positively correlated with each other. Assumption 5 is used to guarantee that degeneracy in the asymptotic distribution of $\gamma' \tilde{Y}_n$ arises only from known degeneracies in A. We note, however, that Assumption 5 does not rule out settings where the solutions to the bounds of the identified set given in equation (5) and (6) are non-unique or degenerate (*i.e.* where the extreme points for θ occur at "flat faces" of the identified set). If A is full-rank, for example, then Assumption 5(A) holds trivially with $T = I$, and thus the mean of the moments $\tilde{\mu}(\bar{\theta})$ is completely unrestricted.[20]

The assumptions stated above are sufficient for the conditions in Proposition 2 in ARP, which establishes uniform size control for the conditional and hybrid tests.

Proposition 3.1. *Suppose Assumptions 2–5 hold. Then the conditional and LF-hybrid tests uniformly control size. That is, for any $\alpha < 0.5$,*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\psi^C_\alpha\left(\hat{\beta}_n, A, d, \theta_P, \frac{1}{n}\hat{\Sigma}_n\right)\right] \leq \alpha.$$

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\psi^{C\text{-}LF}_{\kappa,\alpha}\left(\hat{\beta}_n, A, d, \theta_P, \frac{1}{n}\hat{\Sigma}_n\right)\right] \leq \alpha.$$

3.4. *Uniform asymptotic consistency*

We next provide conditions under which the conditional and hybrid tests are uniformly asymptotically consistent, in the sense that power against fixed alternatives outside the identified set converges uniformly to 1. To establish uniform consistency of the conditional and hybrid tests, we strengthen Assumptions 2 and 3 as follows.

Assumption 6. *Let $W_n = ((\hat{\beta}_n - \beta_P)', (\mathrm{vec}(\hat{\Sigma}_n) - \mathrm{vec}(\Sigma_P))')'$, where $\mathrm{vec}(\Sigma)$ is the vector of the elements of the matrix Σ. We assume*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{f\in BL_1} \left|\left|\mathbb{E}_P\left[f(\sqrt{n}W_n)\right] - \mathbb{E}\left[f(\xi_P^+)\right]\right|\right| = 0,$$

where $\xi_P^+ \sim \mathcal{N}(0, V_P)$, $V_P = \begin{pmatrix} \Sigma_P & V_{P,\beta\Sigma} \\ V_{P,\Sigma\beta} & V_{P,\Sigma} \end{pmatrix}$.

Assumption 7. *For all $P \in \mathcal{P}$, $\Sigma_P \in \mathbf{S}$ and the matrix V_P defined in Assumption 6 lies in a compact set $\mathbf{V}$. Additionally, $(\Sigma_P - V_{P,\beta\Sigma}V^{\dagger}_{P,\Sigma}V_{P,\Sigma\beta})$ has eigenvalues bounded below by $\tilde{\lambda} > 0$, where $\dagger$ denotes the Moore–Penrose inverse.*

Assumption 6 strengthens Assumption 2 to require that $\hat{\beta}_n$ and $\hat{\Sigma}_n$ have a joint normal asymptotic distribution. Although somewhat more restrictive, event-study estimates are often estimated via OLS, and standard covariance estimators for OLS, including cluster-robust variance estimators, produce asymptotically normal estimates as the number of clusters grows large (Hansen, 2007; Stock and Watson, 2008; Hansen and Lee, 2019). We do not impose that the asymptotic distributions of $\hat{\beta}_n$ and $\hat{\Sigma}_n$ are independent, as would occur in linear models if the linear model is correctly specified. Assumption 7 strengthens Assumption 3 to require that the asymptotic distribution of $\hat{\beta}_n$ is not perfectly asymptotically colinear with $\hat{\Sigma}_n$.

20. Some values of A satisfying Assumption 5 may imply that certain pairs of moments cannot simultaneously be binding. For example, the restriction that $|\delta_1| \leq 1$ can be represented as $\delta_1 \leq 1$ and $-\delta_1 \leq 1$, which satisfies Assumption 5(A), but clearly both moments cannot simultaneously bind. Nevertheless Assumption 5(A) is compatible with "flat faces" even when A is not full rank. For example, if Δ corresponds with the restrictions $|\delta_1| \leq 1$ and $|\delta_2| \leq 1$, then the extreme points for $\theta = \tau_2$ occur at flat faces of the identified set for (τ_1, τ_2).

Under the imposed assumptions, we obtain uniform consistency of the conditional and hybrid tests.

Proposition 3.2. *Suppose Assumptions 4–7 hold. Then for any $x > 0$ and $\alpha < 0.5$,*

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}} \mathbb{E}_P \left[\psi_\alpha^C \left(\hat{\beta}_n, A, d, \theta_P^{ub} + x, \frac{1}{n}\hat{\Sigma}_n \right) \right] = 1$$

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}} \mathbb{E}_P \left[\psi_{\kappa,\alpha}^{C\text{-}LF} \left(\hat{\beta}_n, A, d, \theta_P^{ub} + x, \frac{1}{n}\hat{\Sigma}_n \right) \right] = 1,$$

where $\theta_P^{ub} = \sup \mathcal{S}(\beta_P, \Delta)$ is the upper bound of the identified set. The analogous result holds replacing $\theta_P^{ub} + x$ with $\theta_P^{lb} - x$ for $\theta_P^{lb} = \inf \mathcal{S}(\beta_P, \Delta)$.

3.5. *Optimal local asymptotic power*

We next provide conditions under which the conditional test has optimal local asymptotic power. We first state the conditions and our formal results, and then provide several examples highlighting when the assumptions will and will not hold.

3.5.1. Main results. We begin by defining LICQ. Recall that the upper bound of the identified set is given by

$$\theta^{ub}(\beta, \Delta) = l'\beta_{post} - \left(\min_{\delta} l'\delta_{post}, \text{ s.t. } A\delta \le d, \delta_{pre} = \beta_{pre} \right).$$

Since $\delta_{post} = \beta_{post} - \tau_{post}$, we can re-write the upper bound as a maximization over τ_{post},

$$\theta^{ub}(\beta, \Delta) = \max_{\tau_{post}} l'\tau_{post}, \text{ s.t. } -A_{(\cdot,post)}\tau_{post} \le d - A\beta, \tag{16}$$

where $A_{(\cdot,post)}$ contains the columns of A corresponding with δ_{post}. Let τ_{post}^* denote a solution to the optimization for $\theta^{ub}(\beta, \Delta)$ in (16), and let B^* denote the indices of the binding constraints, so that $-A_{(B^*,post)}\tau_{post}^* = d_{B^*} - A_{(B^*,\cdot)}\beta$ and $-A_{(-B^*,post)}\tau_{post}^* < d_{-B^*} - A_{(-B^*,\cdot)}\beta$.

Definition 2 (LICQ)**.** We say that LICQ holds in direction l if there exists a solution τ_{post}^* to (16) such that the gradient of the binding constraints with respect to τ_{post}, $-A_{(B^*,post)}$, has full row rank.[21] We define LICQ in the direction $-l$ analogously for the optimization that replaces max with min in (16).

For $\epsilon > 0$, we define $\mathcal{P}_\epsilon$ to be the set of distributions $P \in \mathcal{P}$ such that LICQ holds in the direction l and the non-binding constraints are slack by at least ϵ, *i.e.* $-A_{(-B^*,post)}\tau_{post}^* < d - A\beta_P - \epsilon$.

Our next result states that for $P \in \mathcal{P}_\epsilon$, the local power of the conditional test converges to the power envelope for tests that control size in the finite sample normal model. To state this result formally, we define $\mathcal{I}_\alpha(\Delta, \Sigma_n)$ to be the collection of confidence sets that control size in the finite sample normal model, *i.e.* confidence sets satisfying (10).

21. The definition of LICQ in Kaido *et al.* (2021) would require that this condition holds *for all* solutions τ_{post}^*. For our results, however, it is sufficient for the condition to hold for *some* solution τ_{post}^*.

Proposition 3.3. *Suppose Assumptions 2–4 hold. Let* $\theta_P^{ub} = \sup \mathcal{S}(\beta_P, \Delta)$. *Then for any* $\epsilon > 0$, $x > 0$, *and* $\alpha < 0.5$,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_\epsilon} \left| \mathbb{E}_P \left[\psi_\alpha^C \left(\hat{\beta}_n, A, d, \theta_P^{ub} + \frac{1}{\sqrt{n}} x, \frac{1}{n} \hat{\Sigma}_n \right) \right] - \rho_\alpha^*(P, x) \right| = 0,$$

where

$$\rho_\alpha^*(P, x) = \lim_{n\to\infty} \sup_{\mathcal{C}_{\alpha,n} \in \mathcal{I}_\alpha\left(\Delta, \frac{1}{n}\Sigma_P\right)} \mathbb{P}_{\hat{\beta}_n \sim N\left(\beta_P, \frac{1}{n}\Sigma_P\right)} \left(\left(\theta_P^{ub} + \frac{1}{\sqrt{n}} x \right) \notin \mathcal{C}_{\alpha,n} \right)$$

is the optimal local asymptotic power of a size-α *test in the finite sample normal model. An analogous result holds for the lower bound under the class of distributions where LICQ holds in direction* $-l$.

Since the LF-hybrid test rejects whenever the conditional test with size $(\alpha-\kappa)/(1-\kappa)$ rejects, it is immediate that the local asymptotic power of the LF-hybrid test is at least as good as the power of the optimal size-$((\alpha-\kappa)/(1-\kappa))$ test.

Corollary 3.1. *Under the conditions of Proposition 3.3,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_\epsilon} \left(\mathbb{E}_P \left[\psi_{\kappa,\alpha}^{C\text{-}LF} \left(\hat{\beta}_n, A, d, \theta_P^{ub} + \frac{1}{\sqrt{n}} x, \frac{1}{n} \hat{\Sigma}_n \right) \right] - \rho_{(\alpha-\kappa)/(1-\kappa)}^*(P, x) \right) \geq 0.$$

We emphasize that Proposition 3.3 and Corollary 3.1 are new, and exploit structure in our context not contained in the more general setting considered in ARP.

3.5.2. Discussion and examples. As discussed in Kaido *et al.* (2021), LICQ and related constraint qualifications have been used frequently in the partial identification literature. Intuitively, LICQ ensures that the bounds of the identified set are differentiable with respect to the means of the moments $(\tilde{\mu}(\bar{\theta}))$, and thus avoids challenges related to estimation and inference for non-differentiable parameters (Hirano and Porter, 2012). Uniform LICQ conditions have been invoked recently by Gafarov (2019) and Cho and Russell (2019), and a related Slater constraint qualification is used in Kaido and Santos (2014). One important distinction between our results and previous results using LICQ is that we do not require LICQ for our size control results (Proposition 3.1). Thus, our tests control size even when LICQ fails (and so the bounds may be non-differentiable), but Proposition 3.3 shows that this does not come at the cost of power asymptotically when indeed LICQ holds.[22]

Figure 1 provides geometric intuition for when LICQ will and will not hold in the case where $\bar{T} = 2$ and the target parameter is the average of the post-treatment effects, $\theta = \frac{1}{2}(\tau_1 + \tau_2)$. In panel (a), there is a unique τ_{post}^* (coloured in red) at which two linearly independent moments bind, so LICQ is satisfied. LICQ is likewise satisfied in panel (b), where the optimal τ_{post}^* is not unique (a so-called "flat-face" problem). This is because at the indicated values τ_{post}^* (coloured in red), there is either one or two linearly independent binding moments. A failure of LICQ is shown in panel (c). In this example, there are three binding moments at τ_{post}^* (coloured in red), so the binding constraints cannot be linearly independent in $\mathbb{R}^2$. Such a situation may arise when there are both smoothness restrictions and sign or shape restrictions that are simultaneously binding at the boundary of the identified set.

22. We view this result as loosely parallel to results in the weak identification literature showing that certain procedures control size under weak identification but are efficient under strong identification (*e.g.* Moreira, 2003).

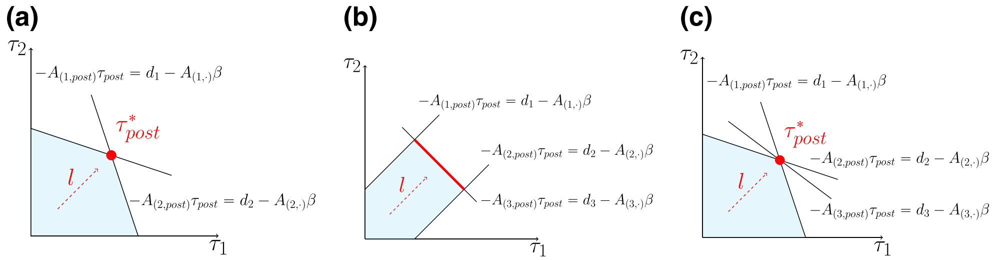

FIGURE 1

Diagram illustrating when LICQ (Assumption 2) will and will not hold in the case where $\bar{T} = 2$.

Note: In each panel, we assume that the rows associated with binding moments are ordered first in the matrix A for ease of notation. The blue shading denotes the identified set for (τ_1, τ_2) and the dashed red arrow points in the direction $l = (\frac{1}{2}, \frac{1}{2})'$. In panel (a), LICQ is satisfied since there is a unique τ^*_{post} (coloured in red) at which two linearly independent moments are binding. In panel (b), even though τ^*_{post} is not unique, LICQ is satisfied as there is either one or two linearly independent binding moments at the values of τ^*_{post} coloured in red. In panel (c), there are three binding moments at τ^*_{post} (coloured in red), and so LICQ is violated.

In the three period DiD model (where there are no nuisance parameters, since $\bar{T} = 1$), LICQ is satisfied when the bounds of the identified set are each determined by one moment. This holds everywhere for $\Delta^{\text{SD}}(M)$ when $M > 0$. It holds almost everywhere for $\Delta^{SDPB}(M)$ when $M > 0$, although it fails when both the sign restrictions and smoothness restrictions are simultaneously binding. (For LICQ to hold with non-binding moments slack by at least ϵ, *i.e.* $P \in \mathcal{P}_\epsilon$, δ_P must not be local to a point at which LICQ fails.) When $M = 0$, both the upper and lower bounds for $\Delta^{\text{SD}}(M)$ and $\Delta^{SDPB}(M)$ are binding, so LICQ fails.

More generally, the result in Proposition 3.3 is under the asymptotic regime where the sampling variation grows small relative to the length of the identified set, and thus the binding and non-binding moments are "far" apart relative to sampling variation. Importantly, it can be shown that the LICQ condition rules out settings where θ is point identified. Thus, the asymptotics considered in Proposition 3.3 may not provide a good approximation to the finite-sample performance of the conditional test in settings where θ is point-identified, or when the length of the identified set is "small" relative to sampling variation.

We are not aware of results analogous to Proposition 3.3 for any test that controls size in the finite-sample normal model. Kaido and Santos (2014) provide an efficiency result under a related Slater constraint qualification condition, but their test does not control size when the constraint qualification fails. It is worth highlighting that if LICQ holds for a particular set of moments, then it also holds if one adds moments that are slack at the optimal τ^*_{post}. Proposition 3.3 thus requires that the asymptotic power of the test is not affected by the inclusion of slack moments. The only other non-trivial tests that we are aware of that control size in the finite-sample normal model and have this form of insensitivity to slack moments are the tests proposed by Cox and Shi (2022). An interesting open question is whether the tests proposed by Cox and Shi (2022) also converge to the power envelope under LICQ.[23]

3.5.3. Extensions. Proposition 3.3 is stated for the case when Δ is a single polyhedron. An immediate corollary, however, is that when $\Delta = \bigcup_{k=1}^{K} \Delta_k$, the conditional test based on the union of confidence sets has optimal local asymptotic power when the Δ_k that determines the identified set bounds is unique and satisfies the conditions of Proposition 3.3. This implies,

23. Extending the results to the Cox and Shi tests is non-trivial given that they use a different test statistic and construct critical values in a different way.

for example, that when $\Delta = \Delta^{RM}(\bar{M})$, the power of the conditional test converges to the power envelope when there is a unique (non-zero) pre-treatment maximum violation, *i.e.* when $\max_{s<0} |\delta_{s+1} - \delta_s| > 0$ and has a unique solution.[24] Likewise, the conditional test has optimal local asymptotic power for $\Delta^{SDRM}(\bar{M})$ when there is a unique maximum non-linearity in the pre-treatment period. Intuitively, this is because the upper bound of the identified set is determined by a single Δ_{k^*} satisfying LICQ, so the conditional test for this Δ_{k^*} has optimal local asymptotic power, whereas our consistency results imply that the tests for the remaining Δ_k that do not determine the identified set bound reject with probability approaching 1. See Corollary 4.1 in the working paper version of this paper for a formal derivation (Rambachan and Roth, 2021).

Proposition 3.3 shows that under LICQ the local asymptotic power of the conditional test converges to the power envelope for tests controlling size in the finite-sample normal model. In the working paper version of this paper, we showed that the power envelope from the finite-sample normal model corresponds with the power-envelope among tests that control size asymptotically and have certain invariance properties using results in Müller (2011) (Proposition E.4 in Rambachan and Roth, 2021).

4. INFERENCE USING FIXED LENGTH CONFIDENCE INTERVALS

We next consider FLCIs based on affine estimators. While the conditional and hybrid confidence sets offer attractive asymptotic power guarantees under asymptotics in which sampling variation grows small relative to the length of the identified set, FLCIs offer finite-sample power guarantees (in the normal model) for certain classes Δ of interest. In certain special cases, FLCIs may thus outperform the ARP tests when sampling variation is large relative to the length of the identified set. For brevity of exposition, we focus on the properties of FLCIs in the case where the finite-sample normal approximation (9) holds exactly with Σ_n known; Armstrong and Kolesár (2020b) provide uniform asymptotic results for FLCIs under conditions similar to those in Section 3.3.

4.1. *Constructing FLCIs*

Following Donoho (1994) and Armstrong and Kolesár (2018, 2020a), we consider FLCIs based on an affine estimator for θ, denoted by $\mathcal{C}_{\alpha,n}(a, v, \chi) := (a + v'\hat{\beta}_n) \pm \chi$, where a and χ are scalars and $v \in \mathbb{R}^{\underline{T}+\bar{T}}$. We minimize the half-length of the confidence interval, χ, subject to the constraint that $\mathcal{C}_{\alpha,n}(a, v, \chi)$ satisfies the coverage requirement (10) in the finite-sample normal model.

Observe that if $\hat{\beta}_n \sim \mathcal{N}(\beta, \Sigma_n)$, then $a + v'\hat{\beta}_n \sim \mathcal{N}\left(a + v'\beta, v'\Sigma_n v\right)$, 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 affine estimator's bias for θ. Observe further that $\theta \in \mathcal{C}_{\alpha,n}(a, v, \chi)$ if and only if $|a + v'\hat{\beta}_n - \theta| \leq \chi$. For fixed values a and v, the smallest value of χ that satisfies (10) is therefore the $1-\alpha$ quantile of the $|\mathcal{N}\left(\bar{b}, v'\Sigma_n v\right)|$ distribution, where $\bar{b}$ is the affine estimator's worst-case bias

$$\bar{b}(a, v) := \sup_{\delta \in \Delta, \tau_{post} \in \mathbb{R}^{\bar{T}}} \left| a + v'\left(\delta + L_{post}\tau_{post}\right) - l'\tau_{post} \right|. \tag{17}$$

24. For this convergence to hold uniformly, the non-binding moments must be slack by ϵ, so we would need that $\max_{s<0} |\delta_{s+1} - \delta_s|$ is at least ϵ greater than the second largest difference.

Let $cv_\alpha(t)$ denote the $1-\alpha$ quantile of the folded normal distribution $|\mathcal{N}(t,\, 1)|$.[25] For fixed a and v, the smallest value of χ satisfying the coverage requirement (10) is thus

$$\chi_n(a, v; \alpha) = \sigma_{v,n} \cdot cv_\alpha(\bar{b}(a, v)/\sigma_{v,n}), \tag{18}$$

where $\sigma_{v,n} := \sqrt{v'\Sigma_n v}$. The optimal (*i.e.* minimum-length) FLCI is constructed by choosing the values of a and v to minimize (18). 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 Kolesár, 2018, 2020a). We denote the $1-\alpha$ level, optimal FLCI by $\mathcal{C}^{FLCI}_{\alpha,n}(\hat{\beta}_n, \Sigma_n) := (a_n + v_n'\hat{\beta}_n) \pm \chi_n$, where $\chi_n := \inf_{a,v} \chi_n(a, v; \alpha)$ and a_n, v_n are the optimal values in the minimization.

4.1.1. Example: $\boldsymbol{\Delta^{SD}(M)}$. Suppose $\theta = \tau_1$. For $\Delta^{SD}(M)$, the affine estimator used by the optimal FLCI takes the form $a + v'\hat{\beta}_n = \hat{\beta}_{n,1} - \sum_{s=-T+1}^{0} w_s(\hat{\beta}_{n,s} - \hat{\beta}_{n,s-1})$, where the weights w_s sum to one (but may be negative). This estimator adjusts the event-study coefficient for $t=1$ by an estimate of the differential trend between $t=0$ and $t=1$ formed by taking a weighted average of the differential trends in periods prior to treatment. The worst-case bias will be smaller if more weight is placed on pre-treatment periods closer to the treatment date, but it may reduce variance to place more weight on earlier pre-periods. The weights w_s are optimally chosen to balance this trade-off.

4.2. *Finite-sample near optimality*

In particular cases of interest, such as when $\Delta = \Delta^{SD}(M)$, the optimal FLCIs introduced above have near-optimal expected length in the finite-sample normal model. The following result, which is an immediate consequence of results in Armstrong and Kolesár (2018, 2020a), 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 8. *Assume (i) Δ is convex and centrosymmetric (i.e. $\tilde{\delta} \in \Delta$ implies $-\tilde{\delta} \in \Delta$), and ii) $\delta \in \Delta$ is such that $(\tilde{\delta} - \delta) \in \Delta$ for all $\tilde{\delta} \in \Delta$.*

Proposition 4.1. *Suppose δ and Δ satisfy Assumption 8. Let $\mathcal{I}_\alpha(\Delta, \Sigma_n)$ denote the class of confidence sets that satisfy the coverage criterion (10) at the $1-\alpha$ level. Then, for any τ with $\tau_{pre} = 0$ and Σ_n positive definite,*

$$\frac{\inf_{\mathcal{C}_{\alpha,n} \in \mathcal{I}_\alpha(\Delta, \Sigma_n)} \mathbb{E}_{\hat{\beta}_n \sim \mathcal{N}(\delta+\tau,\, \Sigma_n)}\left[\lambda(\mathcal{C}_{\alpha,n})\right]}{2\chi_n} \geq \frac{z_{1-\alpha}(1-\alpha) - \tilde{z}_\alpha \Phi(\tilde{z}_\alpha) + \phi(z_{1-\alpha}) - \phi(\tilde{z}_\alpha)}{z_{1-\alpha/2}},$$

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 8 is satisfied for $\Delta^{SD}(M)$ but not for our other ongoing examples. For example, $\Delta^{SDPB}(M)$ is convex but not centrosymmetric, and $\Delta^{RM}(\bar{M})$ is neither convex nor centrosymmetric. Part ii) of Assumption 8 is satisfied whenever parallel trends holds in both the pre-treatment and post-treatment periods ($\delta = 0$) and whenever δ is a linear trend for the case of $\Delta^{SD}(M)$.

FLCIs thus offer attractive guarantees for the case of $\Delta^{SD}(M)$. When $\alpha = 0.05$, the lower bound in Proposition 4.1 evaluates to 0.72, meaning that the expected length of the shortest

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

possible confidence set that satisfies the coverage requirement (10) is at most 28% shorter than the length of the optimal FLCI when the conditions of the proposition hold.

4.3. *(In)Consistency of FLCIs*

As discussed above, these finite-sample guarantees do not apply for several types of restrictions Δ of importance, including those that construct bounds using the maximum pre-treatment violation or incorporate sign and shape restrictions. We now show that the FLCIs can perform poorly under such restrictions. We first provide two illustrative examples, and then state a formal inconsistency result.

4.3.1. Example: $\boldsymbol{\Delta^{SDPB}(M)}$ and $\boldsymbol{\Delta^{SDI}(M)}$. Suppose $\theta = \tau_1$. It can be shown that the worst-case bias of an affine estimator over $\Delta^{SDPB}(M)$ or $\Delta^{SDI}(M)$ is the same as the worst-case bias for that estimator over $\Delta^{\mathrm{SD}}(M)$.[26] Since the construction of the optimal FLCI depends only on the worst-case bias and variance of the affine estimator, it follows that the optimal FLCI constructed using $\Delta^{SDPB}(M)$ or $\Delta^{SDI}(M)$ is the same as the one constructed using $\Delta^{\mathrm{SD}}(M)$. Therefore, the optimal FLCI does not adapt to additional sign or monotonicity restrictions.

4.3.2. Example: $\boldsymbol{\Delta^{RM}(\bar{M})}$. Suppose $\theta = \tau_1$. If $\Delta = \Delta^{RM}(\bar{M})$ and $\bar{M} > 0$, then all affine estimators for τ_1 have infinite worst-case bias, since $\delta \in \Delta^{RM}(\bar{M})$ can have $|\delta_1|$ arbitrarily large if $|\delta_{-1}|$ is also sufficiently large. Thus, the only valid FLCI is the entire real line.

We next provide a formal result on the (in)consistency of the FLCIs. Specifically, we will show that even as the sampling variation Σ_n converges to 0, the optimal FLCI will include fixed points outside of the identified set with positive probability unless certain special conditions are met.[27] Recall from Lemma 2.1 that the identified set $\mathcal{S}(\beta, \Delta)$ is an interval when Δ is convex, with length equal to $\theta^{ub}(\beta, \Delta) - \theta^{lb}(\beta, \Delta) = b^{max}(\beta_{pre}, \Delta) - b^{min}(\beta_{pre}, \Delta)$. Since the length of the identified set only depends on β_{pre} and Δ, denote it by $LID(\beta_{pre}, \Delta)$. Our next result shows that $\mathcal{C}^{FLCI}_{\alpha,n}(\hat{\beta}_n, \Sigma_n)$ is consistent if and only if $LID(\beta_{pre}, \Delta)$ is its maximum possible value, provided that the identified set is not the entire real line (in which case any procedure is trivially consistent).

Assumption 9 (Identified set maximal length and finite)**.** *Suppose $\delta \in \Delta$ is such that $LID(\delta_{pre}, \Delta) = \sup_{\tilde{\delta}_{pre} \in \Delta_{pre}} LID(\tilde{\delta}_{pre}, \Delta) < \infty$, where $\Delta_{pre} = \{\delta_{pre} \in \mathbb{R}^{\underline{T}} : \exists \delta_{post}$ s.t. $(\delta'_{pre}, \delta'_{post})' \in \Delta\}$ is the set of possible values for δ_{pre}.*

Proposition 4.2. *Suppose Δ is convex and $\alpha < 0.5$. Fix $\delta \in \Delta$ and τ with $\tau_{pre} = 0$, and suppose $\mathcal{S}(\delta + \tau, \Delta) \neq \mathbb{R}$. Then (δ, Δ) satisfy Assumption 9 if and only if $\mathcal{C}^{FLCI}_{\alpha,n}(\hat{\beta}_n, \Sigma_n)$ is consistent, meaning that for $\Sigma_n = \Sigma^*/n$,*

$$\lim_{n \to \infty} \mathbb{P}_{\hat{\beta}_n \sim \mathcal{N}(\delta+\tau,\, \Sigma_n)} \left(\theta^{out} \in \mathcal{C}^{FLCI}_{\alpha,n}(\hat{\beta}_n, \Sigma_n)\right) = 0 \quad \text{for all } \theta^{out} \notin \mathcal{S}(\delta + \tau, \Delta).$$

26. Suppose the vector $\bar{\delta}$ maximizes the bias for an affine estimator (a, v) over $\Delta^{\mathrm{SD}}(M)$. The vector that adds a constant slope to $\bar{\delta}$, say $\tilde{\delta}_c = \bar{\delta} + c \cdot (-\underline{T}, \ldots, \bar{T})'$, also lies in $\Delta^{\mathrm{SD}}(M)$, and for c sufficiently large, $\tilde{\delta}_c$ will lie in $\Delta^{SDPB}(M)$. Moreover, the worse-case bias will be the same for δ and $\tilde{\delta}_c$, since if (a, v) has finite worst-case bias it must subtract out a weighted average of the pre-treatment slopes.

27. For ease of exposition, we present a result using "small-Σ" asymptotics in the normal model, as in *e.g.* Kadane (1971) and Moreira and Ridder (2019).

Thus, if Assumption 9 fails, then $\mathcal{C}^{FLCI}_{\alpha,n}(\hat{\beta}_n, \Sigma_n)$ is inconsistent in the strong sense that it includes fixed points outside of the identified set with non-vanishing probability. It follows that there will be some $\delta \in \Delta$ such that the FLCI is inconsistent under δ unless the identified set is always the same length. Proposition 4.2 is new, and may be relevant for other settings in which FLCIs are used.

The intuition for the possible inconsistency of FLCIs is as follows: to ensure that an FLCI satisfies the coverage requirement (10), its length must be at least $\sup_{\tilde{\delta}_{pre}\in\Delta_{pre}} LID(\tilde{\delta}_{pre}, \Delta)$. However, this implies that if in fact $LID(\delta_{pre}, \Delta) < \sup_{\tilde{\delta}_{pre}\in\Delta_{pre}} LID(\tilde{\delta}_{pre}, \Delta)$, then the FLCI is strictly longer than the length of the identified set, regardless of the value of Σ_n, and thus some points outside of the identified set must be covered with non-vanishing probability. This reflects the fact that FLCIs are by construction *fixed length*, and thus their length does not adapt to information in the data about the length of the identified set. By contrast, the length of the conditional/hybrid confidence sets can depend on $\hat{\beta}_{pre}$ and thus "adapts" to the length of the identified set.

In the three-period DiD example, Assumption 9 holds everywhere for $\Delta^{\text{SD}}(M)$ (since the identified set is always the same length, $2M$), for values of δ where the sign restrictions do not bind for $\Delta^{SDPB}(M)$, and nowhere for the polyhedra that form $\Delta^{RM}(\bar{M})$. The restrictiveness of Assumption 9 thus depends greatly on Δ.

The results in this section establish that when certain conditions on Δ are satisfied, optimal FLCIs are consistent and have desirable finite-sample guarantees in terms of expected length. FLCIs are thus attractive for our baseline smoothness class $\Delta^{\text{SD}}(M)$, since they are guaranteed to be consistent and offer attractive finite-sample guarantees. Our inconsistency result shows, however, that FLCIs may perform poorly for other choices of Δ that may be of interest in empirical applications, such as those that construct bounds using a pre-treatment maximum or incorporate sign and monotonicity restrictions.

5. SIMULATION STUDY

In this section, we conduct a simulation study to investigate the performance of the discussed confidence sets across a range of relevant data-generating processes. We find good size control for all of the procedures, and therefore focus in the main text on a comparison of power to provide concrete recommendations on the best approach in practice. In the supplementary material, we present results on size control and other additional simulation results.

5.1. *Simulation design*

Our simulations are calibrated using the estimated covariance matrix from the 12 recently-published papers surveyed in Roth (2022). For any given paper in the survey, we 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. For a chosen mean vector β, we simulate event-study coefficients $\hat{\beta}_s$ from a normal model, $\hat{\beta}_s \sim \mathcal{N}(\beta, \hat{\Sigma})$.[28] In simulation s, we construct nominal 95% confidence sets for the parameter of interest θ using the pair $(\hat{\beta}_s, \hat{\Sigma})$ for each proposed procedure. The parameter of interest is the causal effect in the first post-treatment period ($\theta = \tau_1$); in the supplementary material, we present simulation results in which the parameter of interest

28. We focus on the normal simulations in the main text since it allows for a tractable computation of the optimal excess length of procedures that control size. In the supplementary material, we show that our procedures perform similarly in simulations based on the empirical distribution in the original paper.

TABLE 1
Summary of expected properties for each simulation design

	Parallel Trends		Pulse Pre-Trend	
	$\Delta^{\text{SD}}(M)$	$\Delta^{SDPB}(M)$	$\Delta^{SDRM}(\bar{M})$	$\Delta^{RM}(\bar{M})$
Conditional and Hybrid				
Consistent	✓	✓	✓	✓
Asymptotically (near-)optimal	✓	✓	✓	×
FLCI				
Consistent	✓	×	×	×
Finite-sample near-optimal	✓	×	×	×

is the average causal effect in the post-treatment periods ($\theta = \bar{\tau}_{post}$), with qualitatively similar results.

For a given choice of Δ, we compute the identified set $\mathcal{S}(\beta, \Delta)$ using the formulas provided in Lemma 2.1 and calculate the expected excess length for each of the proposed confidence sets. We benchmark the expected excess length of our proposed confidence sets relative to an efficiency bound for confidence sets that satisfy the uniform coverage requirement.[29] We report the efficiency ratio of each procedure, which is defined as the ratio of the optimal benchmark relative to the average excess length for the procedure. All results are calculated over 1,000 simulations per paper.

We consider four choices of Δ to highlight the performance of our proposed confidence sets across a range of conditions: $\Delta^{\text{SD}}(M)$, $\Delta^{SDPB}(M)$, $\Delta^{RM}(\bar{M})$, and $\Delta^{SDRM}(\bar{M})$. We consider simulations under the assumption of zero treatment effects, so that $\tau = 0$ and thus $\beta = \delta$. We consider two forms for δ. First, we consider the baseline case of parallel trends ($\delta = 0$). Second, we consider a "pulse" pre-trend in which δ_{-1} is non-zero and the remaining elements of δ are zero. Such a pre-trend might arise in practice if there are confounding policy changes or other events close to the time of treatment. These different choices of δ allow us to highlight the relative strengths of the proposed inference procedures. For example, FLCIs have near-optimal expected length when $\delta = 0$ and $\Delta = \Delta^{\text{SD}}(M)$, whereas the conditional test has optimal local asymptotic power under the pulse design when $\Delta = \Delta^{SDPB}(M)$. Table 1 summarizes which of our theoretical results hold for each of the simulation designs when M and $\bar{M}$ are non-zero.

In practice, we find that for $\Delta^{\text{SD}}(M)$ and $\Delta^{SDPB}(M)$, the results depend on M but are qualitatively similar across values of δ. By contrast, for $\Delta^{SDRM}(\bar{M})$ and $\Delta^{RM}(\bar{M})$, the choice of δ is more important than the choice of $\bar{M}$. Therefore, to highlight the most important dimensions for each of the simulation designs, in the main text of the paper we report results for $\Delta^{\text{SD}}(M)$ and $\Delta^{SDPB}(M)$ under different values of M and $\delta = 0$ (parallel trends), whereas for $\Delta^{RM}(\bar{M})$ and $\Delta^{SDRM}(\bar{M})$ we vary the magnitude of the pre-treatment pulse δ_{-1}, holding $\bar{M} = 1$ constant. In the supplementary materials, we report results for additional choices of these parameters.

29. For choices of Δ that are convex (*e.g.* $\Delta^{\text{SD}}(M)$ and $\Delta^{SDPB}(M)$), we benchmark the expected excess length of our proposed confidence sets against a sharp optimal bound over confidence sets that satisfy the finite-sample coverage requirement (10). This optimal bound is provided in the supplementary materials, and follows as a corollary from results in Armstrong and Kolesár (2018) on the optimal expected length of a confidence set satisfying the uniform coverage requirement (10). For choices of Δ that can be written as the union of convex sets (*e.g.* $\Delta^{RM}(\bar{M})$ and $\Delta^{SDRM}(\bar{M})$), we compare the expected excess length of our proposed confidence sets against the maximal optimal bound over each set in the union, which is a potentially non-sharp bound for any confidence set with correct coverage.

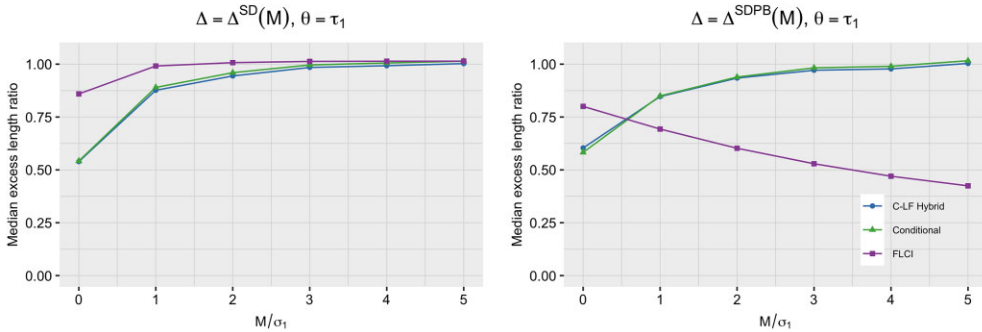

FIGURE 2

Simulation results for $\Delta^{\mathrm{SD}}(M)$ and $\Delta^{SDPB}(M)$: Median efficiency ratios for proposed procedures.

Note: Median efficiency ratios for our proposed confidence sets over $\Delta^{\mathrm{SD}}(M)$ and $\Delta^{SDPB}(M)$ under the assumption of parallel trends and zero treatment effects (*i.e.* $\beta=0$). The efficiency ratio for a procedure is defined as the efficiency bound divided by the procedure's expected excess length. The results for the FLCI are plotted in purple, conditional-LF ("C-LF Hybrid") hybrid in blue, and conditional confidence set in green. Results are averaged over 1,000 simulations for each of the 12 papers surveyed, and the median across papers is reported here.

We report results for three methods for constructing confidence sets: FLCIs, conditional confidence sets, and conditional-least favourable hybrid confidence sets.[30] For $\Delta^{RM}(\bar{M})$ and $\Delta^{SDRM}(\bar{M})$, we omit results for the FLCI since the FLCIs have infinite length.

5.2. *Simulation results*

To compare results easily across the 12 papers in the simulation study, we normalize the units of δ_{-1} and M by the standard deviation of $\hat{\beta}_1$ (denoted σ_1). Large normalized values of M or δ_{-1} correspond with the case where the identified set is large relative to sampling variation, mimicking our asymptotic power results in which sampling variation grows small relative to the identified set. In the graphs below, we report the median value of excess length efficiency across the papers in the survey. The normalization described above implies that the units of the x-axis correspond with the worst-case bias of the naive estimator $\hat{\beta}_1$ divided by its standard error.[31]

5.2.1. Results for $\Delta^{\mathrm{SD}}(M)$. The left panel of Figure 2 plots the efficiency ratio for each procedure as a function of M/σ_1 when $\Delta=\Delta^{\mathrm{SD}}(M)$. All procedures perform well as M/σ_1 grows large with efficiency ratios approaching 1, illustrating our asymptotic (near-)optimality results for this design. However, the FLCIs perform best 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 8 holds. Although the conditional and hybrid confidence sets have efficiency approaching the optimal bound for M/σ_1 large, their efficiency is only about 50% when $M/\sigma_1=0$, in which case θ is point identified and thus LICQ does not hold. The conditional and hybrid confidence sets perform similarly.

30. For the conditional-least favourable hybrid confidence sets, we use a first-stage least-favourable test of size $\kappa=\alpha/10$, following ARP and Romano et al. (2014).

31. For $\hat{\beta}_1$ normally distributed, the worst-case coverage of a conventional 95% confidence interval as a function of the normalized worst-case bias b is $\Phi(1.96+b)-\Phi(-1.96+b)$, which is 0.95 for $b=0$, 0.83 for $b=1$, 0.48 for $b=2$, etc.

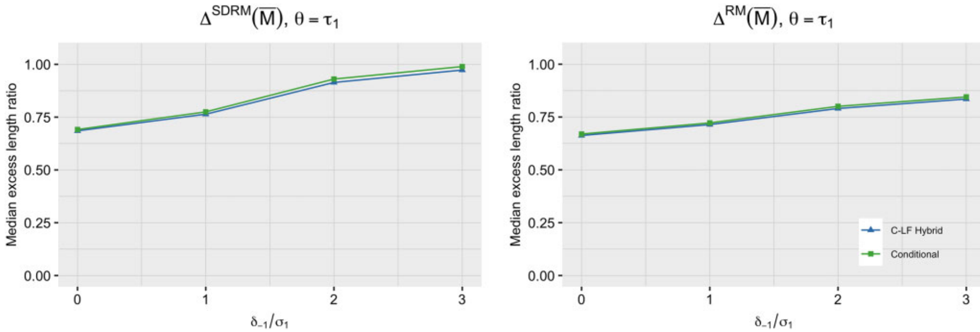

FIGURE 3

Simulation results for $\Delta^{SDRM}(\bar{M})$ and $\Delta^{RM}(\bar{M})$: Median efficiency ratios for proposed procedures.

Note: Median efficiency ratios for our proposed confidence sets over $\Delta^{SDRM}(\bar{M})$ and $\Delta^{RM}(\bar{M})$ with $\bar{M} = 1$ under the assumption of zero treatment effects and a "pulse" pre-trend (*i.e.* $\beta_{-1} = \delta_{-1}$ and $\beta_t = 0$ for all $t \neq -1$). The efficiency ratio for a procedure is defined as the efficiency bound divided by the procedure's expected excess length. The results for the conditional-least favourable ("C-LF") hybrid are plotted in blue, and conditional confidence set in green. Results are averaged over 1,000 simulations for each of the 12 papers surveyed, and the median across papers is reported here.

5.2.2. Results for $\boldsymbol{\Delta^{SDPB}(M)}$. The right panel of Figure 2 plots the efficiency ratio for each procedure as a function of M/σ_1 when $\Delta = \Delta^{SDPB}(M)$. The efficiency ratios for the conditional and hybrid confidence sets are again (near-)optimal as M/σ_1 grows large, highlighting our asymptotic (near-)optimality results for these procedures in this simulation design. By contrast, the efficiency ratios for the FLCIs steadily decrease as M/σ_1 increases, reflecting that the FLCIs are not consistent in this simulation design when $M > 0$. The conditional-LF hybrid confidence sets slightly improve efficiency relative to the conditional confidence sets when M/σ_1 is small and retain near-optimal performance as M/σ_1 grows large.

5.2.3. Results for $\boldsymbol{\Delta^{SDRM}(\bar{M})}$. The left panel of Figure 3 plots the efficiency ratios for the conditional and conditional-least favourable hybrid confidence sets as a function of δ_{-1}/σ_1 when $\Delta = \Delta^{SDRM}(\bar{M})$. We omit results for the optimal FLCI since the optimal FLCI has infinite length for this design. Both procedures perform well as δ_{-1}/σ_1 grows large with efficiency ratios approaching 1, illustrating our asymptotic (near-) optimality result for this design. Both procedures also have similar power curves, with slightly higher power for the conditional.

5.2.4. Results for $\boldsymbol{\Delta^{RM}(\bar{M})}$. The right panel of Figure 3 plots the efficiency ratio for the conditional and conditional-least favourable hybrid confidence sets as a function of δ_{-1}/σ_1 when $\Delta = \Delta^{RM}(\bar{M})$. We again omit results for the optimal FLCI since the optimal FLCI has infinite length for this design. The conditions for our asymptotic (near-) optimality result for unions of convex sets do not hold in this simulation design (as the maximum pre-period violation is not unique). Nonetheless, we find that the efficiency ratios for the conditional and hybrid confidence sets approach about 83% when δ_{-1}/σ_1 grow large. We find it somewhat encouraging that these procedures can have excess length within 17% of the optimum even in cases where LICQ fails. Once again, we also find that the conditional and conditional-least favourable hybrid have similar power.

5.2.5. Takeaways from simulations. Two clear patterns emerge from our simulations. First, the conditional and hybrid confidence sets perform well across a wide range of specifications, with particularly good power when the length of the identified set is large relative to sampling variation. Second, the FLCIs have the best performance for $\Delta^{\mathrm{SD}}(M)$, particularly when

M is small, which aligns with the finite-sample near-optimality results in Section 4. However, FLCIs can perform quite poorly for other classes of Δ.

Overall, we therefore recommend to use the conditional-LF hybrid confidence sets for generic forms of Δ, and optimal FLCIs for the special case of $\Delta^{\mathrm{SD}}(M)$ (or other special cases where the consistency/finite-sample near-optimality of FLCIs is guaranteed). Although the conditional and hybrid approaches perform similarly in our simulations, we recommend the hybrid approach in general based on the guidance provided in ARP. We implement these recommendations in our applications in the next section.

6. PRACTICAL GUIDANCE AND EMPIRICAL ILLUSTRATIONS

6.1. *Practical guidance*

We recommend that researchers use our methods to construct robust confidence intervals under restrictions on the possible violations of parallel trends Δ that are motivated by domain knowledge in their empirical setting. We also suggest that researchers report sensitivity analyses to illustrate the sensitivity of their causal conclusions to alternative assumptions on the possible violations of parallel trends.

6.1.1. Choice of $\boldsymbol{\Delta}$. The choice of Δ should be motivated by economic knowledge about the types of possible confounding factors that could produce non-parallel trends. We now provide some guidance on how the choice of Δ can be motivated by domain knowledge, highlighting some cases where our leading examples, $\Delta^{RM}(\bar{M})$ and $\Delta^{\mathrm{SD}}(M)$, would be sensible choices.

In some empirical settings, researchers may be concerned about differential economic shocks to the treated and control groups that generate violations of parallel trends. If the researcher believes that the magnitude of these differential shocks in the post-treatment period is not too different from the magnitude in the pre-treatment period, then it may be reasonable to assume $\delta \in \Delta^{RM}(\bar{M})$, which explicitly bounds the relative magnitudes of violations of parallel trends in the post-treatment based on observed violations in the pre-treatment period. In other settings, researchers may be worried about violations of parallel trends that arise due to differences in smoothly evolving secular trends that differentially affect the treated and comparison groups. In this case, it may be reasonable to assume $\delta \in \Delta^{\mathrm{SD}}(M)$, which explicitly bounds the extent to which the slope of the difference in trends can vary across consecutive periods. Economic knowledge may imply additional restrictions as well. For example, if the researcher knows of a confounding policy change that would have a positive effect on the outcome, then it is reasonable to further assume that post-treatment difference in trends must be positive (*i.e.* $\delta_t \geq 0$ for $t > 0$).

In our empirical applications below, we illustrate how domain knowledge about the types of possible violations of parallel trends can inform the choice of Δ. We encourage applied researchers to use such domain knowledge to inform the restrictions they impose on the possible choices of parallel trends in their context.

6.1.2. Choice of inference procedure. Based on our theoretical results and Monte Carlo simulations, we recommend the ARP hybrid confidence sets for generic, polyhedral forms of Δ. For the special case of $\Delta^{\mathrm{SD}}(M)$—or other choices of Δ for which the consistency and finite-sample-near-optimality of FLCIs are guaranteed—we recommend FLCIs. Our recommended choice of inference procedure is implemented in the HonestDiD R and Stata packages that accompany the paper.[32] Furthermore, these confidence sets are quick to compute. Each sensitivity analysis plot in the empirical applications below took less than 9 minutes to compute on a 2012 Macbook Pro with a 2.3 GHz Quad-Core Intel Core i7 and 16 GB RAM.

6.1.3. Sensitivity analyses. Once the researcher has chosen a baseline class of restrictions on the possible violations of trends (*e.g.* relative magnitudes bounds $\Delta^{RM}(\bar{M})$ or smoothness bounds $\Delta^{\mathrm{SD}}(M)$), we recommend conducting sensitivity analysis over the associated parameter ($\bar{M} \geq 0$ or $M \geq 0$, respectively) that governs how different the post-treatment violations of parallel trends can be from the pre-trends. It is natural to report both the sensitivity of the researcher's causal conclusion to the choice of this parameter and the "breakdown" parameter value at which particular hypotheses of interest can no longer be rejected; similar "breakdown" concepts appear in the partial identification settings of Horowitz and Manski (1995), Kline and Santos (2013), Manski and Pepper (2018), and Masten and Poirier (2020).[33] We illustrate how one can interpret the magnitudes of the breakdown points in our two empirical illustrations below.

6.2. *Estimating the incidence of a value-added tax cut*

Benzarti and Carloni (2019b, henceforth, BC) 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 % in July of 2009. BC analyse the impact of this change using a dynamic DiD design that compares restaurants to a control group of other market services firms that were not affected by the VAT change, estimating the OLS regression specification

$$Y_{it} = \sum_{s \neq 2008} \beta_s \times 1[t = s] \times D_i + \phi_i + \lambda_t + \epsilon_{it}, \tag{19}$$

where Y_{it} is the log of (before-tax) profits for firm i in year t; D_i is an indicator for whether firm i is a restaurant; ϕ_i and λ_t are firm and year fixed effects; and standard errors are clustered at the regional level. BC's main finding is that the VAT reduction had a large, positive effect on restaurant profits. Figure 4 shows the estimated event-study coefficients $\{\hat{\beta}_s\}$ from specification (19) (Benzarti and Carloni, 2019a). We can formally reject the hypothesis that $\beta_{pre} = 0$ ($p < 0.01$), as there appears to have been a difference in trends between 2006 and 2007. Nevertheless, the post-treatment coefficients for 2009–2011 appear to be substantially larger in magnitude than any of the pre-trends coefficients.

A key concern in this empirical setting is that there may be unobserved, industry-specific or macroeconomic shocks that would have affected restaurants differently from other market-services firms even in the absence of a change in VAT. It seems reasonable to impose that the industry-specific shocks to restaurants in the post-treatment period are not too much larger

32. The latest version of the R and Stata packages are respectively available at http://github.com/asheshrambachan/HonestDiD and https://github.com/mcaceresb/stata-honestdid/.

33. Our main focus in this paper is on constructing robust confidence sets given a particular restriction $\Delta(M)$, rather than inference on the identification breakdown point or breakdown frontier as in *e.g.* Masten and Poirier (2020). Note, however, that if we define $M^* = \min M$ s.t. $0 \in \mathcal{S}(\beta, \Delta(M))$ to be the identification breakdown point for a null effect, and $\hat{M}^* = \min M$ s.t. $0 \in \mathcal{C}(\hat{\beta}_n, \hat{\Sigma}_n; \Delta(M))$ to be the sample breakdown point, then $P(\hat{M}^* \geq M^*) \geq P(0 \in \mathcal{C}(\hat{\beta}_n, \hat{\Sigma}_n; \Delta(M^*))$. It follows that $(-\infty, \hat{M}^*]$ is a valid $(1-\alpha)$-level confidence interval for M^* provided that our conditions for size control are satisfied for $\Delta(M^*)$. We suspect that our results could be extended to allow for uniform coverage of the breakdown frontier under additional regularity conditions, but leave this to future work.

FIGURE 4

Event-study coefficients $\{\beta_s\}$ for log profits, estimated using the event-study specification in (19).

than those in the pre-treatment period—whereas imposing that industry-specific shocks follow a smooth trend seems unreasonable—and so we base our analysis on bounds on relative magnitudes $\Delta^{RM}(\bar{M})$.

The left panel of Figure 5 shows robust confidence sets for the treatment effect in 2009 for $\Delta^{RM}(\bar{M})$ using different values of $\bar{M}$. The figure shows that if we impose $\bar{M}=1$, meaning that we restrict the post-treatment violations of parallel trends to be no larger than the maximal pre-treatment violation of parallel trends, then we obtain a robust confidence set of $[0.07, 0.31]$ for the causal effect on restaurant profits in 2009. This is wider than the original OLS confidence interval which is only valid if parallel trends holds exactly, but nevertheless rules out a null effect on restaurant profits in 2009. Looking further to the right, we see that the "breakdown value" for a null effect is around $\bar{M}=2$. Thus, our conclusion of a significant effect on restaurant profits depends on whether we are willing to restrict that the post-treatment violations of parallel trends can be no more than twice as large as the maximal pre-treatment violation. Given that the first year after the treatment coincided with a large recession in France (2009), it may be plausible that the differential factors affecting restaurants were larger in that year than in the pre-treatment period. Our approach helps formalize how much larger they would need to be to reject the conclusion of a null effect (or other hypotheses).

The right panel of Figure 5 shows analogous results when the estimand is the average causal effect on restaurant profits across all four post-treatment periods ($\bar{\tau}$). When $\bar{M}=1$, our robust confidence set now includes zero, and is about twice as large as for the first-period effect. The intuition for why the confidence sets are larger when looking at $\bar{\tau}$ than τ_{2009} is that $\Delta^{RM}(\bar{M})$ bounds the violation of parallel trends across *consecutive periods* by $\bar{M}$ times the max in the pre-treatment period. Thus, the identified set will be larger for later periods, since the treatment and control groups have more time to diverge (*e.g.* the identified set for the second period will be twice as larger as for the first period).[34] If we are willing to bound the magnitude of economic

34. This intuition holds generally for choices of Δ that bound changes across consecutive periods, but need not necessarily hold for other types of restrictions.

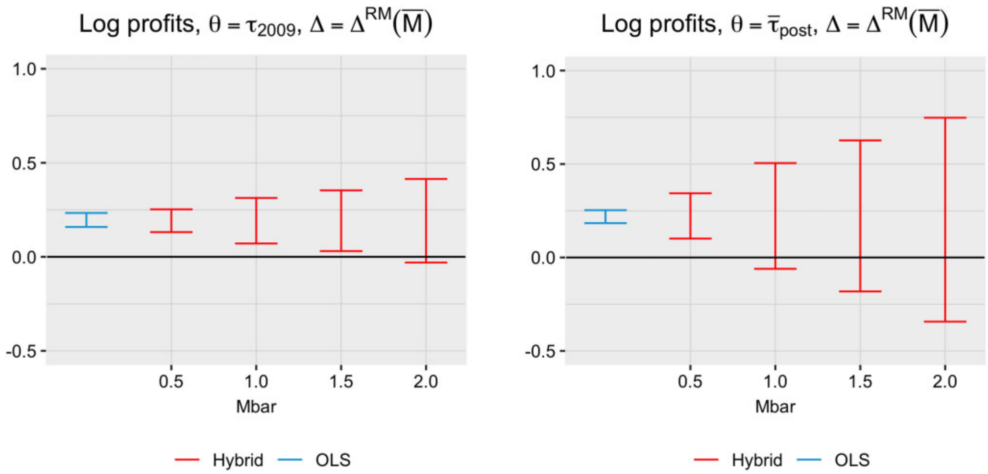

FIGURE 5

Sensitivity analysis for Benzarti and Carloni (2019b).

shocks by the max in the pre-treatment period, we will thus typically obtain wider confidence sets for parameters involving later periods.

Finally, in Supplementary Material, Appendix C, we compare the confidence sets reported in Figure 5 to the sample analogue to the identified set, $\mathcal{S}(\hat{\beta}, \Delta)$, which we show is a Hausdorff consistent estimator of the identified set in this example. We find that the reported confidence sets are between 40 and 80% longer than the estimated identified set, suggesting that both sampling uncertainty and the length of the identified set play an important role in the width of the confidence sets.

6.3. *The effect of duty-to-bargain laws on long-run student outcomes*

Lovenheim and Willen (2019a, henceforth 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. LW examine the impacts of these laws on the adult labour market outcomes of people who were students around the time that these laws were passed, comparing individuals across different states and different birth cohorts to exploit the differential timing of the passage of DTB laws across states. The authors estimate the following OLS regression specification separately for men and women, using data from the American Community Survey (ACS),

$$Y_{sct} = \sum_{r=-11}^{21} D_{scr}\beta_r + X'_{sct}\gamma + \lambda_{ct} + \phi_s + \epsilon_{sct}. \tag{20}$$

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] The event-study coefficients $\{\hat{\beta}_r\}$ estimate the dynamic treatment effects (or placebo

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.

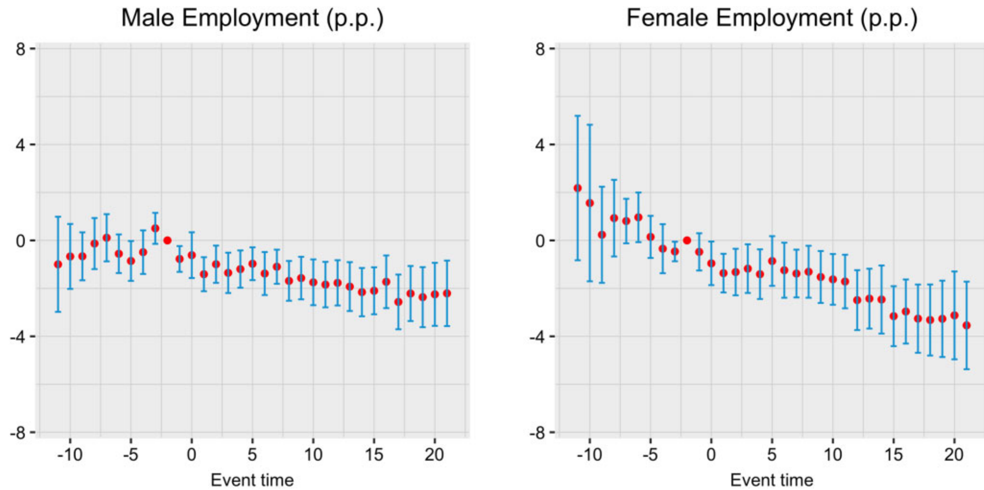

FIGURE 6

Event-study coefficient $\{\beta_r\}$ for employment, estimated using the event-study specification in (20).

effects) r years after DTB passage.[36] The remaining terms include time-varying controls, birth-cohort-by-ACS-year fixed effects, and state fixed effects. We normalize the event-study coefficient β_{-2} to 0.[37] We focus on the results where the outcome is employment.

Figure 6 plots the estimated event-study coefficients $\{\hat{\beta}_r\}$ from specification (20) (Lovenheim and Willen, 2019b). 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. By contrast, the results for women (right panel) suggest a downward-sloping pre-existing trend.

LW write that, the "primary concern in our identification strategy is the existence of secular trends that differ systematically with treatment" (p. 318), such as confounding changes in labour supply or educational attainment. Given that the concern is long-run trends that are likely to evolve smoothly over time, smoothness restrictions of the form Δ^{SD} seem natural in this context. Indeed, in some of their robustness checks, LW estimate models with group-specific linear trends, which roughly corresponds with the case $\Delta^{SD}(0)$.[38] It thus seems natural to consider relaxations of the form $\Delta^{SD}(M)$, which allows for deviations from non-linearity of no more than M between consecutive periods.

Figure 7 reports results for the treatment effect on employment for the cohort 15 years after the passage of a DTB law (as in Table 2 of LW), constructing robust confidence sets about how non-linear the difference in trends can be. In blue, we plot the original OLS confidence

36. Treatment timing in LW is staggered, and therefore the results in Sun and Abraham (2020) imply that β_r can be interpreted as a sensible weighted average of causal effects under parallel trends only if treatment effects are homogeneous across adoption cohorts. For simplicity, we focus on the robustness of the results to violations of parallel trends using the original specification in LW, which is valid under the assumption of homogeneous treatment effects. As discussed in Section 2.1, our sensitivity analysis can also be applied to estimators that are robust to treatment effect heterogeneity.

37. 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 reference period in our analysis.

38. The two are not exactly equivalent, however, because LW include parametric trends into what they call a "parametric event-study" model (see their specification (2)), which imposes that treatment effects are linear in time since treatment, rather than the flexible dynamic event-study specification (20).

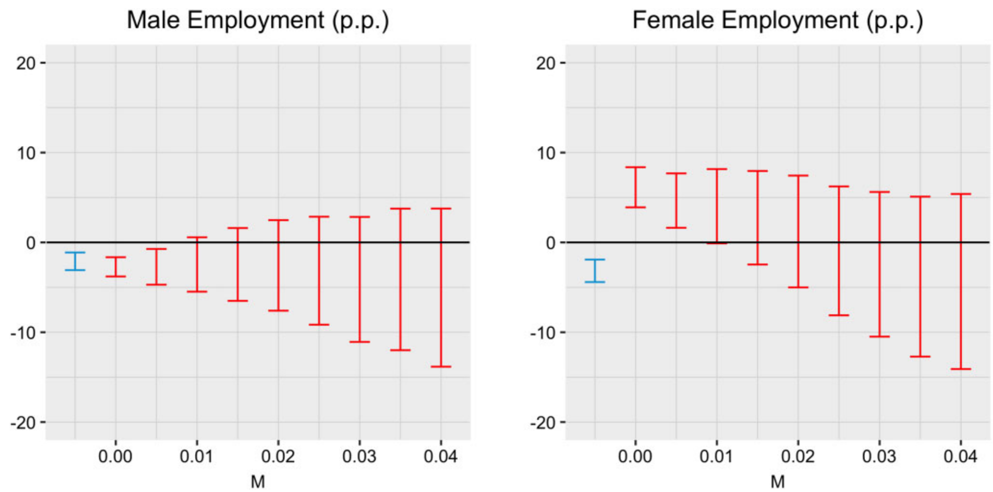

FIGURE 7

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

intervals for $\hat{\beta}_{15}$ from specification (20). In red, we plot FLCIs when $\Delta = \Delta^{SD}(M)$ for different values of M; recall that $M = 0$ corresponds with allowing only for linear violations of parallel trends, and larger values of M allow for larger deviations from linearity. In the analysis for men (left panel), the FLCIs are similar to those from OLS when allowing for violations of parallel trends that are approximately linear ($M \approx 0$), but become wider as we allow for more non-linearity; the breakdown value for a significant effect is $M \approx 0.01$. For women (right panel), 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 6. Indeed, for $M < 0.01$ the robust confidence set contains only positive values. Intuitively, this is because the point estimate for $t = 15$ lies above a linear extrapolation of the negative pre-trend. Thus, if we were to impose the same smoothness restrictions for men as for women, we would either have to reconcile significant effects of opposite signs by gender (if $M < 0.01$) or we would not be able to rule out null effects for both genders ($M \geq 0.01$).

How can we interpret the magnitudes of M in this example? We consider a calibration exercise based on the magnitudes of possible possible confounds: if violations of parallel trends were driven by confounding changes in education quality, what would a given value of M imply about the evolution of those confounds? Chetty *et al.* (2014) estimate that a 1 standard deviation increase in teacher value-added (VA) corresponds with a 0.4 percentage point increase in adult employment. Hence, a value of $M = 0.01$ would correspond with allowing the slope of the differential trend to change by the equivalent of a one-fourtieth of a standard deviation of teacher VA across consecutive periods. Since the robust confidence sets for both men and women begin to include zero around this value of M, the strength with which we can rule out a null effect depends on our assessment of the economic plausibility of such non-linearities.

Finally, in Supplementary Material, Appendix C, we again compare the reported confidence sets to the estimated identified set $\mathcal{S}(\hat{\beta}, \Delta)$. Our results suggest that sampling uncertainty is large relative to the length of the identified set when M is close to zero, but becomes less important as M increases.

7. CONCLUSION

This paper considers the problem of conducting inference in DiD and related designs that is robust to violations of the parallel trends assumption. We introduce a variety of restrictions on the class of possible differences in trends that formalize commonly made arguments in empirical work, generalizing the framework for partial identification in Manski and Pepper (2018). We provide inference procedures that are uniformly valid so long as the difference in trends satisfies these restrictions, and derive novel results on the power of these procedures. We recommend that applied researchers report robust confidence sets under economically motivated restrictions on parallel trends. We also recommend that researchers 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. Such sensitivity analyses make transparent what assumptions are needed in order to draw particular conclusions.

Acknowledgments. This paper was previously titled, “An Honest Approach to Parallel Trends.” We are grateful to Isaiah Andrews, Elie Tamer, and Larry Katz for their invaluable advice and encouragement. We also thank Clément de Chaisemartin, Gary Chamberlain, Raj Chetty, Peter Ganong, Ed Glaeser, Nathan Hendren, Ryan Hill, Ariella Kahn-Lang, Jens Ludwig, Sendhil Mullainathan, Claudia Noack, Frank Pinter, Alex Poirier, Adrienne Sabety, Pedro Sant’Anna, Jesse Shapiro, Neil Shephard, Jann Spiess, Jim Stock, and seminar participants at Brown, Chicago Booth, Dartmouth, Harvard, Michigan, Microsoft, Princeton, Rochester, UCL, Yale, SEA 2021, and ASSA 2022 for helpful comments, and Dorian Carloni for kindly sharing data. We are grateful to Mauricio Cáceres Bravo for his help developing the HonestDiD Stata package. We gratefully acknowledge financial support from the NSF Graduate Research Fellowship under Grant DGE1745303 (Rambachan) and Grant DGE1144152 (Roth).

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

Data Availability Statement

The data and code underlying this research is available on Zenodo at https://doi.org/10.5281/zenodo.7388015.

REFERENCES

ABADIE, A. (2005), “Semiparametric Difference-in-Differences Estimators”, *The Review of Economic Studies*, **72**, 1–19.

ANDREWS, I., ROTH, J. and PAKES, A. Inference for Linear Conditional Moment Inequalities”, *Review of Economic Studies*, forthcoming.

ARMSTRONG, T. and KOLESÁR, M. (2018), “Optimal Inference in a Class of Regression Models”, *Econometrica*, **86**, 655–683.

ARMSTRONG, T. and KOLESÁR, M. (2020a), “Sensitivity Analysis Using Approximate Moment Condition Models”, *Quantitative Economics*, **12**, 77–108.

ARMSTRONG, T. and KOLESÁR, M. (2020b), “Simple and Honest Confidence Intervals in Nonparametric Regression”, *Quantitative Economics*, **11**, 1–39.

ASHENFELTER, O. (1978), “Estimating the Effect of Training Programs on Earnings”, *The Review of Economics and Statistics*, **60**, 47–57.

ATHEY, S. and IMBENS, G. W. (2021), “Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption”, *Journal of Econometrics*, **226**, 62–79.

BENZARTI, Y. and CARLONI, D. (2019a), “Replication Data For: Who Really Benefits from Consumption Tax Cuts? Evidence from a Large VAT Reform in France” (Technical Report, American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]). https://doi.org/10.3886/E114723V1.

BENZARTI, Y. and CARLONI, D. (2019b), “Who Really Benefits from Consumption Tax Cuts? Evidence from a Large VAT Reform in France”, *American Economic Journal: Economic Policy*, **11**, 38–63.

BHULLER, M., HAVNES, T., LEUVEN, E. and MOGSTAD, M. (2013), “Broadband Internet: An Information Superhighway to Sex Crime?”, *The Review of Economic Studies*, **80**, 1237–1266.

BILINSKI, A. and HATFIELD, L. A. (2020), “Nothing to See Here? Non-inferiority Approaches to Parallel Trends and Other Model Assumptions” *arXiv:1805.03273 [stat.ME]*.

BORUSYAK, K. and JARAVEL, X. (2016), "Revisiting Event Study Designs" (SSRN Scholarly Paper ID 2826228, Rochester, NY: Social Science Research Network).

BUGNI, F. A., CANAY, I. A. and SHI, X. (2017), "Inference for Subvectors and Other Functions of Partially Identified Parameters in Moment Inequality Models", *Quantitative Economics*, **8**, 1–38.

CALLAWAY, B. and SANT'ANNA, P. H. C. (2021), "Difference-in-Differences with Multiple Time Periods", *Journal of Econometrics*, **225**, 200–230.

CANAY, I. and SHAIKH, A. (2017), "Practical and Theoretical Advances in Inference for Partially Identified Models" in Honoré, B., Pakes, A., Piazzesi, M. and Samuelson, L. (eds) *Advances in Economics and Econometrics* (Cambridge: Cambridge University Press) 271–306.

CHEN, X., CHRISTENSEN, T. M. and TAMER, E. (2018), "Monte Carlo Confidence Sets for Identified Sets", *Econometrica*, **86**, 1965–2018.

CHERNOZHUKOV, V., NEWEY, W. K. and SANTOS, A. (2015), "Constrained Conditional Moment Restriction Models" *arXiv:1509.06311 [math, stat]*.

CHETTY, R., FRIEDMAN, J. N. and ROCKOFF, J. E. (2014), "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", *American Economic Review*, **104**, 2633–2679.

CHO, J. and RUSSELL, T. M. (2019), "Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments" *arXiv:1810.03180 [econ.EM]*.

COX, G. and SHI, X. (2023), "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models", *The Review of Economic Studies*, **90**, 201–228.

DE CHAISEMARTIN, C. and D'HAULTFŒUILLE, X. (2020), "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects", *American Economic Review*, **110**, 2964–2996.

DE CHAISEMARTIN, C. and D'HAULTFŒUILLE, X. (2021), "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey" (SSRN Scholarly Paper ID 3980758, Rochester, NY: Social Science Research Network).

DETTE, H. and SCHUMANN, M. (2020), "Difference-in-Differences Estimation Under Non-Parallel Trends" (Working Paper).

DOBKIN, C., FINKELSTEIN, A., KLUENDER, R. and NOTOWIDIGDO, M. J. (2018), "The Economic Consequences of Hospital Admissions", *American Economic Review*, **108**, 308–352.

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

FLYNN, Z. (2019), "Inference Based on Continuous Linear Inequalities via Semi-Infinite Programming" (SSRN Scholarly Paper ID 3390788, Rochester, NY: Social Science Research Network).

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

FREYALDENHOVEN, S., HANSEN, C. and SHAPIRO, J. (2019), "Pre-event Trends in the Panel Event-study Design", *American Economic Review*, **109**, 3307–3338.

GAFAROV, B. (2019), "Inference in high-dimensional set-identified affine models" *arXiv:1904.00111 [econ.EM]*.

GOODMAN-BACON, A. (2018), "Public Insurance and Mortality: Evidence from Medicaid Implementation", *Journal of Political Economy*, **126**, 216–262.

GOODMAN-BACON, A. (2021), "Difference-in-Differences with Variation in Treatment Timing", *Journal of Econometrics*, **225**, 254–277.

GREENSTONE, M. and HANNA, R. (2014), "Environmental Regulations, Air and Water Pollution, and Infant Mortality in India", *American Economic Review*, **104**, 3038–3072.

HANSEN, C. B. (2007), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T is Large", *Journal of Econometrics*, **141**, 597–620.

HANSEN, B. E. and LEE, S. (2019), "Asymptotic Theory for Clustered Samples", *Journal of Econometrics*, **210**, 268–290.

HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P. (1998), "Characterizing Selection Bias Using Experimental Data", *Econometrica*, **66**, 1017–1098.

HIRANO, K. and PORTER, J. R. (2012), "Impossibility Results for Nondifferentiable Functionals", *Econometrica*, **80**, 1769–1790.

HO, K. and ROSEN, A. (2017), "Partial Identification in Applied Research: Benefits and Challenges" in Honoré, B., Pakes, A., Piazzesi, M. and Samuelson, L. (eds) *Advances in Economics and Econometrics* (Cambridge: Cambridge University Press) 307–359.

HOROWITZ, J. L. and MANSKI, C. F. (1995), "Identification and Robustness with Contaminated and Corrupted Data", *Econometrica*, **63**, 281–302.

HUDSON, S., HULL, P. and LIEBERSOHN, J. (2017), "Interpreting Instrumented Difference-in-Differences" (Working Paper).

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

KAHN-LANG, A. and LANG, K. (2020), "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications", *Journal of Business and Economic Statistics*, **38**, 613–620.

KAIDO, H. and SANTOS, A. (2014), "Asymptotically Efficient Estimation of Models Defined by Convex Moment Inequalities", *Econometrica*, **82**, 387–413.

KAIDO, H., SANTOS, A., MOLINARI, F. and STOYE, J. (2019), "Confidence Intervals for Projections of Partially Identified Parameters", *Econometrica*, **87**, 1397–1432.

KAIDO, H., SANTOS, A., MOLINARI, F. and STOYE, J. (2022), "Constraint Qualifications In Partial Identification", *Econometric Theory*, **38**, 1–24.

KEELE, L. J., SMALL, D. S., HSU, J. Y. and FOGARTY, C. B. (2019), "Patterns of Effects and Sensitivity Analysis for Differences-in-Differences" *arXiv:1901.01869 [stat.AP]*.

KIM, W., KWON, K., KWON, S. and LEE, S. (2018), "The Identification Power of Smoothness Assumptions in Models with Counterfactual Outcomes", *Quantitative Economics*, **9**, 617–642.

KLINE, P. and SANTOS, A. (2013), "Sensitivity to Missing Data Assumptions: Theory and an Evaluation of the U.S. Wage Structure", *Quantitative Economics*, **4**, 231–267.

KOLESÁR, M. and ROTHE, C. (2018), "Inference in Regression Discontinuity Designs with a Discrete Running Variable", *American Economic Review*, **108**, 2277–2304.

LEAVITT, T. (2020), "Beyond Parallel Trends: Improvements on Estimation and Inference in the Difference-in-Differences Design" (Working Paper).

LEE, J. Y. and SOLON, G. (2011), "The Fragility of Estimated Effects of Unilateral Divorce Laws on Divorce Rates", *The B.E. Journal of Economic Analysis & Policy*, **11**, 1–9.

LOVENHEIM, M. F. and WILLEN, A. (2019a), "The Long-Run Effects of Teacher Collective Bargaining", *American Economic Journal: Economic Policy*, **11**, 292–324.

LOVENHEIM, M. F. and WILLEN, A. (2019b), "Replication Data For: The Long-Run Effects of Teacher Collective Bargaining" (Technical Report, American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]). https://doi.org/10.3886/E116525V1.

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

MANSKI, C. F. (1989), "Anatomy of the Selection Problem", *The Journal of Human Resources*, **24**, 343–360.

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

MANSKI, C. F. (1997), "Monotone Treatment Response", *Econometrica*, **65**, 1311–1334.

MANSKI, C. F. (2003), *Partial Identification of Probability Distributions* (New York, NY: Springer).

MANSKI, C. F. (2007), *Identification for Prediction and Decision* (Cambridge, MA: Harvard University Press).

MANSKI, C. F. (2008), *Identification for Prediction and Decision* (illustrated ed., Cambridge, Mass: Harvard University Press).

MANSKI, C. F. (2013), *Public Policy in an Uncertain World: Analysis and Decisions* (Cambridge, MA: Harvard University Press).

MANSKI, C. F. and PEPPER, J. V. (2013), "Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections", *Journal of Quantitative Criminology*, **29**, 123–141.

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

MASTEN, M. A. and POIRIER, A. (2020), "Inference on Breakdown Frontiers", *Quantitative Economics*, **11**, 41–111.

MOLINARI, F. (2020), "Microeconometrics with Partial Identification", in Durlauf, S. N., Hansen, L. P., Heckman, J. J. and Matzkin, R. L. (eds) *Handbook of Econometrics*, Vol. 7 of *Handbook of Econometrics, Volume 7A*, Chapter 5 (Amsterdam: Elsevier) 355–486.

MOREIRA, M. J. (2003), "A Conditional Likelihood Ratio Test for Structural Models", *Econometrica*, **71**, 1027–1048.

MOREIRA, M. J. and RIDDER, G. (2019), "Efficiency Loss of Asymptotically Efficient Tests in an Instrumental Variables Regression" (SSRN Scholarly Paper ID 3348716, Rochester, NY: Social Science Research Network).

MÜLLER, U. K. (2011), "Efficient Tests Under a Weak Convergence Assumption", *Econometrica*, **79**, 395–435.

NOACK, C. and ROTHE, C. (2020), "Bias-Aware Inference in Fuzzy Regression Discontinuity Designs" *arXiv:1906.04631 [econ.EM]*.

RAMBACHAN, A. and ROTH, J. (2021), "An Honest Approach to Parallel Trends" (Working paper, p. 126).

ROMANO, J. P. and SHAIKH, A. M. (2008), "Inference for Identifiable Parameters in Partially Identified Econometric Models", *Journal of Statistical Planning and Inference*, **138**, 2786–2807.

ROMANO, J. P., SHAIKH, A. M. and WOLF, M. (2014), "A Practical Two-Step Method for Testing Moment Inequalities", *Econometrica*, **82**, 1979–2002.

ROTH, J. (2022), "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends", *American Economic Review: Insights*, **4**, 305–322.

ROTH, J., SANT'ANNA, P. H. C., BILINSKI, A. and POE, J. (2022), "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature" *arXiv:2201.01194 [econ, stat]*.

SANT'ANNA, P. H. C. and ZHAO, J. B. (2020), "Doubly Robust Difference-in-Differences Estimators", *Journal of Econometrics*, **219**, 101–122.

SCHRIJVER, A. (1986), *Theory of Linear and Integer Programming* (Chicester: Wiley-Interscience).

STOCK, J. and WATSON, M. (2008), "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression", *Econometrica*, **76**, 155–174.

SUN, L. and ABRAHAM, S. (2021), "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects", *Journal of Econometrics*, **225**, 175–199.

TAMER, E. (2010), "Partial Identification in Econometrics", *Annual Review of Economics*, **2**, 167–195.

VAN DER VAART, A. W. and WELLNER, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics* (New York, NY: Springer).

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

YE, T., KEELE, L., HASEGAWA, R. and SMALL, D. S. (2020), "A Negative Correlation Strategy for Bracketing in Difference-in-Differences with Application to the Effect of Voter Identification Laws on Voter Turnout" *arXiv:2006.02423 [stat.ME]*.