I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16202

Fixed Effects and Causal Inference

Daniel L. Millimet
Marc F. Bellemare

JUNE 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16202

Fixed Effects and Causal Inference

Daniel L. Millimet
Southern Methodist University and IZA

Marc F. Bellemare
University of Minnesota

JUNE 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Fixed Effects and Causal Inference*

Across many disciplines, the fixed effects estimator of linear panel data models is the default method to estimate causal effects with nonexperimental data that are not confounded by time-invariant, unit-specific heterogeneity. One feature of the fixed effects estimator, however, is often overlooked in practice: With data over time $t \in \{1,...,T\}$ for each unit of observation $i \in \{1,...,N\}$, the amount of unobserved heterogeneity the researcher can remove with unit fixed effects is weakly decreasing in T. Put differently, the set of attributes that are time-invariant is not invariant to the length of the panel. We consider several alternatives to the fixed effects estimator with $T > 2$ when relevant unit-specific heterogeneity is not time-invariant, including existing estimators such as the first-difference, twice first-differenced, and interactive fixed effects estimators. We also introduce several novel algorithms based on rolling estimators. In the situations considered here, there is little to be gained and much to lose by using the fixed effects estimator. We recommend reporting the results from multiple linear panel data estimators in applied research.

JEL Classification: C23, C51, C52

Keywords: panel data, fixed effects, first-differences, interactive fixed effects, unobserved heterogeneity, time-varying individual effects

Corresponding author:
Daniel L. Millimet
Department of Economics
Box 0496
Southern Methodist University
Dallas, TX 75275-0496
USA

E-mail: millimet@smu.edu

* The authors are grateful to Jeff Bloem, Bernhard Dalheimer, Paul Glewwe, Jason Kerwin, John Mullahy, Christian Vossler, and Le Wang for comments, as well as seminar participants at the New York Econometrics Camp including Badi Baltagi, Hugo Jales, Ivan Korolev, Carlos Lamarche, and James MacKinnon. All remaining errors are our own.

You keep using that word. I do not think it means what you think it means.

– Inigo Montoya, *The Princess Bride* (1987)

1 Introduction

Ever since Mundlak (1961) tried to explain agricultural productivity differentials by including a separate dummy variable for each farmer in the data to control for "management bias," the fixed effects estimator (FE) has been used by researchers to control for time-invariant, unit-specific heterogeneity (see, e.g., Balestra and Nerlove, 1966; Hsiao, 2007). Researchers continue to rely on FE for causal inference in panel data settings (Angrist and Pischke, 2009; Imai and Kim, 2019). Imai and Kim (2021, p. 405), for instance, refer to FE as the "default methodology for estimating causal effects with panel data," while Hill et al. (2020, p. 367) state that "fixed-effects models for panel data are now widely recognized as powerful analytic tools for longitudinal data analysis."

From a causal inference perspective, the advantage of the FE estimator over pooled ordinary least squares (POLS) is that the FE estimator, by removing both observed *and* unobserved time-invariant unit-specific heterogeneity, requires an arguably more palatable assumption to be unbiased. Unbiasedness of the FE estimator requires strict exogeneity with respect to the time-varying error, whereas POLS requires exogeneity with respect to a composite error that includes more than just the time-varying error, as we discuss below.

One important limitation of the FE estimator seems to be rarely acknowledged in practice, however: Unobserved heterogeneity that is time-invariant is not invariant to the length of the panel. In this paper, we explore the consequences of ignoring this limitation in empirical research and provide guidance for applied researchers.

The notion that time-invariant, unobserved heterogeneity depends crucially on the length of the panel is made explicit in Mundlak (1961, p. 44)

> "Instead of beginning by conceptualizing what we mean by management we shall assume that whatever management is, it does not change considerably over time; and for short periods, say a few years, it can be assumed to remain constant."

and in Mundlak (1978, p. 82)

> "[I]t would be unrealistic to assume that the individuals do not change in a differential way as the model assumes ... [I]t is more realistic to assume that individuals do change differentially but at a pace that can be ignored for short time intervals."

Yet, Mundlak's point has become lost over time.

With data over time $t \in \{1, ..., T\}$ for each unit $i \in \{1, ..., N\}$, researchers eagerly take advantage of increased data availability and use longer and longer panels (Backhouse and Cherrier, 2017).[1,2] McKenzie

[1] We focus here on balanced panels, but our core point also applies to unbalanced panels.

[2] While it is possible for $T \to \infty$ because the researcher uses higher-frequency data for the same length of time (e.g., T goes from 5 to 10 because data over a five-year period are collected semi-annually rather than annually), these are not the situations we have in mind. Rather, we consider here only those cases where T increases because additional rounds of data are collected, but for which the frequency remains the same (e.g., T goes from 5 to 10 because the data cover a 10-year period instead of a five-year period).

(2012) even advocates for longer T when researchers collect their own experimental data. The fact that some or all of what is time-invariant when, say, $T = 5$ may no longer be time-invariant when $T = 10$ is consistently overlooked.

For example, individual preferences—long considered to be exogenous and fixed—are increasingly viewed by economists as endogenous, being influenced by shocks, individual circumstances, groups, and institutions (e.g., Becker and Mulligan, 1997; Bowles, 1998; Fershtman and Segal, 2018; Liebenehm, Degener and Strobl, forthcoming). Similarly, in empirical analyses of individual labor market or educational outcomes, a worker's "innate ability, perseverance, motivation, and industriousness," are unlikely to be constant over the life cycle (Bai, 2009, p. 1233). Ding and Lehrer (2014, p. 69) state that "researchers often implicitly assume that both the impact and stock of unobserved ability are constant over time" but this "appears inconsistent with a rapidly growing body of scientific evidence which indicates that the impacts and development of these unobserved factors vary substantially over the life cycle." Likewise, in cross-country regressions, there is likely little that remains fixed over time for a given country as borders, climate, institutions, and even topography are all time-varying.

Nevertheless, researchers typically neither alter their interpretation of the fixed effects nor acknowledge the ever-stronger assumption required for unbiasedness as T grows. Hill et al. (2020, p. 363) state:

> "[W]e often assume (without direct confirmation) that certain characteristics (e.g., biology, personality, or culture) do not change over time ... [I]t is often unclear whether some characteristics are actually fixed or variable. So what exactly are fixed-effects models controlling? ... [T]he answer to this question is often ambiguous. Uncertainty along these lines could introduce a host of theoretical and empirical problems."

Moreover, the issue also arises even when relevant attributes remain time-invariant as T grows if the *effects* of these attributes are time-varying.[3] Hill et al. (2020, p. 364) continue:

> "Researchers typically state and restate how fixed-effects models control for time-invariant characteristics, but this is only true if those variables have the same effects at each point in time. If the coefficients for supposed time-invariant characteristics vary with time, they become equivalent to time-varying characteristics."

As such, there is an important trade-off that needs to be reckoned with. That is, though the variance of the FE estimator will shrink as T grows, the estimator will be biased and inconsistent if the fixed effects remove less unobserved heterogeneity and this invalidates the required strict exogeneity assumption. In the limit, the FE and POLS estimators may even coincide as $T \to \infty$ if there is *no* relevant unit-specific unobserved heterogeneity that is time-invariant *and* has a constant effect across all time periods.[4]

We assess this bias-variance trade-off by extending the standard linear panel data model to include what we refer to as *regime*-specific heterogeneity in the unit fixed effects. Here, a regime represents a group of consecutive time periods S, with $S < T$, over which relevant unobserved heterogeneity for a particular unit

[3] Through the paper, we use the term *relevant* to refer to unobserved heterogeneity that is not independent of the covariates.

[4] We are not referring to $T \to \infty$ in the usual asymptotic sense where the underlying data-generating process is fixed as T grows. Rather, we are referring here to a situation where the data-generating process itself changes as T grows because the definition of the fixed effect changes.

is constant. [5] Our setup is analogous to structural breaks in the unit fixed effects, but with unknown break dates that may vary across units. It is also similar to the hierarchical structure considered in Papke and Wooldridge (forthcoming), except that we are considering aggregation of unit-specific heterogeneity over regimes rather than groups of units.

When relevant unobserved differences across units are constant across regimes, but not the entire sample, the strict exogeneity assumption will fail and FE will be biased and inconsistent. We discuss several estimators that not only improve on the FE estimator in this situation, but are also easy to implement in applied research. These include the well-known first-difference (FD) and twice first-differenced (TFD) estimators, as well as a novel estimator which we dub the rolling first-differences (RFD) estimator. We also discuss several extensions to RFD including rolling twice first-differenced (RTFD) and rolling fixed effects (RFE). RFD proceeds by estimating a series of rolling regressions by FD where only two time periods are retained at each step and then aggregating the estimates into a final estimate using minimum-distance estimation (MDE). The RTFD estimator proceeds by estimating a series of rolling regressions by TFD where only three time periods are retained at each step. The RFE estimator is identical to RFD except that more than two time periods are retained at each step and the FE estimator is employed. Our consideration of rolling estimators is analogous to the proposed use of rolling regressions for time series models in Cai and Juhl (forthcoming).

Through simulations and replications of Rose (2004), Tomz, Goldstein and Rivers (2007), and James (2015), we show that FD and RFD both outperform FE when the standard linear panel data model is misspecified, but generally not when it is correctly specified. Thus, the estimators we discuss sacrifice (some) efficiency for (much) robustness. This is also the case in Aquaro and Čížek (2013) when dealing with outliers, and in Pesaran and Zhou (2018) when considering the asymptotic efficiency of POLS relative to FE in the presence of sparse unit fixed effects. Moreover, we find that the RFD estimator outperforms the FD estimator in some settings. Finally, we find that other existing estimators—namely, TFD, RTFD, and the interactive fixed effects (IFE) estimator from Bai (2009)—work well in certain situations. Overall, we recommend that researchers report FD, RFD, TFD, and IFE estimates in addition to FE when using panel data with more than two time periods, particularly when the sample spans several periods. The RFE estimator should prove useful when relevant unobserved heterogeneity is likely to be time-invariant over regimes with $2 < S < T$.

While we are not the first to introduce relevant heterogeneity into the standard linear panel data specification, as we discuss in Section 2, our analysis makes three important contributions. First, we clarify the trade-off involved with ever-longer panel data when using the standard FE estimator—a trade-off that pits the efficiency gain from more data against unbiasedness and consistency. Second, we provide guidance to empirical researchers regarding simple alternatives to the rote application of the FE estimator in static two-way fixed effects (TWFE) models. Third, we replicate articles by Rose (2004) as well the comment on Rose (2004) by Tomz, Goldstein and Rivers (2007), and by James (2015) and find that using the FD, TFD, RFD, RFE, and IFE estimators lead to conclusions that are qualitatively different from those authors' reported FE results. Moreover, in the former replication, we empirically resolve a puzzle in the international trade literature.

Prior to continuing, a few remarks are warranted. We recognize that not only might there be temporal

[5] Of course, if $S = 1$, then each period corresponds to a new regime, implying that no relevant unit-specific unobserved heterogeneity is time-invariant.

heterogeneity in the unit fixed effects, but the slope coefficients as well. We mostly abstract from this for two reasons. First, as we discuss in Section 2, there is a fairly large literature that already addresses heterogeneity in the slope coefficients while ignoring temporal heterogeneity in the unit fixed effects. As such, we complement that literature. Second, while it is a context-specific empirical question, temporal heterogeneity in the unit fixed effects strikes us as a potentially more important concern as T grows. Each slope coefficient in a linear panel data model represents the effect of a well-defined observed characteristic of units. Thus, it is possible to use institutional knowledge to garner some insights into the possibility that the coefficient is heterogeneous across space and/or time. In contrast, the unit fixed effect is a composite of many unobserved attributes that are left entirely unspecified by the researcher. Consequently, there is little basis upon which to justify the assumption that this composite of unobserved attributes *and* the effect of this composite are time-invariant. Finally, our novel RFD, RTFD, and RFE estimators make it possible to allow for temporal heterogeneity in both the slope coefficients and the unit fixed effects. Each entails a series of rolling regressions, estimating separate parameters over a short time interval. Rather than assuming common parameters over time and aggregating using MDE, one can examine the various regime-specific coefficient estimates and simply not aggregate.

The remainder of this paper is organized as follows. Section 2 discusses the relevant econometric literature. Section 3 introduces the static linear panel data setup, clarifies what is at stake as T increases, and compares the FE, FD, RFD estimators, as well as others. In section 4, we illustrate these estimators via simulation and two replications. Section 5 concludes.

2 Literature Review

As stated previously, we are not the first to introduce greater heterogeneity into the standard linear panel data specification. A growing econometric literature extends the FE estimator by relaxing the assumption that relevant unit-specific heterogeneity is time-invariant. But models in this literature—referred to as *time-varying individual effects*—impose restrictions on how the unit fixed effects vary over time. Moreover, these models have yet to gain popularity among applied researchers. Mundlak (1978), for his part, introduces unit-specific time trends into the standard panel data model—a practice that is now widespread (see, e.g., Autor, 2003). More recently, IFE models allow for common time-varying factor loadings on the unit-specific fixed effects (e.g., Ahn, Hoon Lee and Schmidt, 2001, 2013; Han, Orea and Schmidt, 2005; Bai, 2009). This setup allows for time-invariant unit-specific heterogeneity to have homogeneous but time-varying effects as well as time-invariant, unit-specific responses to common period-specific shocks. Thus, IFE models relax the traditional FE estimator, but in a fairly rigid manner.

Another class of models extends the standard specification by allowing for (unknown) group membership or structural breaks (e.g., Bonhomme and Manresa, 2015; Boldea, Drepper and Gan, 2020; Lumsdaine, Okui and Wang, forthcoming; Kaddoura and Westerlund, forthcoming). Bonhomme and Manresa (2015) allows for group-specific time effects. Boldea, Drepper and Gan (2020) examine a model with multiple structural breaks where each regime has unique values of the slope parameters and the unit fixed effects. Going one step further, Lumsdaine, Okui and Wang (forthcoming) also allow for multiple structural breaks, but each regime is partitioned into groups with common slope coefficient and group-specific fixed effects. Their setup also permits units to change groups at the structural break as well. Kaddoura and Westerlund (forthcoming) consider the IFE model where the slope coefficients are allowed to experience structural breaks.

As with IFE, these studies make great strides in moving beyond the traditional FE estimator, but they have some practical limitations. First, models based on structural breaks assume a common break date for all units.[6] Second, models permitting multiple structural breaks with or without groups become computationally intensive as the number of breaks or groups increase since estimation typically involves searching over all possible break dates as well as possible group assignments.

The bias–efficiency trade-off that emerges as T increases when unit fixed effects capture less unobserved heterogeneity also relates to several other literatures exploring failures of the standard linear panel data framework. For example, Bramati and Croux (2007) and Aquaro and Čížek (2013) discuss the bias of the FE estimator in the presence of outliers and consider robust alternatives, particularly ones utilizing temporal differencing rather than mean-differencing. Our allowance for changes in the unit fixed effects across regimes, where regimes of time-invariant heterogeneity may differ across units, may behave similarly to the outliers considered in this literature. Bonhomme and Manresa (2015), Papke and Wooldridge (forthcoming), and Lewis et al. (2022) consider the optimal choice of fixed effects with or without slope heterogeneity in a static linear panel framework where units are clustered into groups. Additionally, many studies examine heterogeneity of various forms in the slope coefficients. Sarafidis and Weber (2015) and Gibbons, Serrato and Urbancic (2019) consider a panel data model with group-specific slope coefficients estimated using the traditional FE estimator that ignores this slope heterogeneity. Cai and Juhl (forthcoming) consider a time series model with time-varying slope coefficients. The authors develop correct inference when comparing estimates across rolling regressions to test for the presence of such heterogeneity. Keane and Neal (2020) allow for spatial and temporal slope heterogeneity and propose a new estimation algorithm referred to as the mean observation OLS procedure. See also Su and Chen (2013), Ando and Bai (2016), Baltagi, Feng and Kao (2016), Su, Shi and Phillips (2016), Liu et al. (2020), Mehrabani (forthcoming), among others.[7]

We contribute to this literature by clarifying the trade-off involved with ever-longer panel data when using the standard FE estimator for applied researchers, as well as providing guidance to empirical researchers concerning simple alternatives to the FE estimator.

3 Panel Data Estimation

We first present the general FE setup and explain how the FE estimator's usefulness for causal identification decreases as T grows. We then discuss solutions to this problem: FD, TFD, RFD, and RFE estimators. Finally, we discuss other existing estimators that may prove useful in certain situations.

3.1 Setup

In the standard static panel data model, the assumed data-generating process (DGP) is given by

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}, \quad i = 1, ..., N; \; t = 1, ..., T \tag{1}$$

where x_{it} is a $1 \times K$ vector of time-varying observed attributes (which may include period fixed effects), α_i is a unit-specific fixed effect, and ε_{it} is a mean-zero error. Researchers describe α_i as capturing all

[6] Jiang and Kurozumi (forthcoming) develop a test of the common break date assumption.

[7] Sun and Shapiro (2022) discusses slope heterogeneity in the context of linear panel data models assessing treatment exposure and connects this to the recent literature on difference-in-differences with staggered adoption.

time-invariant attributes of unit i, whether observed or unobserved, whereas ε_{it} captures all *time-varying*, unobserved determinants of y (that are not constant across all units if x includes period fixed effects). Our focus here is on what exactly is meant by "time-invariant" and "time-varying" in this context—terms that are often casually invoked by researchers when justifying the unbiased estimation of causal parameters.

Estimation of Equation (1) is most often accomplished using the FE estimator, also known as the within estimator and equivalent to the least squares dummy variables (LSDV) estimator for all T (Mundlak, 1961). This entails using POLS after mean-differencing to estimate the model given by

$$\begin{aligned} y_{it} - \overline{y}_i &= (x_{it} - \overline{x}_i)\,\beta + (\varepsilon_{it} - \overline{\varepsilon}_i) \\ \ddot{y}_{it} &= \ddot{x}_{it}\beta + \ddot{\varepsilon}_{it}, \end{aligned} \tag{2}$$

where two dots over a variable indicate deviation from its unit-specific mean. Given a random sample, $\{y_{it}, x_{it}\}_{i=1,...,N;t=1,...,T}$, the estimate of β, $\widehat{\beta}_{FE}$, requires $\mathtt{E}\left[\varepsilon_{it}|\mathbf{x}_i, \alpha_i\right] = 0$ for all t for unbiasedness, where $\mathbf{x}_i$ is a $T \times K$ matrix of covariates for unit i. This condition, known as strict exogeneity, implies that x_{it} is independent of the time-varying error term, ε_{is}, in every period $s = 1, ..., T$ conditional on α_i.[8]

Our interest is in the role played by T. The presumption often made by applied researchers is that larger T is *better* than smaller T. In studies with $N >> T$, which is typically the case in applied microeconomics, asymptotic results rely on $N \to \infty$ with T fixed. As such, the only advantage of increasing T is efficiency; sample size grows by N with each additional time period assuming a balanced panel, yet only one degree of freedom is lost if time effects are included (Hsiao, 2007).[9] While efficiency is important, when the objective is causal inference, unbiasedness and consistency are equally, if not more, salient (Angrist and Pischke, 2009).

Despite this, an overlooked consequence of larger T is how it impacts the strict exogeneity assumption. Specifically, the unit fixed effect, α_i, captures all unobserved attributes of unit i that are time-invariant *over the sample period*. Similarly, the time-varying error component, ε_{it}, captures all unobserved attributes of unit i that are not time-invariant *over the sample period*. Thus, as T increases, it is plausible that some previously unobserved, time-invariant attributes now become time-varying. Alternatively, it is possible that there is no change in these attributes, but the effects of these attributes change *over the sample period*. In either case, the strict exogeneity assumption is less likely to hold as T increases since conditioning on α_i is less impactful and more unobserved heterogeneity is relegated to ε_{it}. In this sense, the claim in Thombs (2022, p. 1) is quite prophetic: "As interest in big data increases and the temporal depth of widely used data sets expands, large N, large T data are becoming more ubiquitous, making it paramount that researchers have the appropriate tools to analyze these data."

To formalize the role of T, we introduce some notation. Define $\alpha_i^{t,t'}$ as unobserved attributes of unit i that are time-invariant over the regime spanning t to t', where $t \geq 1$, $t' \leq T$, and $t < t'$. Now, suppose a researcher obtains access to additional periods of data such that T increases from T to T'. As a result of

[8] In contrast, estimation of Equation (1) by POLS requires (contemporaneous) exogeneity, given by $\mathtt{E}\left[\alpha_i + \varepsilon_{it}|x_{it}\right] = 0$ for all t.

[9] In dynamic panels, there are gains to larger T beyond efficiency as the bias of FE estimates diminishes with T (Nickell, 1981).

the longer panel, the DGP in Equation (1), where $\alpha_i \equiv \alpha_i^{1,T}$, becomes

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha_i^{1,T'} + \alpha_{it} + \varepsilon_{it}, \quad i = 1, ..., N;\ t = 1, ..., T' \\ &\equiv x_{it}\beta + \alpha_i^{1,T'} + \eta_{it}, \end{aligned} \tag{3}$$

where α_{it} are unobserved attributes of unit i that are *not time-invariant* over the regime spanning periods 1 to T', but are *time-invariant* over the regime spanning periods 1 to T or $T+1$ to T'. Thus, the attributes captured by $\alpha_i^{1,T'}$ are at best the same attributes captured by $\alpha_i^{1,T}$, such that $\alpha_i^{1,T'} \subseteq \alpha_i^{1,T}$. Similarly, $\alpha_i^{1,T'} \subseteq \alpha_i^{T+1,T'}$.

If the FE estimator is used to estimate Equation (3), the strict exogeneity assumption is given by $\mathtt{E}\left[\eta_{it}|\widetilde{\mathbf{x}}_i, \alpha_i^{1,T'}\right] = 0$, $t = 1, ..., T'$, where $\widetilde{\mathbf{x}}_i$ is a $T' \times K$ matrix of covariates for unit i. This is a stronger condition than that required in the original sample given by $\mathtt{E}\left[\varepsilon_{it}|\mathbf{x}_i, \alpha_i^{1,T}\right] = 0$, $t = 1, ..., T$, as the former requires $\mathtt{E}\left[\alpha_{it}|\widetilde{\mathbf{x}}_i, \alpha_i^{1,T'}\right] = 0$, $t = 1, ..., T'$. This makes it clear that researchers must balance the efficiency gain from larger T against the stronger strict exogeneity assumption needed for unbiasedness. This trade-off is typically overlooked by researchers.[10]

The stronger strict exogeneity assumption required in the augmented sample is testable using a Hausman test. Define $\widehat{\beta}_{FE}(s, s')$ as the FE estimate of β using only periods $t = s, ..., s'$. If $\mathtt{E}\left[\alpha_{it}|\widetilde{\mathbf{x}}_i, \alpha_i^{1,T'}\right] = 0$ (and the other standard assumptions required by FE hold), then $\widehat{\beta}_{FE}(1, T)$ and $\widehat{\beta}_{FE}(1, T')$ are both unbiased and consistent estimates of β. The former, however, is inefficient. If this assumption does not hold, then the former remains unbiased and consistent while the latter does not. Thus, the null hypothesis, $\mathtt{H_o} : \mathtt{E}\left[\alpha_{it}|\widetilde{\mathbf{x}}_i, \alpha_i^{1,T'}\right] = 0$, can be tested against the alternative that this expectation is not zero using a standard Hausman test. If the null is not rejected, then the benefit from increasing the length of the panel arguably outweighs the cost, under the usual caveats associated with a failure to reject the null hypothesis.

In practice, however, this Hausman test is not implementable as described because the researcher simply has access to panel data with T' periods and does not know the period T at which to break the sample. This is similar to a Hausman and McFadden (1984) test of the independence of irrelevant alternatives assumption in a multinomial logit model, in which case the test involves comparing estimates from the full sample with the complete choice set to estimates based on a restricted choice set but test results are not invariant to the choice of the restricted set.

Another means of assessing the role of T in FE estimation of the linear panel data model borrows from the setup in Kaddoura and Westerlund (forthcoming). The authors consider a static linear panel data model with time-varying slope heterogeneity but with constant slopes over unknown sub-periods. Here, we do the same, but for α_i rather than β. Formally, we replace α_i with α_{it} and write the model as

$$y_{it} = x_{it}\beta + \alpha_{it} + \varepsilon_{it}, \quad i = 1, ..., N;\ t = 1, ..., T \tag{4}$$

[10] The same potentially holds for changes in N. Increasing N implies that period fixed effects, if they are included in the model as they are in the popular TWFE model, may control for fewer period-specific attributes. There is a distinction between increasing N versus increasing T, however, as the former entails drawing additional observations from the *same* underlying population while the latter entails *expanding* the population to include additional time periods. We leave this discussion for future research.

where α_{it} takes on $M_i + 1$ distinct values for unit i, where

$$\alpha_{it} = \alpha_{ij(i)} \tag{5}$$

for $t = T_{j(i)-1}, ..., T_{j(i)} - 1$, $j = 1, ..., M_i + 1, M_i \in [0, T-1]$, $T_0 = 1$, and $T_{M_i+1} = T$. The notation $j(i)$ indicates the $j^{\texttt{th}}$ regime for unit i. Thus, unit i has $M_i + 1$ regimes over which α_{it} is constant, with regimes starting in periods $T_0, T_{1(i)}, ..., T_{M_i(i)}$. If $M_i = 0$ for all i, then $\alpha_{ij(i)} = \alpha_i$ and the model reverts to the standard linear panel data model. Conversely, if $M_i = T - 1$, then every period contains a unique value of α_{it}. This specification allows each unit to have a different numbers of regimes over which α_{it} is constant, as well as for a different set of time periods spanned by each regime. For example, if $T = 5$, one unit may experience two regimes where, say, α_{it} takes on one value for $t = 1, 2$ and a different value for $t = 3, 4, 5$, whereas another unit experiences three regimes where, say, α_{it} takes on one value for $t = 1$, another value for $t = 2, 3$, and yet another value for $t = 4, 5$.

With this setup, $\widehat{\beta}_{FE}$ is obtained by applying POLS to the following estimating equation

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + \left(\alpha_{it} - \sum_{j=1}^{M_i+1} \omega_{ij}\alpha_{ij} \right) + \ddot{\varepsilon}_{it}, \tag{6}$$

where

$$\omega_{ij} = \frac{T_j - T_{j-1}}{T},$$

is the share of sample periods where $\alpha_{it} = \alpha_{ij}$. Equation (6) makes it clear that mean-differencing does not fully remove α_{it} if $M_i > 1$. It only removes the part of α_{it} that is constant over the entire sample, $\alpha_i^{1,T}$. As such, unbiasedness of the FE estimator requires $\texttt{E}\left[\alpha_{it} + \varepsilon_{it} | \mathbf{x}_i, \alpha_i^{1,T}\right] = 0$. In words, we now require strict exogeneity of x with respect to both α_{it} and ε_{it} conditional on $\alpha_i^{1,T}$.

When $\hat{\beta}_{FE}(1,T)$ is biased due to a failure of the strict exogeneity assumption in the presence of time-varying, unit-specific unobserved heterogeneity, alternative estimators may perform better. We consider several possibilities. In Section 3.2 we discuss the existing FD and TFD estimators. In Section 3.3 we present several novel estimation algorithms based on rolling regressions. In Section 3.4 we discuss a few other existing estimators.

3.2 First-Differences and Twice First-Differenced Estimators

The FE estimator is identical to the FD estimator when $T = 2$. Moreover, irrespective of T, Wooldridge (2010) and others state that the strict exogeneity assumption is also identical across the estimators.[11] Technically, however, the strict exogeneity assumption required by FD is weaker: For the model in Equation (1), FD requires $\texttt{E}\left[\varepsilon_{it} | x_{i,t-1}, x_{it}, x_{i,t+1}, \alpha_i\right] = 0$ for all t.

But a more substantive difference arises as T increases. Consider the following DGP

$$y_{it} = x_{it}\beta + \alpha_i^{s,s'} + \varepsilon_{it}, \quad i = 1, ..., N; \; t = s, ..., s' \tag{7}$$

where $s < s'$. This is the standard linear panel data model except defined over a generic period defined

[11] Assumption FD.1 is "Same as Assumption FE.1" (Wooldridge, 2010, p. 316).

by s and s'. Given the definition of $\alpha_i^{s,s'}$, the unit fixed effect will capture fewer attributes and ε_{it} more attributes as $s' - s \to \infty$. Thus, setting $s' = s + 1$ yields an unbiased estimate of β under the weakest possible strict exogeneity condition. Specifically, retaining only periods s and $s + 1$, the FD estimator is obtained by estimating

$$\Delta y_{i,s+1} = \Delta x_{i,s+1}\beta + \Delta\varepsilon_{i,s+1} \tag{8}$$

using OLS, where Δ is the first-difference operator. Unbiasedness requires $\mathtt{E}\left[\varepsilon_{i,s}|x_{i,s}, x_{i,s+1}, \alpha_i^{s,s+1}\right] = 0$ and $\mathtt{E}\left[\varepsilon_{i,s+1}|x_{i,s}, x_{i,s+1}, \alpha_i^{s,s+1}\right] = 0$. By focusing on only two consecutive periods, the strict exogeneity assumption is the weakest possible because $\alpha_i^{s,s+1}$ removes the most unobserved heterogeneity possible.

Given a panel of length T, stacking (8) using $s = 1, ..., T - 1$ and estimating the model using POLS yields the FD estimator. When there is unobserved heterogeneity that is not time-invariant over the full sample period, but is time-invariant over different regimes, FD is likely to be superior to FE. This arises because each first-difference removes all unit-specific attributes that are constant over the regime spanning any two consecutive periods.

The stronger strict exogeneity assumption required for FE can be tested. Following Laporte and Windmeijer (2005), a test for the for the failure of these additional restrictions is given by a test of equality of β_{FE} and β_{FD}. This can be accomplished by estimating the following stacked regression via POLS

$$\begin{pmatrix} \ddot{y}_{it} \\ \Delta y_{it} \end{pmatrix} = \begin{pmatrix} \ddot{x}_{it} \\ \Delta x_{it} \end{pmatrix}\beta + \begin{pmatrix} 0 \\ \Delta x_{it} \end{pmatrix}\varphi + u_{it} \tag{9}$$

and testing $\mathtt{H}_o : \varphi = 0$ using a two-sided alternative and a cluster-robust estimator of the covariance matrix (Papke and Wooldridge, forthcoming; Spierdijk, 2023). Of course, rejection of the null may also occur under other sources of mis-specification that differentially affect FE and FD. Moreover, choosing an estimator on the basis of this test risks introducing problems associated with pre-testing (Papke and Wooldridge, forthcoming).

Let us consider a few examples. First, we assess the impact of an increase in the length of the panel from T to T' as in Section 3.1. When T increases, the model changes from Equation (1) to Equation (3), where α_{it} is given by

$$\alpha_{it} \equiv \begin{cases} \alpha_i^{1,T} - \alpha_i^{1,T'} & \text{if } t \leq T \\ \alpha_i^{T+1,T'} - \alpha_i^{1,T'} & \text{if } t > T \end{cases} \tag{10}$$

Recalling that $\eta_{it} \equiv \alpha_{it} + \varepsilon_{it}$, the strict exogeneity assumption required for FE is

$$\mathtt{E}\left[\alpha_{it} + \varepsilon_{it}|\widetilde{\mathbf{x}}_i, \alpha_i^{1,T'}\right] = 0 \quad \forall t \ . \tag{11}$$

In comparison, first-differencing the model specified in Equations (3) and (10) yields

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta\eta_{it}. \tag{12}$$

But since

$$\Delta\eta_{it} = \begin{cases} \Delta\varepsilon_{it} & \text{if } t \neq T + 1 \\ \Delta\varepsilon_{it} + \alpha_i^{T+1,T'} - \alpha_i^{1,T} & \text{if } t = T + 1 \end{cases} \tag{13}$$

it follows that for FD to be unbiased only requires

$$\begin{aligned} &\mathtt{E}\left[\varepsilon_{it}|x_{i,t+1},x_{it},x_{i,t-1},\alpha_i^{1,T},\alpha_i^{T+1,T'},\alpha_i^{1,T'}\right]=0 \quad \forall t, \text{ and} \\ &\mathtt{E}\left[\alpha_i^{T+1,T'}-\alpha_i^{1,T}|x_{it},x_{i,t-1},\alpha_i^{1,T'}\right]=0 \qquad\qquad \text{if } t=T+1\ . \end{aligned} \tag{14}$$

We can relax the second requirement in Equation (14) by including an intercept in the first-differenced specification in Equation (12). This alters the requirement to

$$\mathtt{E}\left[\alpha_i^{T+1,T'}-\alpha_i^{1,T}|x_{it},x_{i,t-1},\alpha_i^{1,T'}\right]=c \quad \text{if } t=T+1 \tag{15}$$

for some unknown c.[12]

Comparing Equations (11) with (14) and (15) reveals a second difference in the requirements for FE and FD to be unbiased: Whereas FE requires x_{it} to be conditionally stochastically independent of *all* values of α_{it}, FD requires only $x_{i,T}$ and $x_{i,T+1}$ to be conditionally stochastically independent of values of α_{iT} and $\alpha_{i,T+1}$. This weaker requirement on the FD estimator makes it an attractive estimator relative to the FE estimator when the goal is one of causal inference.

The distinction between the requirements for FE versus FD are more than a mere technicality. For a second example, we return to the setup in Equations (4) and (5) which is borrowed from Kaddoura and Westerlund (forthcoming). In this case, α_{it} changes more frequently than just in period T. Yet, FD purges *all* unit-specific attributes that are constant over the regime spanning any two consecutive periods. Specifically, first-differencing the model in Equations (4) and (5) yields Equation (12) except the error term is now

$$\Delta\eta_{it}=\begin{cases}\Delta\varepsilon_{it} & \text{if } t\neq T_{j(i)} \\ \Delta\varepsilon_{it}+\alpha_{i,j+1}-\alpha_{ij} & \text{if } t=T_{j(i)}\end{cases} \tag{16}$$

for $j=1,...,M_i$, where $T_{j(i)}$ are the regime "seams" during the sample period for unit i (i.e., they denote the initial period of each regime). Here, unbiasedness of FD requires $x_{i,T_{j(i)}-1}$ and $x_{i,T_{j(i)}}$ to be conditionally stochastically independent *only* from unobserved attributes that are time-varying across seam j, $j=1,...,M_i$, but FE requires x_{it} to be conditionally stochastically independent from *all* unobserved attributes that vary at any point in the sample period.

As a third example, we consider a the situation where α_{it} follows a random walk. The DGP is

$$y_{it}=x_{it}\beta+\alpha_{it}+\varepsilon_{it}, \quad i=1,...,N;\ t=2,...,T \tag{17}$$

where α_{it} is given by

$$\alpha_{it}=\alpha_{i,t-1}+\varphi_{it} \tag{18}$$

and φ_{it} is a mean zero error term. This case is interesting for two reasons. First, α_{it} is *never* constant from one period to the next if $\mathtt{Var}(\varphi)>0$. Second, the model becomes the standard linear panel model as $\mathtt{Var}(\varphi)\to 0$.

[12] Including an intercept in the first-differenced equation is likely to be preferred since in the standard linear panel model it is not assumed that $\mathtt{E}[\alpha_i]=0$. As such, we do not wish to assume that $\mathtt{E}[\Delta\alpha_{it}]=0$.

First-differencing Equation (17) yields

$$\Delta y_{it} = \Delta x_{it}\beta + \varphi_{it} + \Delta\varepsilon_{it}. \tag{19}$$

POLS estimation of Equation (19) yields unbiased estimates if $\mathtt{E}\left[\varphi_{it} + \Delta\varepsilon_{it}|\Delta x_{it}\right] = 0$ (or c if an intercept is included, in which case α_{it} is allowed to follow a common deterministic trend). The FE estimator, however, requires $\mathtt{E}\left[\ddot{\eta}_{it}|\ddot{x}_{it}\right] = 0$, where

$$\begin{aligned} \ddot{\eta}_{it} &= \ddot{\alpha}_{it} + \ddot{\varepsilon}_{it} \\ &= \ddot{\alpha}_{i,t-1} + \ddot{\varphi}_{it} + \ddot{\varepsilon}_{it}\ . \end{aligned} \tag{20}$$

Thus, FE requires x_{it} to be conditionally stochastically independent of *all* values of α_{it} and ε_{it}, which implies the full history of φ_{it}. In contrast, FD requires independence from only the contemporaneous and lead values of φ_{it} and the lagged, contemporaneous, and lead values of ε_{it}.

Finally, consider another case where α_{it} is never constant across time periods, but now follows a unit-specific time trend as in Mundlak (1978). The DGP is given by Equation (17) where α_{it} is

$$\alpha_{it} = \alpha_i^0 + \alpha_i^1 t. \tag{21}$$

In this case, the model becomes the standard linear panel model when $\alpha_i^1 = 0$ for all i.[13]

First-differencing yields

$$\Delta y_{it} = \Delta x_{it}\beta + \alpha_i^1 + \Delta\varepsilon_{it}. \tag{22}$$

POLS estimation of Equation (22) yields unbiased estimates if $\mathtt{E}\left[\alpha_i^1 + \Delta\varepsilon_{it}|\Delta x_{it}\right] = 0$ (or c if an intercept is included). The FE estimator, however, requires $\mathtt{E}\left[\ddot{\eta}_{it}|\ddot{x}_{it}\right] = 0$, where

$$\ddot{\eta}_{it} = \alpha_i^1 \ddot{t} + \ddot{\varepsilon}_{it}. \tag{23}$$

Thus, FE requires x_{it} to be conditionally stochastically independent of α_i^1 and *all* values of ε_{it}. In contrast, FD requires independence from the α_i^1 and the lagged, contemporaneous, and lead values of ε_{it}.

Of course, it is well-known that estimating the model with unit-specific fixed effects and time trends by POLS after *twice* first-differencing is unbiased under what, technically, is still a weaker strict exogeneity assumption than estimating Equation (22) by FE as both α_i^0 and α_i^1 are removed from the estimating equation. This suggests that even in the prior cases, the TFD estimator may outperform both FE and FD in some practical situations if α_{it} follows and approximately linear time trend. For this reason, we also consider the TFD estimator.

In sum, given a panel of length T, FE requires strict exogeneity of x_{it} with respect to α_{it} (and ε_{it}) conditional on $\alpha_i^{1,T}$, whereas FD requires strict exogeneity of x_{it} with respect to only the change in α_{it} (and the lagged, contemporaneous, and lead values of ε_{it}). By removing the part of α_{it} that is persistent from one period to the next, the strict exogeneity assumption required by FD and TFD is at worst equivalent to that required by FE and at best (and more likely) much weaker. Moreover, the difference in the strengths

[13] The model also reduces to the standard linear panel data model if the model includes time fixed effects and $\alpha_i^1 = a$ for all i and some constant a.

of the respective strict exogeneity assumptions is likely to increase with T. This is consistent with the quote from Mundlak (1978) mentioned in Section 1, which points to a fundamental difference between linear panel data models over short versus long time intervals.

In light of the foregoing, one might wonder why the FE estimator is the default in empirical studies. We believe the answer lies in the fact that the FE estimator is more efficient than the FD estimator when ε_{it} is conditionally homoskedastic and not serially correlated (Aquaro and Čížek, 2013). In fact, it is the asymptotically efficient estimator among the class of estimators relying on strict exogeneity under these conditions (Wooldridge, 2010, Section 10.6). The focus on efficiency, however, both comes at the neglect of the potential for bias and inconsistency and is at odds with the goal of causal inference that is often the object of applied work following the Credibility Revolution (Angrist and Pischke, 2010).

3.3 Rolling Estimators

Building on the logic in Sections 3.1 and 3.2, we introduce a new estimation algorithm which we refer to as the rolling first-differences (RFD) estimator, can be obtained by *separately* estimating Equation (7) for $s = 1, ..., T-1$ with $s' = s+1$. This yields $T-1$ estimates of β, $\widehat{\beta}_{FD}(s, s+1)$ (which are identical to $T-1$ FE estimates). If an intercept is included in the FD equations, then $T-1$ estimates of c are also obtained. A final estimate can be obtained via model averaging

$$\widehat{\beta}_{RFD} = \sum_{s=1}^{T-1} \nu_s \widehat{\beta}_{FD}(s, s+1), \tag{24}$$

where ν_s is the weight given to $\widehat{\beta}_{FD}(s, s+1)$, as shown in the last line of Algorithm 1, and the weights sum to one (and similarly for c).

The preferred approach to combining the $T-1$ estimates is by MDE using the optimal weight matrix (Wooldridge, 2010, Section 14.5). In this case, ν_s is a function of the variances and covariances of the individual estimates. In Algorithm 1 we simplify this computation by ignoring the covariances (see, e.g. Mullahy, 2016).[14] Note, it is trivial to alter Algorithm 1 to obtain a rolling twice first-differenced (RTFD) estimator (with or without an intercept), where $T-2$ TFD estimates, $\widehat{\beta}_{TFD}(s, s+2)$, are averaged.

There are two question at this point. First, is there an advantage to RFD over FD (or RTFD over TFD), since the RFD and FD estimators eliminate the same unobserved attributes? As mentioned in Section 3.2, FE is asymptotically efficient among the class of estimators relying on strict exogeneity when ε_{it} is conditionally homoskedastic and not serially correlated. RFD is based on separate FD estimates over two time periods rather than a single FD estimate, where FD and FE are equivalent when $T = 2$. This suggests a possible efficiency gain when using RFD instead of FD. Offsetting this is the inefficiency that likely results from the use of rolling regressions if the unit fixed effects are constant for all units over the full sample periods (Cai and Juhl, forthcoming). We explore this possibility in Section 4.

Another advantage of RFD relative to FD, however, arises with parameter heterogeneity. As discussed in Section 2, there is reason to suspect in many empirical applications that not only might α_i not be time-invariant as T grows, but β as well. If one suspects this to be the case, the separate estimates, $\widehat{\beta}_{FD}(s, s+1)$, can be examined and equality of the parameters can be tested by estimating the following stacked regression

[14] Alternatively, one might use a clustered bootstrap to obtain the empirical variances and covariances of the estimates.

via POLS

$$\begin{pmatrix} \hat{\beta}_{FD}(1,2)) \\ \hat{\beta}_{FD}(2,3) \\ \vdots \\ \hat{\beta}_{FD}(T-1,T) \end{pmatrix} = \begin{pmatrix} \beta_{RFD(1,2)} \\ \beta_{RFD(1,2)} \\ \vdots \\ \beta_{RFD(1,2)} \end{pmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{pmatrix} 0 \\ \varphi_2 \\ \vdots \\ \varphi_{T-1} \end{pmatrix} + \zeta_s \tag{25}$$

weighted by the inverse of the standard errors and testing $\mathtt{H}_o : \varphi = 0$ using a two-sided alternative. If one suspects cross-sectional (with or without temporal) heterogeneity in β, where β_{gt} is the vector of parameters for group g at time t, this can also be incorporated into the RFD estimator by applying the various group estimators at each FD step. This accomplishes the same result as in Boldea, Drepper and Gan (2020) without have to test for the temporal structural breaks.[15]

A final advantage of RFD relative to FD is that RFD precludes the need to contemplate the use of clustered standard errors. As discussed in Abadie et al. (2022), decisions over clustering when using longitudinal data are not clear. Since RFD reduces the estimation to $T-1$ cross-sectional regressions, however, this obviates the need to cluster.

Algorithm 1 Rolling First Differences (RFD) Estimator

1: **while** $s = 1, ..., T-1$ **do**
2: Apply the FD estimator (with or without an intercept in the differenced equation) using data from two periods, $\{y_{it}, x_{it}\}_{i=1,...,N;t=s,s+1}$.
3: Collect the coefficient estimates and robust ariance-covariance matrix: $\widehat{\beta}_{FD}(s, s+1)$, $\widehat{\Sigma}_{\beta_{FD}}(s, s+1)$, where $(s, s+1)$ refers to the sample periods used in the estimation.
4: **end while**
5: Test equality of $\beta_{FD}(s, s+1)$ $\forall s$ (if desired)
6: Estimate β using the minimum distance estimator (if desired)

$$\hat{\beta}_{RFD} = \arg\min_{\beta} \quad \left(\hat{\boldsymbol{\beta}} - \beta\right)' W^{-1} \left(\hat{\boldsymbol{\beta}} - \beta\right),$$

where

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_{FD}(1,2) \\ \vdots \\ \widehat{\beta}_{FD}(T-1,T) \end{bmatrix}; \quad W = \begin{bmatrix} \widehat{\Sigma}_{\beta_{FD}}(1,2) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \widehat{\Sigma}_{\beta_{FD}}(T-1,T) \end{bmatrix}.$$

Or, alternatively, estimate via OLS

$$\hat{\beta}_{FD}(s, s+1) = \beta_{RFD} + \zeta_s$$

weighted by the inverse of $\sqrt{\mathtt{Var}\left(\widehat{\beta}_{FD}(s, s+1)\right)}$.

[15] A more formal comparison with Boldea, Drepper and Gan (2020) we leave for future research.

Second, is there an efficiency gain by rolling over periods spanning more than two periods? Algorithm 2 outlines a rolling fixed effects (RFE) estimator that generalizes RFD to the case where one rolls through FE estimates applied to j consecutive periods, $j \geq 2$. The additional restrictions that are required for RFE with $j > 2$ can be tested using a stacked regression as above. Specifically, a test for the for the failure of these additional restrictions is given by a test of equality of $\beta_{RFE(j)}$ and $\beta_{RFE(j')}$ for any $2 \leq j < j'$, where $\beta_{RFE(j)}$ refers to the estimator obtained using Algorithm 2 for a particular choice of j, can be accomplished by estimating the following stacked regression via POLS

$$\begin{pmatrix} \hat{\beta}_{FE}(s, s+(j-1)) \\ \hat{\beta}_{FE}(s, s+(j'-1)) \end{pmatrix} = \begin{pmatrix} \beta_{RFE(j)} \\ \beta_{RFE(j)} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \varphi + \zeta_s \tag{26}$$

weighted by the inverse of the standard errors and testing $\mathtt{H}_o : \varphi = 0$ using a two-sided alternative. Failure to reject the null suggests an efficiency gain from setting j to j' in Algorithm 2 (subject to the usual caveats concerning pre-testing). Moreover, we can stack more than two values of j (as in Spierdijk (2023)) and perform several different tests of equality using

$$\begin{pmatrix} \hat{\beta}_{FE}(s, s+(j_1-1)) \\ \hat{\beta}_{FE}(s, s+(j_2-1)) \\ \vdots \\ \hat{\beta}_{FE}(s, s+(j_J-1)) \end{pmatrix} = \begin{pmatrix} \beta_{RFE(j_1)} \\ \beta_{RFE(j_1)} \\ \vdots \\ \beta_{RFE(j_1)} \end{pmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{pmatrix} 0 \\ \varphi_{j_2} \\ \vdots \\ \varphi_{j_J} \end{pmatrix} + \zeta_s \tag{27}$$

as above. One could also include the standard FE estimator based on the full sample period in the stacked regression as well. In the interest of brevity, we do not consider the RFE estimator in Section 4.

Algorithm 2 Rolling Fixed Effects (RFE) Estimator

1: Choose the number of periods to be pooled, j, where $j \in [3, T]$
2: **while** $s = 1, ..., T - (j-1)$ **do**
3: Apply the FE estimator using data from j periods, $\{y_{it}, x_{it}\}_{i=1,...,N;t=s,s+(j-1)}$.
4: Collect the coefficient estimates and robust variance-covariance matrix: $\widehat{\beta}_{FE}(s, s+(j-1))$, $\widehat{\Sigma}_{\beta_{FE}}(s, s+(j-1))$, where $(s, s+(j-1))$ refers to the sample periods used in the estimation.
5: **end while**
6: Estimate β using the minimum distance estimator (MDE)

$$\hat{\beta}_{RFE(j)} = \arg\min_{\beta} \quad \left(\hat{\boldsymbol{\beta}} - \beta\right)' W^{-1} \left(\hat{\boldsymbol{\beta}} - \beta\right),$$

where

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_{FE}(1, j) \\ \vdots \\ \widehat{\beta}_{FE}(T-(j-1), T) \end{bmatrix}; \quad W = \begin{bmatrix} \widehat{\Sigma}_{\beta_{FE}}(1, j) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \widehat{\Sigma}_{\beta_{FE}}(T-(j-1), T) \end{bmatrix}.$$

Or, alternatively, estimate via OLS

$$\hat{\beta}_{FE}(s, s+(j-1)) = \beta_{RFE(j)} + \zeta_s$$

weighted by the inverse of $\sqrt{\texttt{Var}\left(\widehat{\beta}_{FE}(s, s+(j-1))\right)}$.

3.4 Alternative Estimators

Prior to turning to empirics, discussion of a few alternative estimators is warranted. First, as mentioned above, when α_{it} changes only occasionally, an example of which is given in Equation (10), the model is one with a structural break in the unit fixed effects. Thus, one might consider applying tests of (partial) structural breaks with unknown break dates. This is potentially more efficient than FD, RFD, and RTFD since it entails using FE over each regime around the break dates, and potentially more efficient than RFE if the "true" break dates can be estimated. We do not consider this approach here beyond what we already propose with RFE, however, since it requires the regimes over which relevant unobserved heterogeneity to be identical for all units for computational reasons and even then feasibility is a concern. In contrast, FD, RFD, and RTFD remove all unobserved heterogeneity that is constant across any pair of consecutive periods. As such, relevant unobserved heterogeneity need not be constant over the same regime for all units.

Allowing for structural breaks in the unit fixed effects at different times across units has parallels to the block-concentrated outliers considered in Bramati and Croux (2007), where the authors consider outliers to be more common in some units than others. In our case, relevant unobserved heterogeneity may be time-invariant over the full sample period for some units, but highly volatile for a few units. FD and RFD

are likely more robust in this situation than estimators that impose common break dates.

A second alternative is the interactive fixed effects (IFE) estimator proposed in Bai (2009), given by

$$y_{it} = x_{it}\beta + \lambda_t'\alpha_i + \varepsilon_{it}, \quad i = 1, ..., N; \; t = 1, ..., T \tag{28}$$

where λ_t is a vector of common factors. This specification allows for unit-specific responses to common temporal shocks as well as time-varying effects of time-invariant unit-specific unobserved heterogeneity. The additional flexibility of the IFE estimator may offer advantages over the traditional FE or FD estimators in certain situations such as the inclusion of unit-specific time trends in Mundlak (1978). In a more general setting with unknown, unit-specific structural breaks in the fixed effects, however, IFE imposes a common factor structure for all units and is therefore generally less robust than FD, RFD, and RTFD. That said, the estimator can accommodate more than one factor λ_t, and thus we consider one- and two-factor IFE estimators in Section 4.

Finally, other alternatives include panel estimators that are robust to outliers, such as the MM-estimator (Yohai, 1987) and the S-estimator (Rousseeuw and Yohai, 1984). We consider both these estimators and the least trimmed squares (LTS) estimator discussed in Aquaro and Čížek (2013) as a structural break in the unit fixed effects may mimic outliers as typically defined based on the covariates. As all three perform poorly in the simulations we consider, results from these estimators are omitted for brevity.

3.5 Asymptotic Comparisons

Before turning to simulations and replications, Pesaran and Zhou (2018) offer an interesting framework that is useful for thinking about the issue in this paper from a different perspective. The authors compare the asymptotic efficiency of POLS and FE (as $N \to \infty$ for fixed T) depending on the *fraction* of sample units with a unit fixed effect. Specifically, their model is given by

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}, \quad i = 1, ..., N; \; t = 1, ..., T \tag{29}$$

where

$$\begin{aligned} \alpha_i &= \alpha + \zeta_i, \quad i = 1, ..., N \\ \zeta_i &= \begin{cases} \widetilde{\zeta}_i & \text{if } i = 1, ..., \left[N^\delta\right] \\ 0 & \text{if } i = \left[N^\delta\right] + 1, ..., N \end{cases} \end{aligned} \tag{30}$$

$\delta \in [0, 1]$, and $\left\{\widetilde{\zeta}_i, i = 1, ..., \left[N^\delta\right]\right\}$ is a sequence of random variables with mean zero and finite variance. Equation (30) permits unit-specific fixed effects for only $\left[N^\delta\right]$ units in the sample; the remainder have a common intercept.

Under a certain assumptions, the authors show that $\widehat{\beta}_{POLS}$ is consistent for β if $\delta < 1$ and is asymptotically normal if $\delta < 0.5$. Moreover, POLS is asymptotically more efficient than FE if $\delta < 0.5$. In this case, the fraction of the sample with unit fixed effects goes to zero as $N \to \infty$ for fixed T such that $\texttt{plim}_{N\to\infty}\widehat{\beta}_{POLS} \to \beta$ and $\texttt{AsyVar}\left(\widehat{\beta}_{POLS}\right) < \texttt{AsyVar}\left(\widehat{\beta}_{FE}\right)$.

While our discussion to this point has focused solely on the finite sample requirements of FE, FD, TFD, and our rolling estimators, we are also interested in the asymptotic efficiency of these estimators (as $N \to \infty$

for fixed T) as the *fraction* of units in the sample with time-varying unit fixed effects changes. The model is given by

$$y_{it} = x_{it}\beta + \alpha_{it} + \varepsilon_{it}, \quad i = 1, ..., N; \; t = 1, ..., T \tag{31}$$

where

$$\begin{aligned} \alpha_{it} &= \alpha_i + \zeta_{it}, \quad i = 1, ..., N \\ \zeta_{it} &= \begin{cases} \widetilde{\zeta}_{it} & \texttt{if } i = 1, ..., \left[N^{\delta}\right] \\ 0 & \texttt{if } i = \left[N^{\delta}\right] + 1, ..., N \end{cases} \end{aligned} \tag{32}$$

and everything else is defined analogously to the above. Equation (32) permits time-varying unit-specific fixed effects for only $\left[N^{\delta}\right]$ units in the sample; the remainder have a time-invariant unit fixed effect. As in Pesaran and Zhou (2018), it is likely that $\widehat{\beta}_{FE}$ and $\widehat{\beta}_{FD}$ are both consistent, but $\widehat{\beta}_{FE}$ is more efficient, for sufficiently small δ as $N \to \infty$ for fixed T. We explore this in Section 4.

4 Examples

We begin by briefly discussing the results of Monte Carlo simulations that illustrate the foregoing. We then discuss the results of two replications.

4.1 Monte Carlo Study

Details and full results from our Monte Carlo study are relegated to Appendix A. We compare several estimators: (i) FE, (ii) FD, (iii) TFD, (iv) RFD, (iv) RTFD, and (v) IFE (with one and two factors).[16] All FD and TFD estimators are used with and without an intercept. To start, we use four experimental designs. The first two designs are based on Equation (10). In both designs we allow for two structural breaks in α_{it}. In the first design, the break dates are common to all units. In the second design, the break dates are unit-specific. The next two designs are based on Equations (18) and (21), where the former allows the relevant unobserved heterogeneity to follow a random walk and the latter allows for the relevant unobserved heterogeneity to follow unit-specific time trends. In simulations allowing for unit-specific time trends, we also consider an additional estimator, twice FD, since this estimator is unbiased in this case.

In these four designs, we set $N = 500$ and consider three panel lengths, $T = 5, 10, 20$. We also vary the standard deviation of the temporal heterogeneity, σ, in α_{it} from zero (such that $\alpha_{it} = \alpha_i$) to 0.5. We evaluate the estimators in terms of the bias and root mean squared error (RMSE) of the estimated coefficient on a single, unit- and time-varying covariate, where the true parameter value is $\beta = 1$, from 200 replications for each experimental design and configuration.

We obtain a several interesting results from these designs. First, when $\alpha_{it} = \alpha_i$ for all time periods, the performance of all estimators is essentially the same. The lone exception is in the case of unit-specific time trends (DGP3), where the IFE estimators have similar bias but a much larger RMSE. Second, the performance of the FE estimator (and IFE1 and IFE2) worsens as temporal heterogeneity in the unit fixed effects increases. When $T = 5$, the deterioration starts when the standard deviation of α_{it} is 0.2 or higher.

[16] We also consider robust estimators from the literature on outliers in fixed effect models such as the MM-estimator and S-estimator. The results were not encouraging and thus we omit them here.

When $T = 10$ or 20, the deterioration starts when the standard deviation of α_{it} is 0.1 or higher. Thus, for panel lengths that are now the norm in empirical research, the assumption of time-invariant unobserved heterogeneity is crucial.

Third, as temporal heterogeneity in the unit fixed effects increases, a clear ranking in terms of performance emerges. In all experimental designs, FE and IFE1 perform the worst. The performances of FD, TFD, RFD, and RTFD are comparable until the standard deviation of α_{it} exceeds 0.2. When the standard deviation exceeds this level, TFD and RTFD perform marginally better than FD and RFD. Thus, when the time-varying fixed effects vary significantly over time, TFD performs better due to its removal of all unit-specific heterogeneity that is constant across two consecutive periods, as well as all unit-specific heterogeneity that is constant across two consecutive first-differences. The performance of IFE2 varies across the experimental designs. When α_{it} arises due to structural breaks (DGP1 and DGP2) or it follows a random walk (DGP4), IFE2 does better than FE and IFE1, but noticeably worse than the remaining estimators. When α_{it} follows a unit-specific time trend, IFE2 achieves comparable performance to TFD and RTFD until the standard deviation of α_{it} reaches 0.5. Even at this level of temporal heterogeneity, IFE2 still outperforms FD and RFD.

Next, we re-do the previous four experimental designs except now mirroring the setup in Pesaran and Zhou (2018). In all cases, we set $T = 10$ and $\sigma = 0.5$, but we vary the proportion of the units with variable α_{it}. Specifically, δN units have $\alpha_{it} \neq \alpha_i$, where $\delta = 0.45, 0.475, 0.50, 0.525, 0.55, 0.60$ and $N = 100, 500, 5000$. Two main results emerge. First, in the two designs allowing for structural breaks in α_{it} for those units with time-varying fixed effects, FE performs best in terms of RMSE for sufficiently for all N when $\delta \leq 0.5$. When $N = 100$ FE has the smallest RMSE when $\delta = 0.55$. However, FE performs poorly in terms of bias. Thus, when only a few units have time-varying fixed effects, FE is much more efficient than the alternatives considered.

Second, in the other two designs—unit-specific time trends and time-varying fixed effects following a random walk—both FE and IFE1 perform very poorly in terms of bias and RMSE for all N. In the case of unit-specific time trends, the performance of all other estimators have essentially identical RMSE. In the random walk case, IFE2, FD, and RFD perform the best, especially with smaller δ. TFD and RTFD perform marginally worse, but still much better than FE and IFE1.

Overall, the simulation results suggest that reliance on the FE estimator offers small efficiency gains when the model is correctly specified: all relevant unit-specific unobserved heterogeneity is time invariant. In some cases, FE continues to perform well in terms of RMSE when the number of units with time-varying unit fixed effects is small. However, in other cases, FE performs quite poorly even when the majority of units have time invariant fixed effects. Alternative estimators such as FD, RFD, TFD, and RTFD perform nearly as well in the cases where FE performs best, yet are much more robust to the presence of time-varying unit-specific heterogeneity for all units or even just some. As such, researchers should avoid relying on FE because they believe that unobserved heterogeneity is time invariant for "most" units.

4.2 Replications

To illustrate the foregoing using real-world applications, we conduct two replication exercises. We first follow Imai and Kim (2019) by replicating Rose (2004), who look at the relationship between country-level participation in the General Agreement on Tariffs and Trade (GATT, which became the World Trade

Organization in 1995) and the extent of participation in international trade, as well as the subsequent comment on Rose (2004) by Tomz, Goldstein and Rivers (2007). We then replicate the results in James (2015), who looks at the relationship between state-level resource-based government revenue on the one hand and taxation, spending, and savings at the state level on the other hand. In both replications, we show that the choice of estimator—FE versus FD or RFD—can have important consequences not just for the magnitude of the coefficients of interest, but also for their sign and significance. We further present the results of iterated FE estimators with one (IFE1) or two factors (IFE2).

4.2.1 Imai and Kim (2019)

Table I replicates results in Imai and Kim (2019), who in turn replicate the analysis in Rose (2004) and Tomz, Goldstein and Rivers (2007) that look at the relationship between country-level GATT participation and bilateral trade. The results in Table I use a sample of 175 countries over the period 1948-1994. The estimates here rely on dyad-specific fixed effects (i.e., fixed effects at the level of trading country pairs) in a standard "gravity" model specification; Rose (2004) and Tomz, Goldstein and Rivers (2007) also consider separate country-specific fixed effects. In Tomz, Goldstein and Rivers (2007) and many other studies estimating gravity models justify the inclusion of country fixed effects following the theoretical model in Anderson and van Wincoop (2003), which "implies the presence of a 'multilateral resistance' term that can be approximated using country and time fixed effects." The model in Anderson and van Wincoop (2003), however, indicates that the multilateral resistance terms, which are functions of underlying prices, are country- and time-specific. Thus, country- or dyad-specific fixed effects cannot control for multilateral resistance.

Each column in Table I denotes a different estimator and each panel denotes a different specification. In Panels A and B the variable of interest (or treatment variable) is formal GATT membership, where Panel B includes year FEs. In Panels C and D, the variable of interest is informal GATT membership, where Panel D includes year FEs. In each panel, we show the estimated coefficient on GATT membership as well as the result of a Laporte and Windmeijer (2005) test of equality of coefficients between the FE and FD specifications.

The results are striking. First, the FE and FD estimates are statistically different at the $p < 0.01$ level in all four panels. In Panels A and B, however, the difference in the point estimates is small in magnitude. This is not the case in Panels C and D. Now, the FD point estimates are roughly 2.5 to 4.5 times smaller. Thus, the choice between FE and FD fundamentally alters the conclusions. Second, the twice-differenced estimator, the various RFD estimators, and the IFE2 estimator are in line with the FD estimates. The IFE1 estimates are most similar to FE.

Finally, whereas the FE estimates are economically different in Panels A and B versus Panels C and D, the FD estimates are qualitatively the same across all four panels. This is a critical finding as it resolves a puzzle from these earlier papers. Tomz, Goldstein and Rivers (2007, p. 2011) write: "It is difficult to explain why the effect should be larger for nonmember participants than formal members, given that both had essentially the same rights and obligations." Chang and Lee (2011) similarly report larger effects for nonmbember participants using matching estimators. FD does not find any such difference in the effects. This is consistent with Eicher and Henn (2011), who also find no difference when accounting for more commonly omitted attributes.

4.2.2 James (2015)

Tables II and III replicate results in James (2015), who looks at the relationship between state-level resource-based government revenue on the one hand and taxation, spending, and savings at the state-level on the other hand using annual data from 1958-2008. The goal is to test a theoretical model that predicts that a benevolent government will reduce taxes and increase both spending and savings in response to an exogenous increase in resource-based government revenue. Despite the sample covering 51 years, James (2015, p. 243) claims that "time-invariant, state-specific characteristics such as average population density, political preferences, wealth, unemployment, culture, and institutional quality are captured by state fixed effects." Clearly, this is not the case. For example, Caughey and Warshaw (2016) shows that political preferences as measured by state policy liberalism varies tremendously over the past century, at least for select states. Frank (2009) documents the intrastate temporal variation in income inequality. The unemployment rate in California varied from a low of 4.8% to a high of 11.1% between 1976 and 2008.[17]

The results in Table II use the full sample of all 50 states. The results in Table III omit Alaska but are otherwise identical. Each column in Tables II and III denotes a different estimator, and each panel denotes a different dependent variable. In each panel, we show the estimated coefficient on resource-based government revenue as well as the result of a Laporte and Windmeijer (2005) test of equality of coefficients between the FE and FD specifications.

Our analysis leads to a few salient findings. First, the FE and FD estimates in Table II are statistically different at the $p < 0.05$ level in three out of five cases. The estimates are even of opposite sign for education expenditures (Panel D). Moreover, while statistically different at only the $p < 0.11$ level, the FE and FD estimates are also of opposite sign for nonresource revenue (Panel A), only the FD estimate is statistically significant (at the $p < 0.01$ level), and the FD estimate is more than three times as large in absolute value. Second, the divergence between the FE and FD estimates is even more pronounced in Table III when Alaska is omitted. While the FE estimates are statistically significant at the $p < 0.05$ level for all outcomes, the FE and FD estimates continue to be statistically different at the $p < 0.05$ level in three out of five cases. Moreover, the FD estimates are statistically indistinguishable from zero in three out of five cases. Thus, the choice between FE and FD fundamentally alters the conclusions.

Third, the twice-differenced estimator and the various RFD estimators are much more similar to the FD estimates than the FE estimates. In Table II, in particular, FE is a clear outlier among this group of estimators. Fourth, the IFE estimates are a mixed bag, but they tend to look closer to the FE estimates than the FD or RFD estimates, at least in the case of the IFE1 estimate. The IFE2 estimates often differ in magnitude from the IFE1 estimate.

Finally, a striking pattern emerges when examining the instrumental variable (IV) estimates in James (2015). Omitting the details, the author worries about the endogeneity of resource revenues even with the inclusion of state FEs and thus combines IV with FE (IV-FE) in the specifications omitting Alaska. Interestingly, in four of five cases the FD estimates are much closer to the IV-FE estimates than the FE estimates. This is consistent with FD removing more relevant unobserved heterogeneity than FE (but not accounting for as much unobserved heterogeneity as IV-FE). While certainly not a general result, in this case FD, twice-differencing, and the RFD estimators isolate much of the same (or similar) exogenous variation in resource-based government revenue as does the instrument. For example, for total expenditures (Panel

[17] See https://fred.stlouisfed.org/series/CAUR.

C), the point estimates change from 0.43 (FE) to 0.19 (FD) to -0.03 (IV-FE). For education expenditures (Panel D), the point estimates change from 0.15 (FE) to 0.02 (FD) to -0.03 (IV-FE). Finally, for public savings (Panel E), the point estimates change from 0.32 (FE) to 0.55 (FD) to 0.76 (IV-FE).

5 Conclusion

More data is clearly "better" when everything else is held constant. But rarely is everything else held constant. In this paper, we document the consequences for causal identification of increasing the number of time periods T when using the FE estimator in the presence of time-varying unit fixed effects. Specifically, we highlight the oft-overlooked fact that as $T \to \infty$, the FE estimator is of decreasing usefulness for causal inference because the amount of time-invariant heterogeneity is decreasing in T.

To remedy this, we have explored several solutions. The first is the well-known FD estimator, which is identical to the FE estimator when $T = 2$ but differs from it then $T > 2$. The second is the TFD estimator. Finally, we propose novel estimation algorithms—dubbed RFD and RTFD—based on the concept of rolling regressions. RFD conducts a series of two-period regressions and aggregates the results. RTFD aggregates the results from a series of three-period regressions.

We then illustrate the performance in practice of various estimators through simulations and by replicating the empirical findings in Rose (2004) and Tomz, Goldstein and Rivers (2007) and in James (2015). The FE estimator is at best marginally more efficient when the temporal heterogeneity in the unit fixed effects is very low or nonexistent. This is a well-known result (when the unit fixed effects are time-invariant) and explains the popularity of FE in empirical research, but the dramatic deterioration in the performance of FE as temporal heterogeneity in the unit fixed effects increases is perhaps surprising *a priori*. Moreover, FD, TFD, and our rolling estimators are generally very similar both in simulations and the replications. It is FE (and IFE1) that are notable outliers. Finally, when the temporal heterogeneity in the unit fixed effects is relatively high, TFD and RTFD perform even better than FD and RFD as even more unobserved heterogeneity is removed.

One the basis of our results, we make the following recommendations for applied researchers interested in causal inference with panel data where $i \in \{1, ..., N\}$ and $t \in \{1, ..., T\}$, with $N >> T$:

1. When $T > 2$, report FD, TFD, RFD, and RTFD estimates for sensitivity in addition to standard FE results. If the estimates are similar across all estimators, then the FE results are likely robust and more efficient. If they are not similar, investigate the source(s) of the discrepancy. Absent a compelling justification, the FD, TFD, RFD, and RTFD estimates ought to be given greater weight given their superior performance in terms of RMSE.

2. While the FD, TFD, RFD, and RTFD estimators are clearly superior to the FE estimator from a causal inference perspective, in practice FD and RFD and then TFD and RTFD results tend to be very similar, differing mainly in their standard errors. We thus advocate presenting all estimation results together.

3. When $T = 2$, there is no need to report FD and RFE estimates in addition to FE estimates, as all three estimators are identical. TFD and RTFD require $T > 2$.

4. At the data collection stage, seek to assemble higher-frequency data (i.e., more sub-periods within each period t) rather than (or in addition to) longer (i.e., larger T) panels. This is consistent with McKenzie (2012), who advocates that researchers conducting field experiments take multiple measurements of noisy outcomes characterized by a low degree of autocorrelation (e.g., agricultural yields, business profits, household expenditures).[18]

Our analysis brings to light a few questions for future research. First, as we discuss, a RFE estimator may be worth considering in certain situations. Second, the rolling estimators can be combined with existing approaches to allow for group heterogeneity in the slope coefficients (and/or time fixed effects) within each rolling regression. This not only allows for group heterogeneity at every time period, but it also allows the group structure to change over time. Third, our analysis may have implications for testing for structural breaks in panel data. In the usual setup where one wishes to perform a Chow test, it makes no difference whether one splits the sample at the suspected break date or includes a full set of interactions and estimates a pooled model. But with unit fixed effects in the model, splitting the sample allows the unit fixed effects to capture unobserved heterogeneity that is time-invariant before the break date as well as after the break date. Pooling the sample only removes unobserved heterogeneity that is time-invariant within units over the entire sample period. Fourth, as mentioned in Chan and Mátyás (2022), it may be advantageous to use machine learning techniques to choose elements of α_{it} to include in the model under a sparsity assumption. Finally, given the popularity of two-way fixed effects estimators (TWFE) for causal inference in difference-in-difference designs, as well as the now well-known issues created by staggered adoption, investigating the performance of rolling estimators in that context seems warranted (see, e.g., de Chaisemartin and d'Haultfoeuille, 2020, forthcoming; Wooldridge, 2021).

References

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2022. "When should you adjust standard errors for clustering?"

Ahn, Seung Chan, Young Hoon Lee, and Peter Schmidt. 2001. "GMM estimation of linear panel data models with time-varying individual effects." *Journal of Econometrics*, 101(2): 219–255.

Ahn, Seung Chan, Young Hoon Lee, and Peter Schmidt. 2013. "Panel data models with multiple time-varying individual effects." *Journal of Econometrics*, 174(1): 1–14.

Anderson, James E, and Eric van Wincoop. 2003. "Gravity with gravitas: a solution to the border puzzle." *American Economic Review*, 93(1): 170–192.

Ando, Tomohiro, and Jushan Bai. 2016. "Panel data models with grouped factor structure under unknown group membership." *Journal of Applied Econometrics*, 31(1): 163–191.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

[18]The difference is that McKenzie (2012) argues for higher-frequency data in an effort to increase statistical power, causal inference being all but guaranteed with a field experiment.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2010. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of economic perspectives*, 24(2): 3–30.

Aquaro, M, and P Čížek. 2013. "One-step robust estimation of fixed-effects panel data models." *Computational Statistics & Data Analysis*, 57(1): 536–548.

Autor, David H. 2003. "Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing." *Journal of labor economics*, 21(1): 1–42.

Backhouse, Roger E, and Béatrice Cherrier. 2017. "The age of the applied economist: the transformation of economics since the 1970s." *History of Political Economy*, 49(Supplement): 1–33.

Bai, Jushan. 2009. "Panel Data Models With Interactive Fixed Effects." *Econometrica*, 77(4): 1229–1279.

Balestra, Pietro, and Marc Nerlove. 1966. "Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas." *Econometrica*, 34(3): 585–612.

Baltagi, Badi H., Qu Feng, and Chihwa Kao. 2016. "Estimation of heterogeneous panels with structural breaks." *Journal of Econometrics*, 191(1): 176–195.

Becker, Gary S., and Casey B. Mulligan. 1997. "The endogenous determination of time preference." *Quarterly Journal of Economics*, 112(3): 729–758.

Boldea, Otilia, Bettina Drepper, and Zhuojiong Gan. 2020. "Change point estimation in panel data with time-varying individual effects." *Journal of Applied Econometrics*, 35(6): 712–727.

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped patterns of heterogeneity in panel data." *Econometrica*, 83(3): 1147–1184.

Bowles, Samuel. 1998. "Endogenous preferences: the cultural consequences of markets and other economic institutions." *Journal of Economic Literature*, 36(1): 75–111.

Bramati, Maria Caterina, and Christophe Croux. 2007. "Robust estimators for the fixed effects panel data model." *The Econometrics Journal*, 10(3): 521–540.

Cai, Zongwu, and Ted Juhl. forthcoming. "The distribution of rolling regression estimators." *Journal of Econometrics*.

Caughey, Devin, and Christopher Warshaw. 2016. "The Dynamics of State Policy Liberalism, 1936–2014." *American Journal of Political Science*, 60(4): 899–913.

Chan, Felix, and László Mátyás. 2022. "Linear econometric models with machine learning." In *Econometrics with Machine Learning*. , ed. Felix Chan and László Mátyás, Chapter 1, 1–37. Springer Nature Switzerland AG.

Chang, Pao-Li, and Myoung-Jae Lee. 2011. "The WTO trade effect." *Journal of International Economics*, 85(1): 53–71.

de Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–2996.

de Chaisemartin, Clément, and Xavier d'Haultfoeuille. forthcoming. "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey."

Ding, Weili, and Steven F Lehrer. 2014. "Understanding the role of time-varying unobserved ability heterogeneity in education production." *Economics of Education Review*, 40: 55–75.

Eicher, Theo S, and Christian Henn. 2011. "In search of WTO trade effects: Preferential trade agreements promote trade strongly, but unevenly." *Journal of International Economics*, 83(2): 137–153.

Fershtman, Chaim, and Uzi Segal. 2018. "Preferences and social influence." *American Economic Journal: Microeconomics*, 10(3): 124–142.

Frank, Mark W. 2009. "Inequality and growth in the United States: evidence from a new state-level panel of income inequality measures." *Economic Inquiry*, 47(1): 55–68.

Gibbons, Charles E, Juan Carlos Suárez Serrato, and Michael B Urbancic. 2019. "Broken or fixed effects?" *Journal of Econometric Methods*, 8(1). Article 20170002.

Han, Chirok, Luis Orea, and Peter Schmidt. 2005. "Estimation of a panel data model with parametric temporal variation in individual effects." *Journal of Econometrics*, 126(2): 241–267.

Hausman, Jerry, and Daniel McFadden. 1984. "Specification tests for the multinomial logit model." *Econometrica*, 52(5): 1219–1240.

Hill, Terrence D, Andrew P Davis, J Micah Roos, and Michael T French. 2020. "Limitations of fixed-effects models for panel data." *Sociological Perspectives*, 63(3): 357–369.

Hsiao, Cheng. 2007. "Panel data analysis—advantages and challenges." *Test*, 16: 1–22.

Imai, Kosuke, and In Song Kim. 2019. "When should we use unit fixed effects regression models for causal inference with longitudinal data?" *American Journal of Political Science*, 63(2): 467–490.

Imai, Kosuke, and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis*, 29(3): 405–415.

James, Alexander. 2015. "US State Fiscal Policy and Natural Resources." *American Economic Journal: Economic Policy*, 7(3): 238–257.

Jiang, Peiyun, and Eiji Kurozumi. forthcoming. "A new test for common breaks in heterogeneous panel data models." *Econometrics and Statistics*.

Kaddoura, Yousef, and Joakim Westerlund. forthcoming. "Estimation of panel data models with random interactive effects and multiple structural breaks when T is fixed." *Journal of Business & Economic Statistics*.

Keane, Michael, and Timothy Neal. 2020. "Climate change and U.S. agriculture: Accounting for multidimensional slope heterogeneity in panel data." *Quantitative Economics*, 11(4): 1391–1429.

Laporte, Audrey, and Frank Windmeijer. 2005. "Estimation of panel data models with binary indicators when treatment effects are not constant over time." *Economics Letters*, 88(3): 389–396.

Lewis, Daniel, Davide Melcangi, Laura Pilossoph, and Aidan Toner-Rodgers. 2022. "Approximating grouped fixed effects estimation via fuzzy clustering regression."

Liebenehm, Sabine, Nele Degener, and Eric Strobl. forthcoming. "Rainfall shocks and risk aversion: Evidence from Southeast Asia." *American Journal of Agricultural Economics*.

Liu, Ruiqi, Zuofeng Shang, Yonghui Zhang, and Qiankun Zhou. 2020. "Identification and estimation in panel models with overspecified number of groups." *Journal of Econometrics*, 215(2): 574–590.

Lumsdaine, Robin L, Ryo Okui, and Wendun Wang. forthcoming. "Estimation of panel group structure models with structural breaks in group memberships and coefficients." *Journal of Econometrics*.

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics*, 99(2): 210–221.

Mehrabani, Ali. forthcoming. "Estimation and identification of latent group structures in panel data." *Journal of Econometrics*.

Mullahy, John. 2016. "Estimation of multivariate probit models via bivariate probit." *The Stata Journal*, 16(1): 37–51.

Mundlak, Yair. 1961. "Empirical production function free of management bias." *Journal of Farm Economics*, 43(1): 44–56.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica*, 46(1): 69–85.

Nickell, Stephen. 1981. "Biases in dynamic models with fixed effects." *Econometrica*, 49(6): 1417–1426.

Papke, Leslie E, and Jeffrey M Wooldridge. forthcoming. "A simple, robust test for choosing the level of fixed effects in linear panel data models." *Empirical Economics*.

Pesaran, M Hashem, and Qiankun Zhou. 2018. "To pool or not to pool: revisited." *Oxford Bulletin of Economics and Statistics*, 80(2): 185–217.

Rose, Andrew K. 2004. "Do we really know that the WTO increases trade?" *American Economic Review*, 94(1): 98–114.

Rousseeuw, Peter, and Victor Yohai. 1984. "Robust regression by means of S-estimators." In *Robust and Nonlinear Time Series Analysis*. , ed. Jürgen Franke, Wolfgang Härdle and Douglas Martin, 256–272. New York, NY:Springer US.

Sarafidis, Vasilis, and Neville Weber. 2015. "A partially heterogeneous framework for analyzing panel data." *Oxford Bulletin of Economics and Statistics*, 77(2): 274–296.

Spierdijk, Laura. 2023. "Assessing the consistency of the fixed-effects estimator: a regression-based Wald test." *Empirical Economics*, 64(4): 1599–1630.

Su, Liangjun, and Qihui Chen. 2013. "Testing homogeneity in panel data models with interactive fixed effects." *Econometric Theory*, 29(6): 1079–1135.

Su, Liangjun, Zhentao Shi, and Peter C B Phillips. 2016. "Identifying latent structures in panel data." *Econometrica*, 84(6): 2215–2264.

Sun, Liyang, and Jesse M Shapiro. 2022. "A linear panel model with heterogeneous coefficients and variation in exposure." *Journal of Economic Perspectives*, 36(4): 193–204.

Thombs, Ryan P. 2022. "A guide to analyzing large N, large T panel data." *Socius*, 8: 1–15.

Tomz, Michael, Judith L Goldstein, and Douglas Rivers. 2007. "Do We Really Know That the WTO Increases Trade? Comment." *American Economic Review*, 97(5): 2005–2018.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* . 2 ed., MIT Press.

Wooldridge, Jeffrey M. 2021. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." *Available at SSRN 3906345*.

Yohai, Victor J. 1987. "High breakdown-point and high efficiency robust estimates for regression." *The Annals of statistics*, 642–656.

TABLE I: Replication: Imai and Kim (2019)

	FE	FD	Twice FD	RFD (cons)	RFD (no cons)	Twice RFD (cons)	Twice RFD (no cons)	IFE1	IFE2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Formal GATT Membership — No Year FEs									
GATT (Formal)	-0.048**	0.051**	0.063*	0.006	0.021	0.007	0.008		
	(0.024)	(0.024)	(0.033)	(0.017)	(0.020)	(0.022)	(0.022)		
LW Test		p=0.000							
Panel B. Formal GATT Membership — With Year FEs									
GATT (Formal)	0.036	0.037	0.061*	0.006	0.006	0.007	0.007	0.056**	0.023
	(0.024)	(0.024)	(0.033)	(0.017)	(0.017)	(0.022)	(0.022)	(0.024)	(0.021)
LW Test		p=0.000							
Panel C. Informal GATT Membership — No Year FEs									
GATT (Participate)	0.147***	0.066***	0.060*	0.041***	0.064***	0.021	0.025		
	(0.030)	(0.025)	(0.036)	(0.014)	(0.014)	(0.021)	(0.021)		
LWTest		p=0.000							
Panel D. Informal GATT Membership — With Year FEs									
GATT (Participate)	0.227***	0.044*	0.054	0.041***	0.041***	0.021	0.021	0.262***	0.014
	(0.030)	(0.026)	(0.036)	(0.014)	(0.014)	(0.021)	(0.021)	(0.029)	(0.027)
LW Test		p=0.000							

Formal membership includes only formal GATT members as in Rose (2004); informal includes nonmember participants as in Tomz, Goldstein and Rivers (2007). Other controls include include Generalized System of Preferences, log product real GDP, log product real GDP per capita, regional free trade agreement, currency union, and currently colonized. LW = Laporte and Windmeijer (2005) test of equality of FE and FD. FE = fixed effects. FD = first-differences. RFD = rolling first differences. IFE1 = interactive fixed effects (1 factor). IFE2 = interactive fixed effects (2 factors). cons/no cons refers to the inclusion of a constant in the first-differenced or twice-differenced specifications. IFE1 and IFE2 are omitted in Panels A and C since IFE always includes time-varying factor(s). * p <.10, ** p< .05, *** p<.01.

TABLE II: Replication: James (2015) Full Sample

	FE	FD	Twice FD	RFD (cons)	RFD (no cons)	Twice RFD (cons)	Twice RFD (no cons)	IFE1	IFE2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Nonresource Revenue									
Resource Revenue	0.006	-0.020***	-0.014**	-0.027***	-0.025***	-0.019**	-0.025***	-0.028	-0.018
	(0.020)	(0.005)	(0.005)	(0.009)	(0.003)	(0.009)	(0.002)	(0.020)	(0.018)
LW Test		p= 0.106							
Panel B. Income Tax Revenue									
Resource Revenue	0.018*	0.022***	0.013***	0.014*	0.001	0.004	-0.001	0.022***	0.072***
	(0.010)	(0.001)	(0.001)	(0.007)	(0.001)	(0.007)	(0.001)	(0.008)	(0.019)
LW Test		p=0.671							
Panel C. Total Expenditures									
Resource Revenue	0.397***	-0.003	-0.020***	-0.008	-0.056***	-0.013*	-0.033***	0.328***	0.109**
	(0.006)	(0.003)	(0.002)	(0.012)	(0.010)	(0.008)	(0.001)	(0.021)	(0.048)
LW Test		p= 0.000							
Panel D. Education Expenditures									
Resource Revenue	0.063***	-0.010***	-0.019***	-0.003	-0.034***	-0.015***	-0.018***	0.061***	0.075***
	(0.007)	(0.001)	(0.001)	(0.005)	(0.005)	(0.003)	(0.002)	(0.008)	(0.007)
LW Test		p=0.000							
Panel E. Public Savings									
Resource Revenue	0.609***	0.983***	1.006***	1.001***	1.043***	0.996***	1.006***	0.669***	0.224***
	(0.021)	(0.007)	(0.005)	(0.013)	(0.006)	(0.009)	(0.001)	(0.038)	(0.068)
LW Test		p=0.000							

Notes: LW = Laporte and Windmeijer (2005) test of equality of FE and FD. FE = fixed effects. FD = first-differences. RFD = rolling first differences. IFE1 = interactive fixed effects (1 factor). IFE2 = interactive fixed effects (2 factors). cons/no cons refers to the inclusion of a constant in the first-differenced or twice-differenced specifications. Time fixed effects included in all models. * p <0.10, ** p<0.05, *** p<0.01.

TABLE III: Replication: James (2015) Omit Alaska

	FE	FD	Twice FD	RFD (cons)	RFD (no cons)	Twice RFD (cons)	Twice RFD (no cons)	IFE1	IFE2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Nonresource Revenue									
Resource Revenue	-0.249***	-0.255	-0.340	-0.124	-0.137	-0.212*	-0.229*	-0.316***	-0.224
	(0.061)	(0.169)	(0.327)	(0.094)	(0.105)	(0.114)	(0.114)	(0.061)	(0.219)
LW Test		p=0.969							
Panel B. Income Tax Revenue									
Resource Revenue	-0.104**	-0.010	0.016	0.009	0.001	0.019	0.012	-0.067*	-0.076
	(0.039)	(0.030)	(0.040)	(0.016)	(0.014)	(0.015)	(0.011)	(0.036)	(0.056)
LW Test		p= 0.039							
Panel C. Total Expenditures									
Resource Revenue	0.429***	0.191**	0.000	0.101	0.198**	-0.118*	-0.121	0.429***	0.563**
	(0.063)	(0.085)	(0.172)	(0.071)	(0.088)	(0.060)	(0.088)	(0.090)	(0.232)
LW Test		p= 0.004							
Panel D. Education Expenditures									
Resource Revenue	0.147***	0.017	-0.042	-0.016	-0.023	-0.056	-0.078*	0.156***	0.136**
	(0.036)	(0.039)	(0.077)	(0.034)	(0.035)	(0.039)	(0.040)	(0.037)	(0.063)
LW Test		p=0.012							
Panel E. Public Savings									
Resource Revenue	0.322***	0.553***	0.661***	0.761***	0.544***	0.931***	0.755***	0.214***	0.230***
	(0.038)	(0.139)	(0.197)	(0.112)	(0.115)	(0.112)	(0.132)	(0.060)	(0.076)
LW Test		p= 0.133							

Notes: LW = Laporte and Windmeijer (2005) test of equality of FE and FD. FE = fixed effects. FD = first-differences. RFD = rolling first differences. IFE1 = interactive fixed effects (1 factor). IFE2 = interactive fixed effects (2 factors). cons/no cons refers to the inclusion of a constant in the first-differenced or twice-differenced specifications. Time fixed effects included in all models. * p <0.10, ** p<0.05, *** p<0.01.

Fixed Effects and Causal Inference

Supplemental Appendix

Daniel L. Millimet & Marc F. Bellemare

June 1, 2023

A Monte Carlo Study

We simulate data from several experimental designs. The first data-generating process (DGP1) allows for three regimes and imposes common break dates:

$$\begin{aligned}
y_{it} &= \beta x_{it} + \mathtt{I}(t \leq T_1)\alpha_i^{1,T_1} + \mathtt{I}(T_1 < t \leq T_2)\alpha_i^{T_1+1,T_2} + \mathtt{I}(t > T_2)\alpha_i^{T_2+1,T} \\
&\quad + \lambda_t + \varepsilon_{it}, \qquad i = 1, ..., N; \quad t = 1, ..., T \\
x_{it} &\sim \mathtt{N}(\mu_{it}, 1) \\
\mu_{it} &= \mathtt{I}(t \leq T_1)\alpha_i^{1,T_1} + \mathtt{I}(T_1 < t \leq T_2)\alpha_i^{T_1+1,T_2} + \mathtt{I}(t > T_2)\alpha_i^{T_2+1,T} \\
\alpha_i^{1,T_1} &\sim \mathtt{N}(\alpha_i, \sigma) \\
\alpha_i^{T_1+1,T_1} &\sim \mathtt{N}(\alpha_i, \sigma) \\
\alpha_i^{T_2+1,T} &\sim \mathtt{N}(\alpha_i, \sigma) \\
\alpha_i &\sim \mathtt{N}(0, 1) \\
\lambda_t &\sim \mathtt{N}(t, 1) \\
\varepsilon_{it} &\sim \mathtt{N}(0, 1)
\end{aligned} \tag{A.1}$$

where $T_1 = \mathtt{int}(T/3)$ and $T_2 = \mathtt{int}(2T/3)$. This is analogous to Kaddoura and Westerlund (forthcoming) except it is the unit fixed effects that vary across regimes rather than the slopes. We set $N = 1000$ and $\beta = 1$ in all experiments. We vary T and σ. Specifically, we consider $T \in \{5, 10, 15\}$ and $\sigma \in \{0, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50\}$. Notice when $\sigma = 0$, the unit fixed effect is again α_i for all t.

DGP2 allows for three regimes with unit-specific break dates:

$$\begin{aligned}
y_{it} &= \beta x_{it} + \mathtt{I}(t \leq T_{1i})\alpha_i^{1,T_{1i}} + \mathtt{I}(T_{1i} < t \leq T_{2i})\alpha_i^{T_{1i}+1,T_{2i}} + \mathtt{I}(t > T_{2i})\alpha_i^{T_{2i}+1,T} \\
&\quad + \lambda_t + \varepsilon_{it}, \qquad i = 1, ..., N; \quad t = 1, ..., T \\
x_{it} &\sim \mathtt{N}(\mu_{it}, 1) \\
\mu_{it} &= \mathtt{I}(t \leq T_{1i})\alpha_i^{1,T_{1i}} + \mathtt{I}(T_{1i} < t \leq T_{2i})\alpha_i^{T_{1i}+1,T_{2i}} + \mathtt{I}(t > T_{2i})\alpha_i^{T_{2i}+1,T} \\
\alpha_i^{1,T_1} &\sim \mathtt{N}(\alpha_i, \sigma) \\
\alpha_i^{T_1+1,T_1} &\sim \mathtt{N}(\alpha_i, \sigma) \\
\alpha_i^{T_2+1,T} &\sim \mathtt{N}(\alpha_i, \sigma) \\
\alpha_i &\sim \mathtt{N}(0, 1) \\
\lambda_t &\sim \mathtt{N}(t, 1) \\
\varepsilon_{it} &\sim \mathtt{N}(0, 1)
\end{aligned} \tag{A.2}$$

where $T_{1i} = \mathtt{int}(B_{1i})$ and $T_{2i} = \mathtt{int}(B_{2i})$ and

$$\begin{aligned}
B_{1i} &= \mathtt{U}(0.33 \cdot T - 1, 0.33 \cdot T + 2) \\
B_{2i} &= \mathtt{U}(B_{1i} + 1, T)
\end{aligned}$$

This is identical to DGP1 except now the regimes over which the unit fixed effects are constant are unit-

specific. Values of β, σ, and T are unchanged from DGP1.

DGP3 is from Mundlak (1978) and includes a unit-specific time trend:

$$\begin{aligned} y_{it} &= \beta x_{it} + \alpha_{it} + \lambda_t + \varepsilon_{it}, \qquad i = 1, ..., N;\ t = 1, ..., T \\ x_{it} &\sim \mathtt{N}(\alpha_{it}, 1) \\ \alpha_{0i} &\sim \mathtt{N}(0, 3) \\ \alpha_{1i} &\sim \mathtt{N}(0, \sigma) \\ \alpha_{it} &= \alpha_{0i} + \alpha_{1i} t \\ \lambda_t &\sim \mathtt{N}(t, 1) \\ \varepsilon_{it} &\sim \mathtt{N}(0, 1) \end{aligned} \tag{A.3}$$

Values of β, σ, and T are unchanged from DGP1.

DGP4 allows for the unit FE to follow a random walk:

$$\begin{aligned} y_{it} &= \beta x_{it} + \alpha_{it} + \lambda_t + \varepsilon_{it}, \qquad i = 1, ..., N;\ t = 1, ..., T \\ x_{it} &\sim \mathtt{N}(\alpha_{it}, 1) \\ \alpha_{i1} &\sim \mathtt{N}(0, 1) \\ \alpha_{it} &\sim \mathtt{N}(\alpha_{it-1}, \sigma), \qquad t > 1 \\ \lambda_t &\sim \mathtt{N}(t, 1) \\ \varepsilon_{it} &\sim \mathtt{N}(0, 1) \end{aligned} \tag{A.4}$$

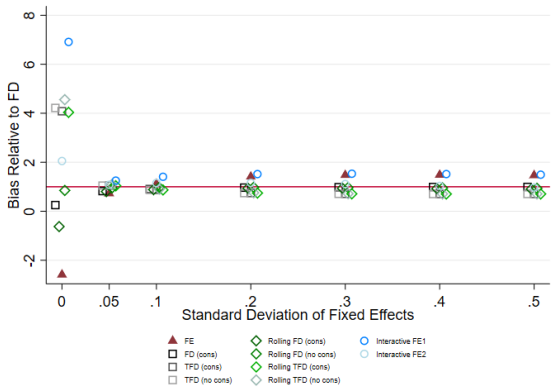
Values of β, σ, and T are unchanged from DGP1.

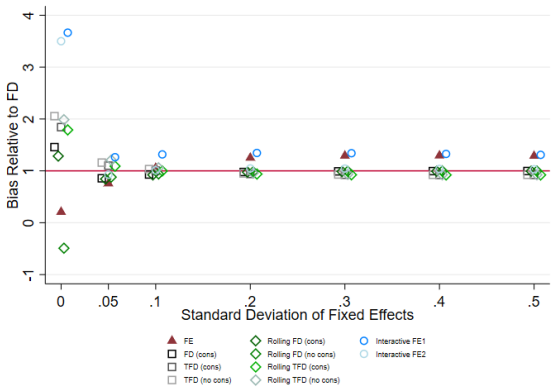
DGP5, DGP6, DGP7, and DGP8 are identical to DGP1-4 except follow Pesaran and Zhou (2018), allowing for time-varying FE for only some units. Specifically, the time-varying FE in each design is

$$\widetilde{\alpha}_{it} = \begin{cases} \alpha_i \sim \mathtt{N}(0, \tau) & i = 1, ..., \left[N^{\delta}\right] \\ \alpha_{it} & i = \left[N^{\delta}\right] + 1, ..., N \end{cases} \tag{A.5}$$

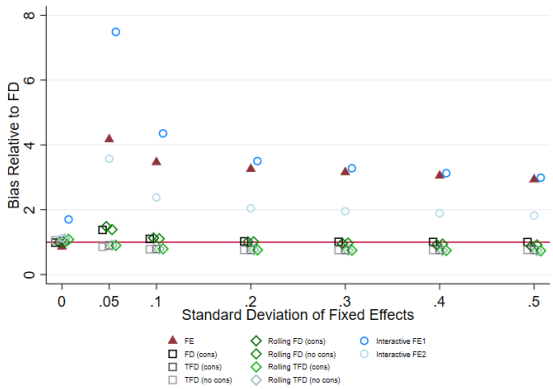
where τ equals one except in DGP7 where τ equals three. We set $T = 10$, $\delta = \{0.45, 0.475, 0.50, 0.525, 0.55, 0.60\}$, and $N = \{100, 500, 5000\}$. This implies that between 7 and 15 units have time-varying fixed effects when $N = 100$, 16 and 41 units when $N = 500$, and 46 and 165 when $N = 5000$.

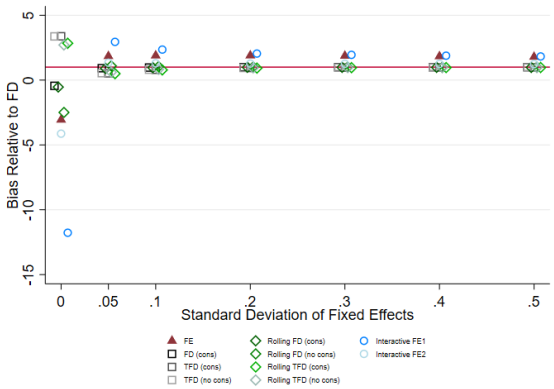
For each experiment, we conduct 200 simulations and report the mean bias and the root mean squared error (RMSE). We display the results graphically, both relative to the FD estimator (except for DGP3 which is relative to twice FD) and in absolute terms.



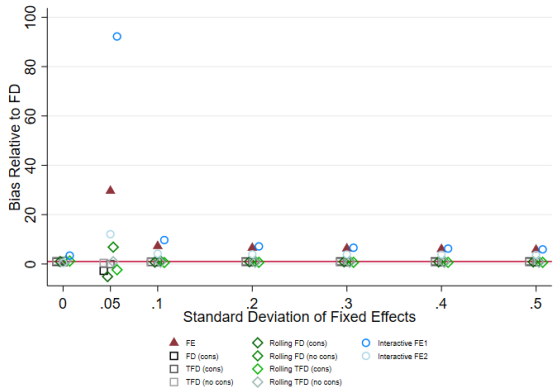


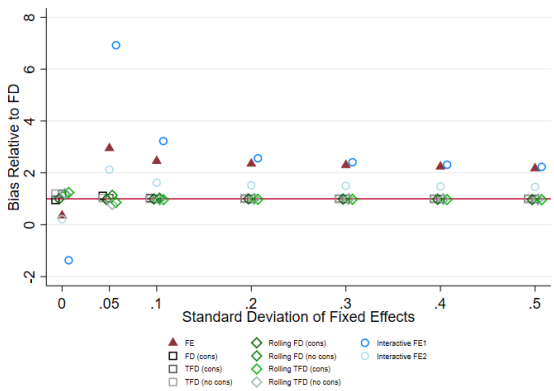
(A) $T = 5$





(B) $T = 10$

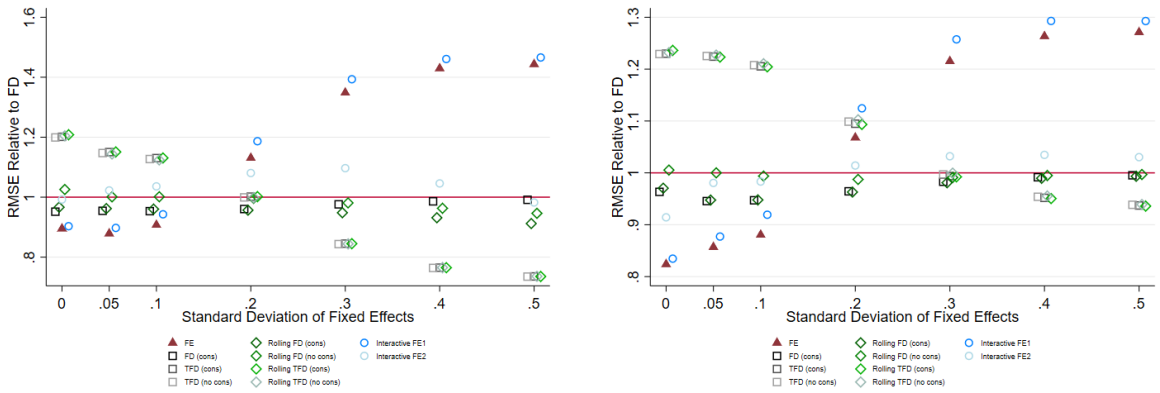




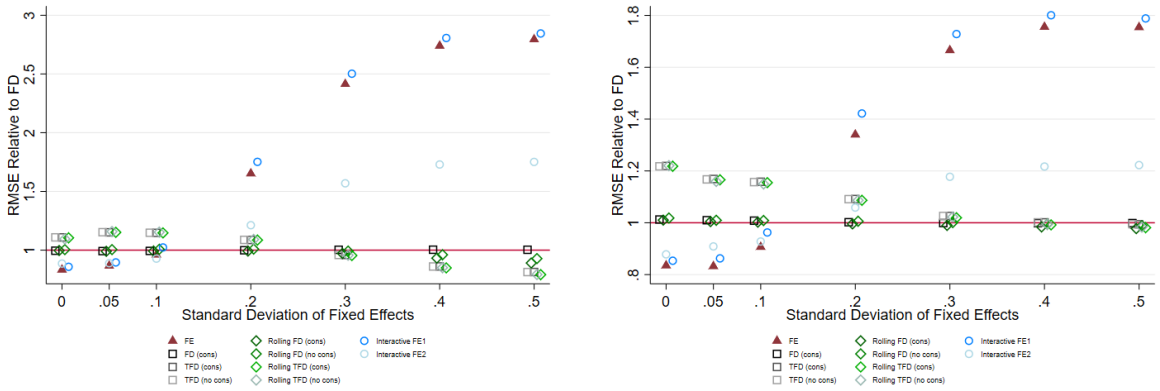
(C) $T = 20$

FIGURE A.1: Simulation Results: Bias

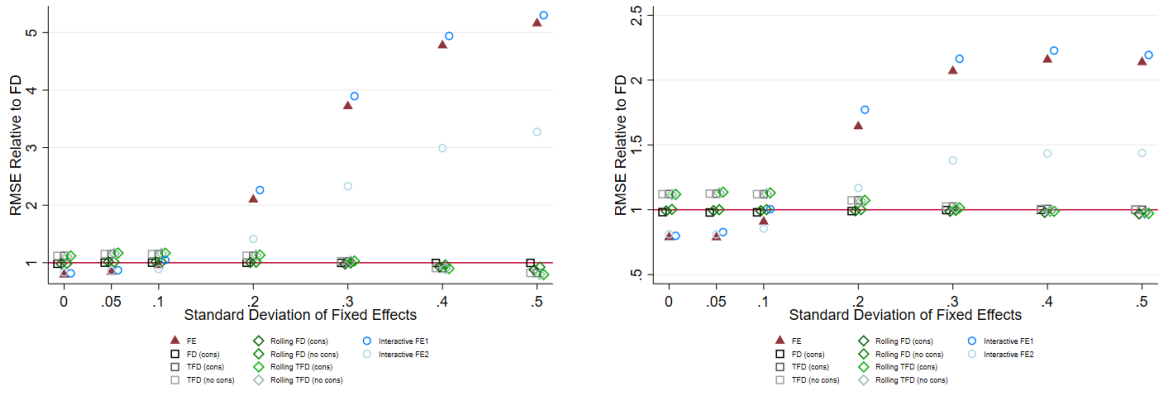
Notes: Performance is relative to the first-difference estimator. DGP1 in left column. DGP2 in right column.



(A) $T = 5$



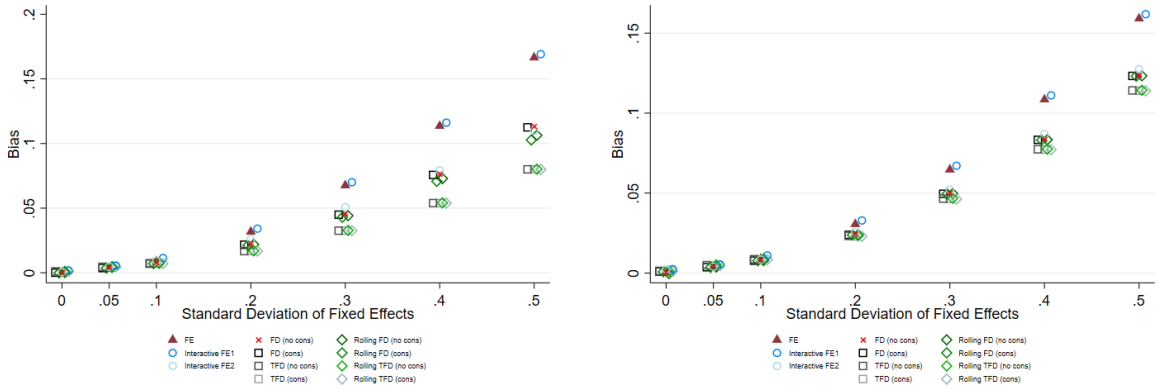
(B) $T = 10$



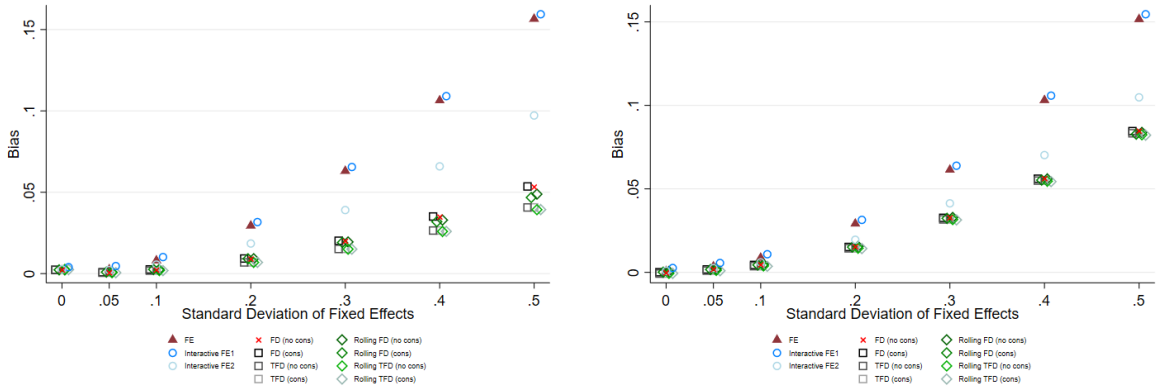
(C) $T = 20$

FIGURE A.2: Simulation Results: Root Mean Squared Error

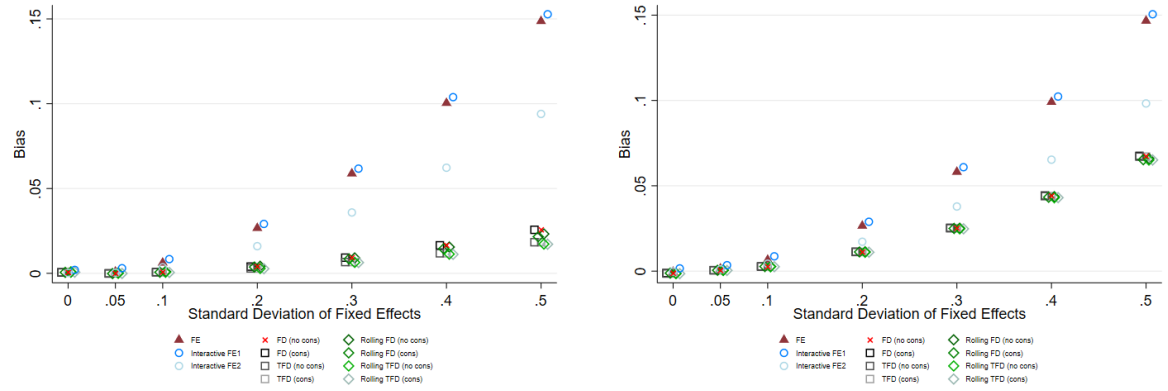
Notes: Performance is relative to the first-difference estimator. DGP1 in left column. DGP2 in right column.



(A) $T = 5$



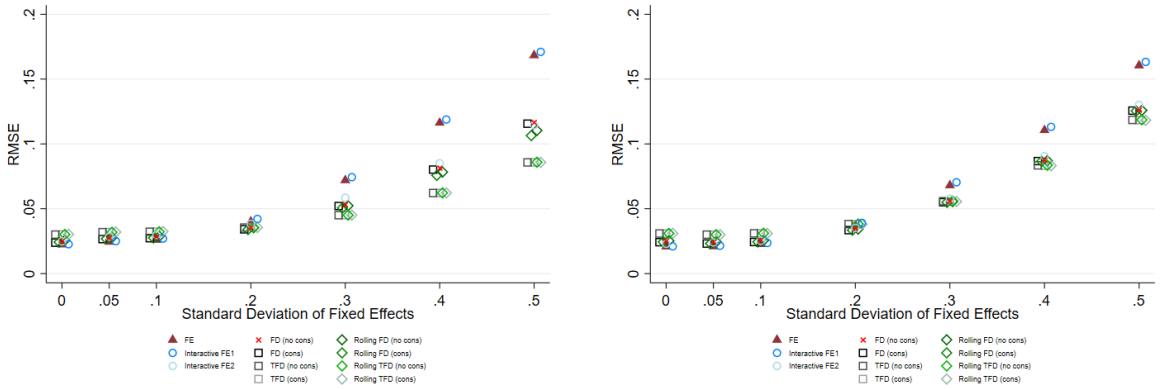
(B) $T = 10$



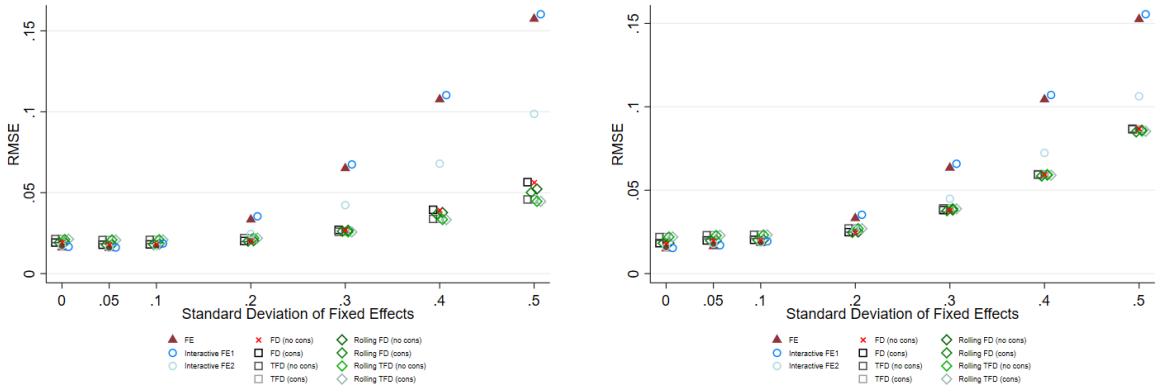
(C) $T = 20$

FIGURE A.3: Simulation Results: Bias

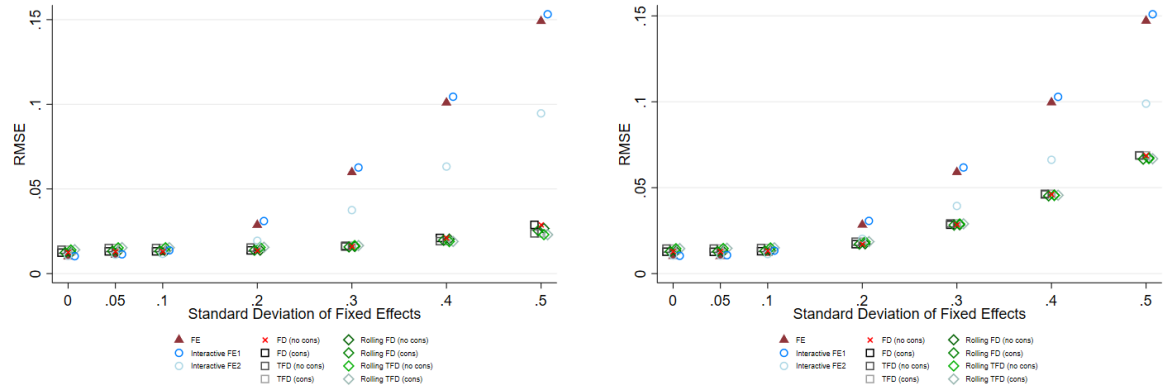
Notes: DGP1 in left column. DGP2 in right column.



(A) $T = 5$



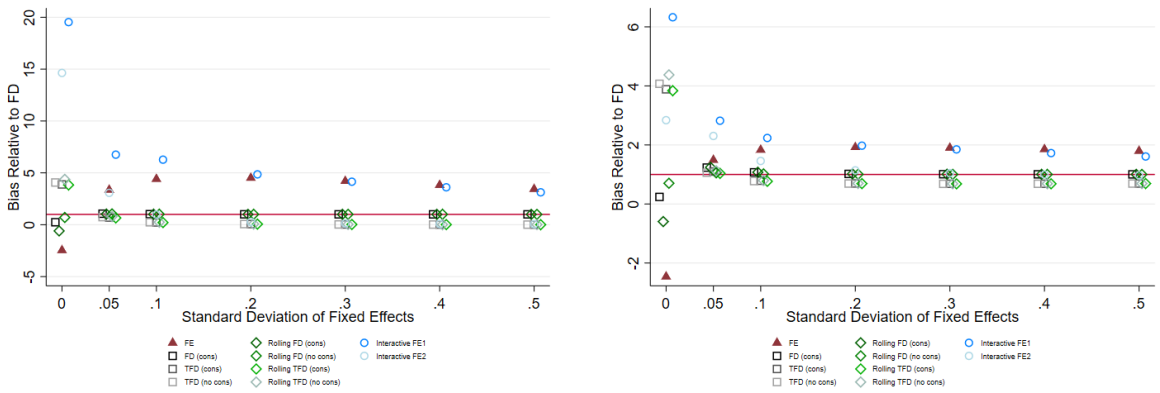
(B) $T = 10$



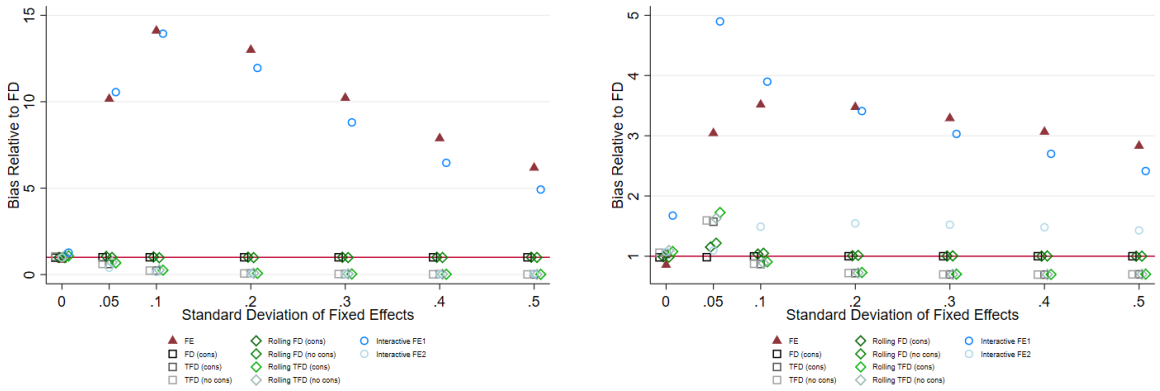
(C) $T = 20$

FIGURE A.4: Simulation Results: Root Mean Squared Error

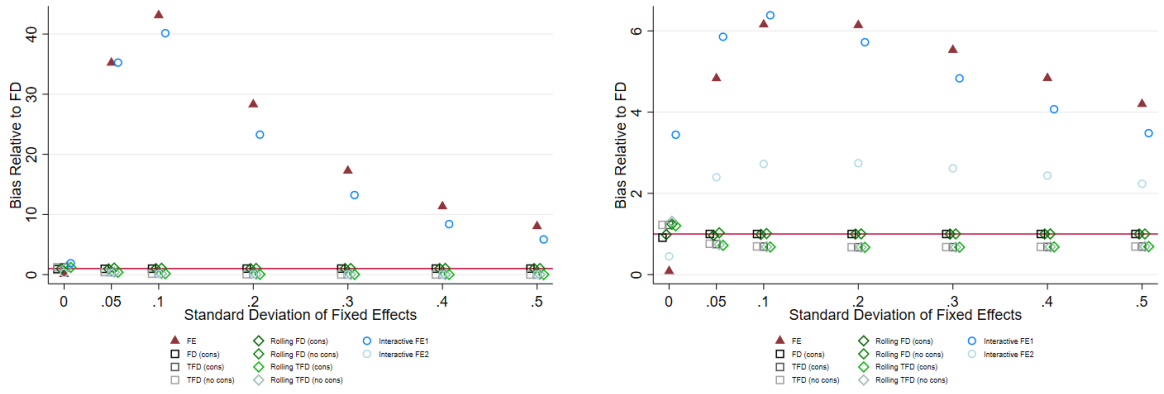
Notes: DGP1 in left column. DGP2 in right column.



(A) $T = 5$



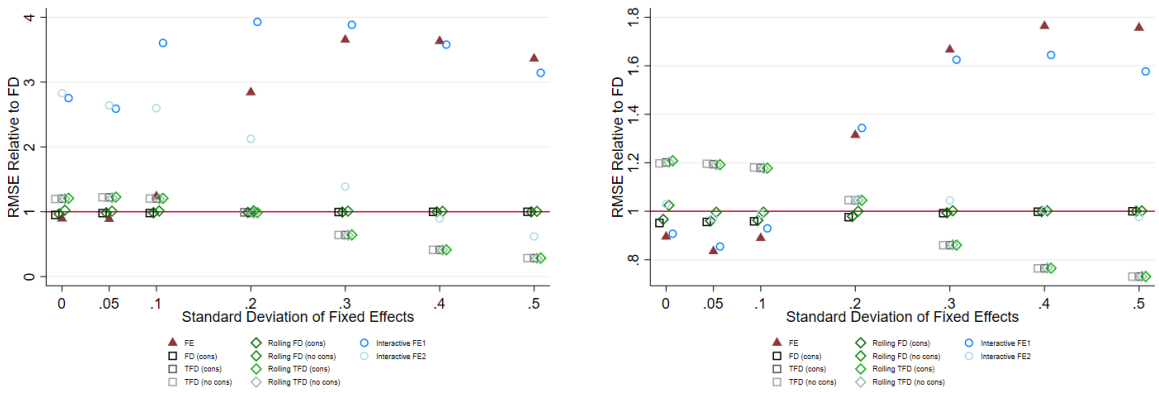
(B) $T = 10$



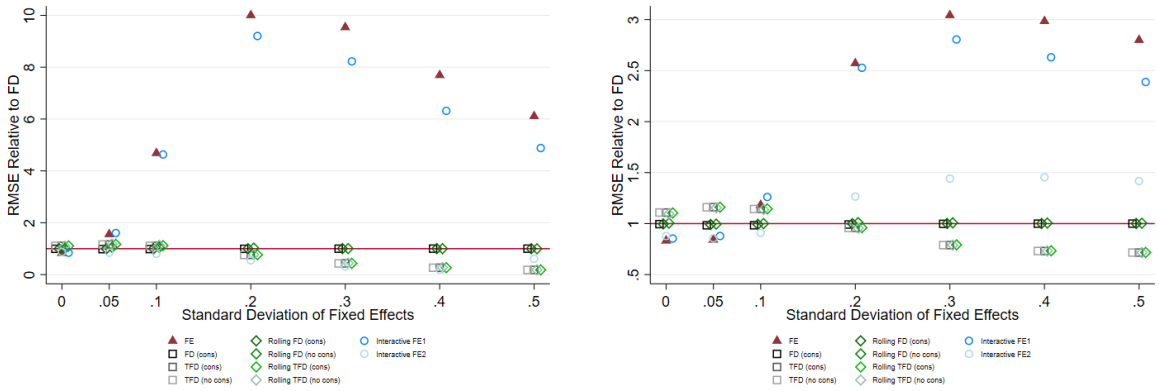
(C) $T = 20$

FIGURE A.5: Simulation Results: Bias

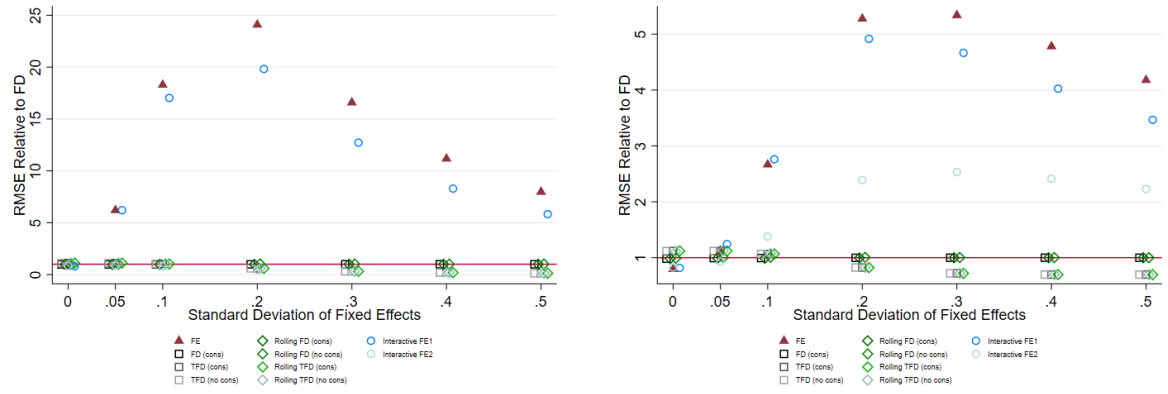
Notes: Performance is relative to the first-difference estimator. DGP3 in left column. DGP4 in right column.



(A) $T = 5$



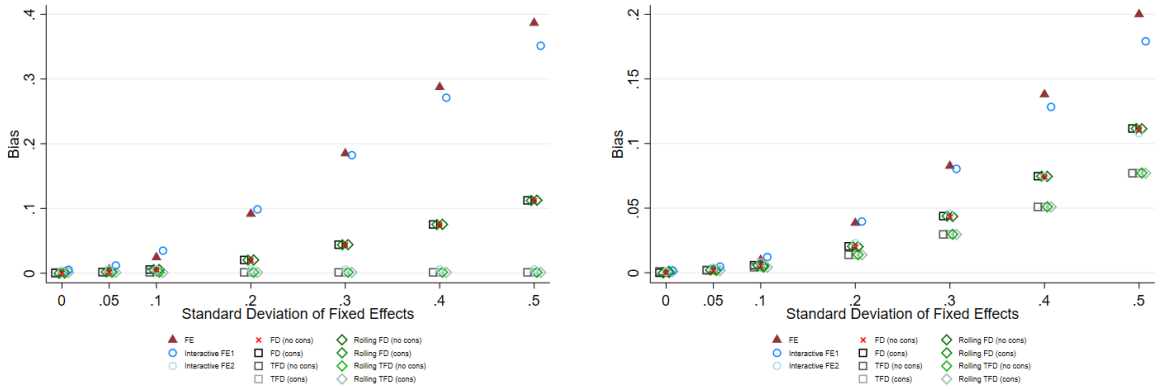
(B) $T = 10$



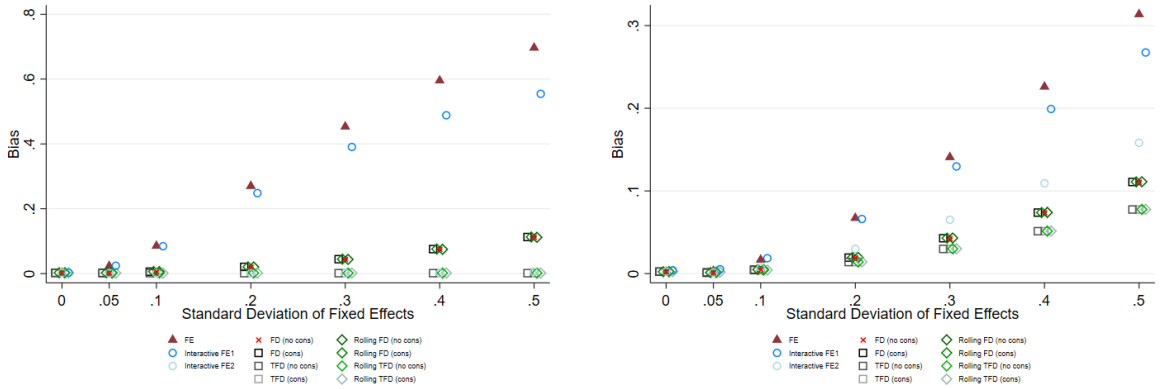
(C) $T = 20$

FIGURE A.6: Simulation Results: Root Mean Squared Error

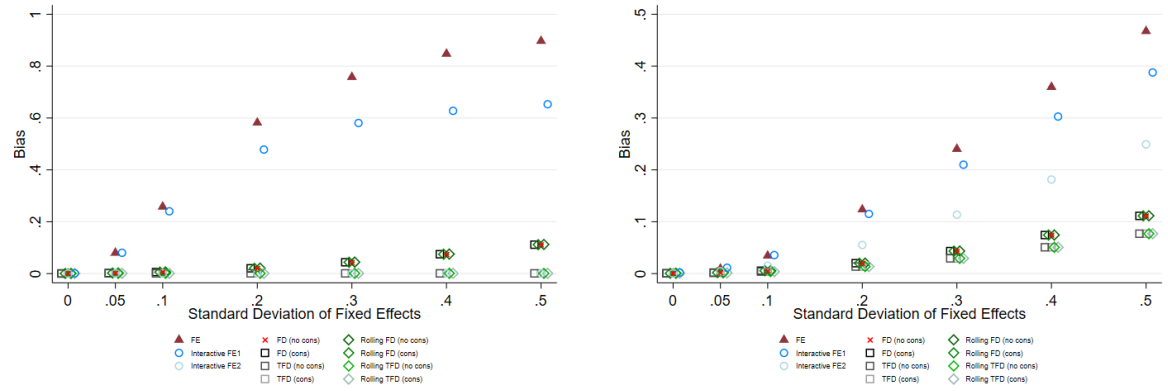
Notes: Performance is relative to the first-difference estimator. DGP3 in left column. DGP4 in right column.



(A) $T = 5$



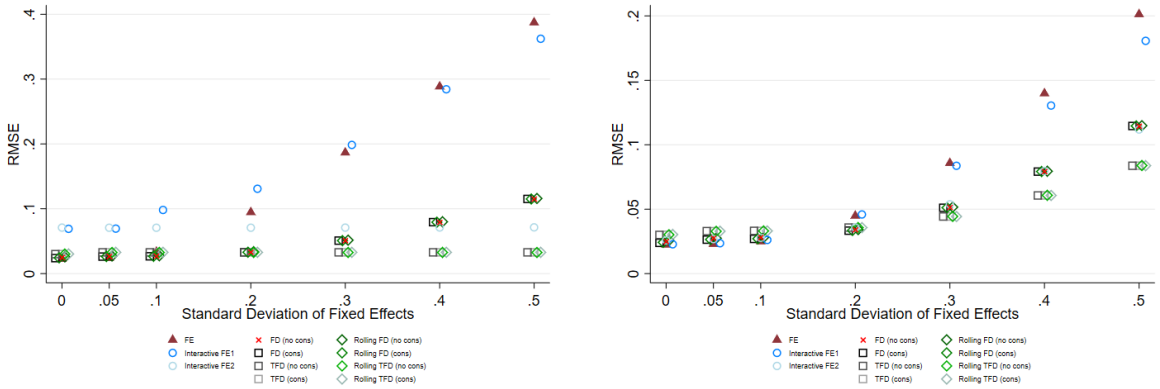
(B) $T = 10$



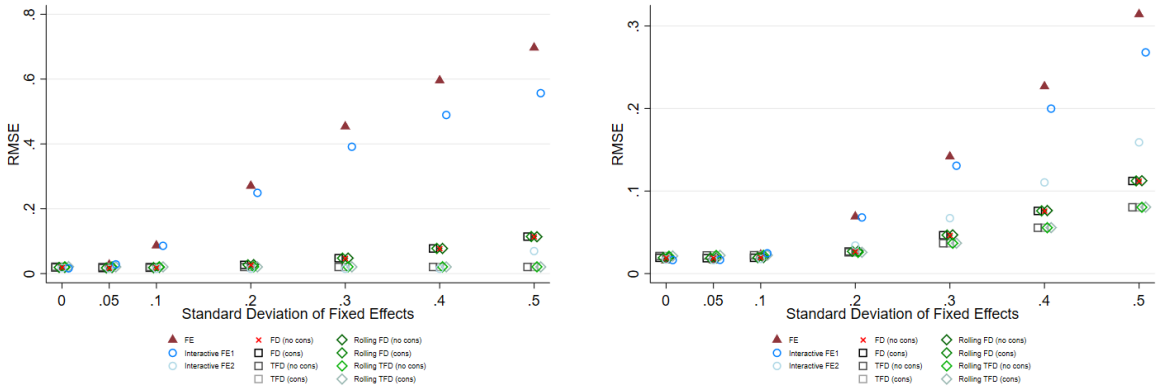
(C) $T = 20$

FIGURE A.7: Simulation Results: Bias

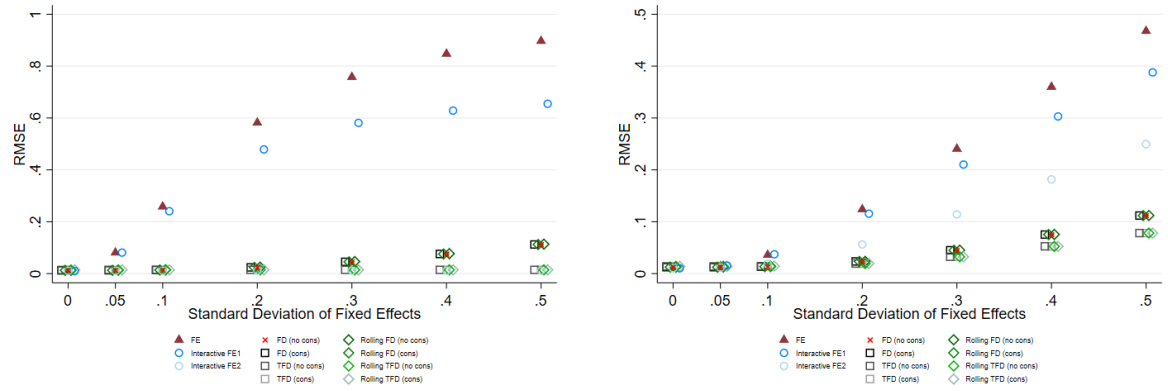
Notes: DGP3 in left column. DGP4 in right column.



(A) $T = 5$



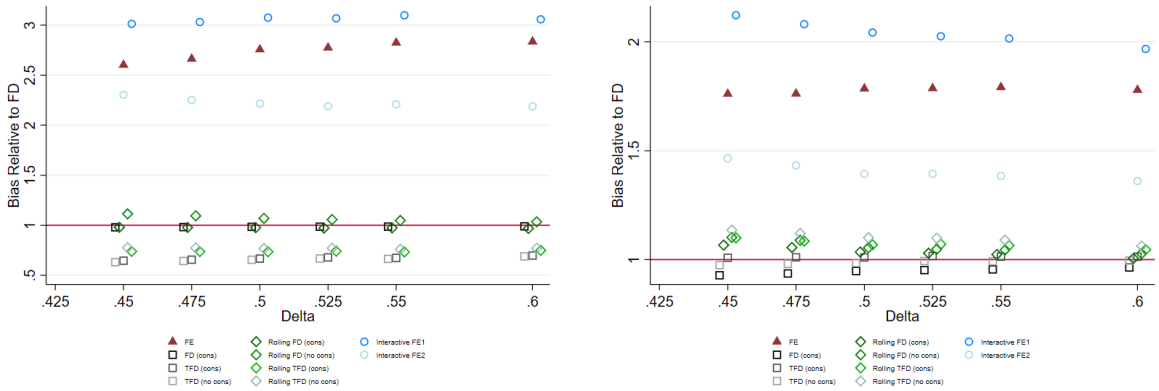
(B) $T = 10$



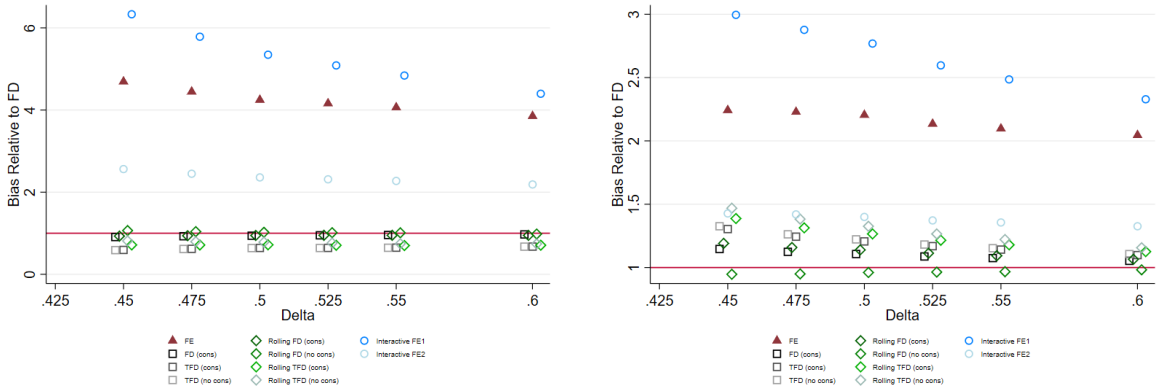
(C) $T = 20$

FIGURE A.8: Simulation Results: Root Mean Squared Error

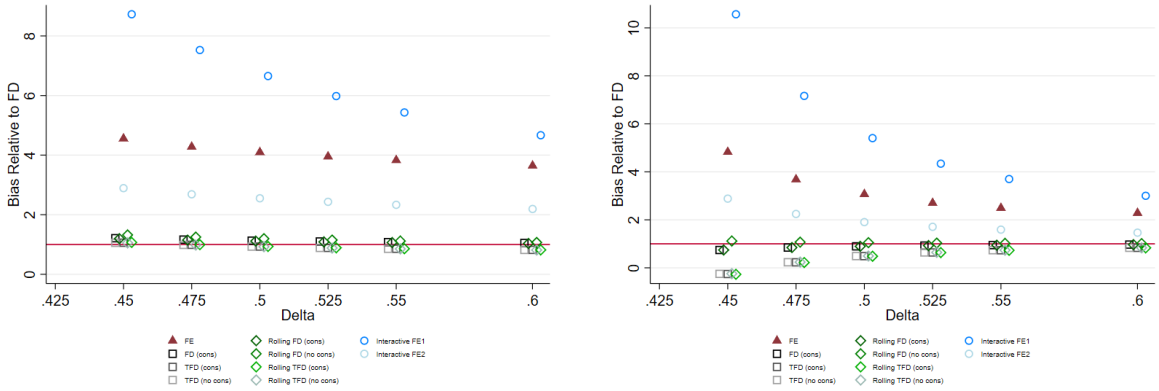
Notes: DGP3 in left column. DGP4 in right column.



(A) $N = 100$



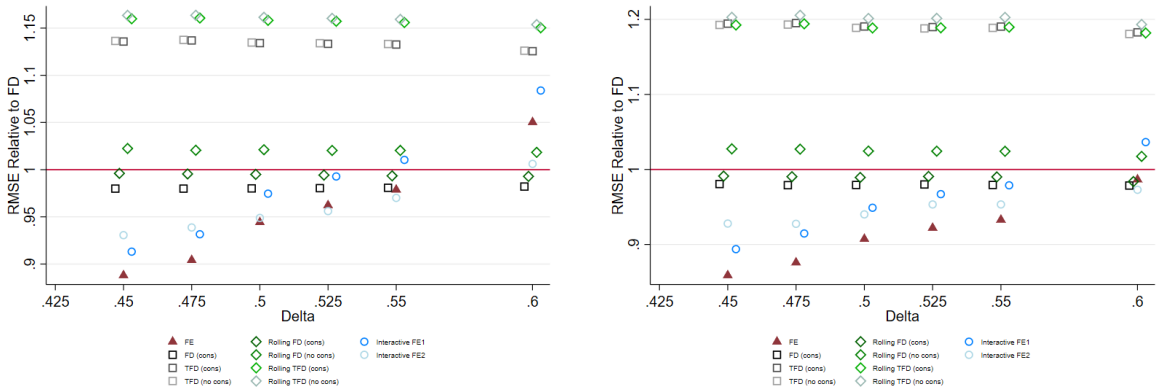
(B) $N = 500$



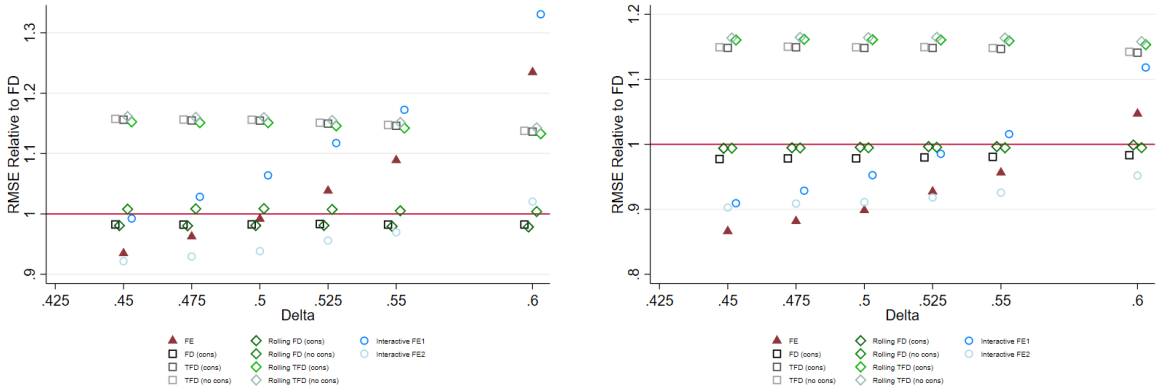
(C) $N = 5000$

FIGURE A.9: Simulation Results: Bias

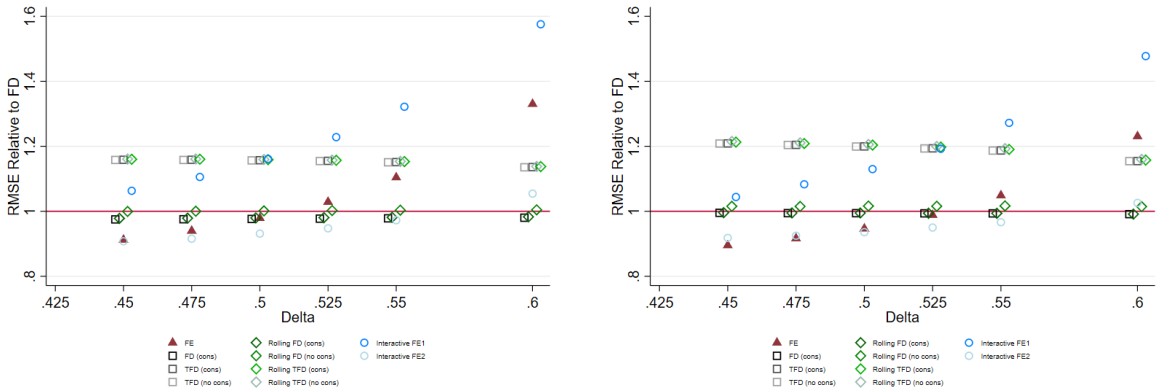
Notes: Delta is the fraction of units with time-varying fixed effects. Performance is relative to the first-difference estimator. DGP5 in left column. DGP6 in right column.



(A) $N = 100$



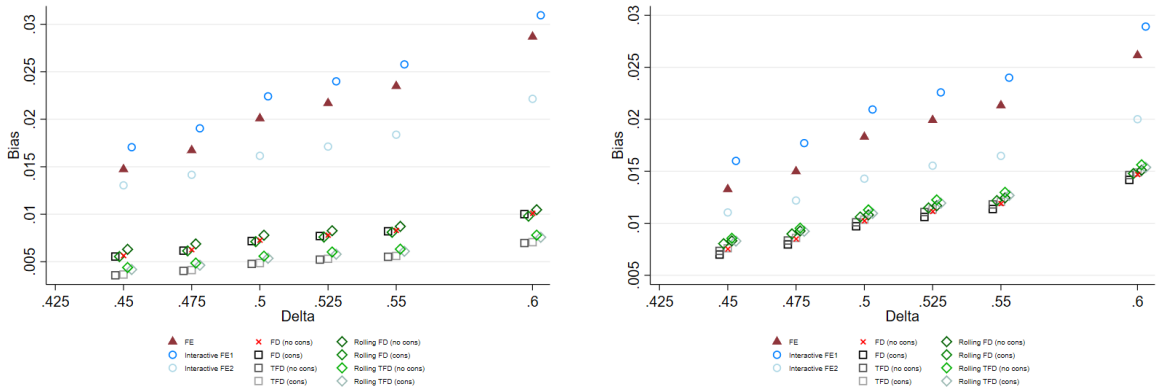
(B) $N = 500$



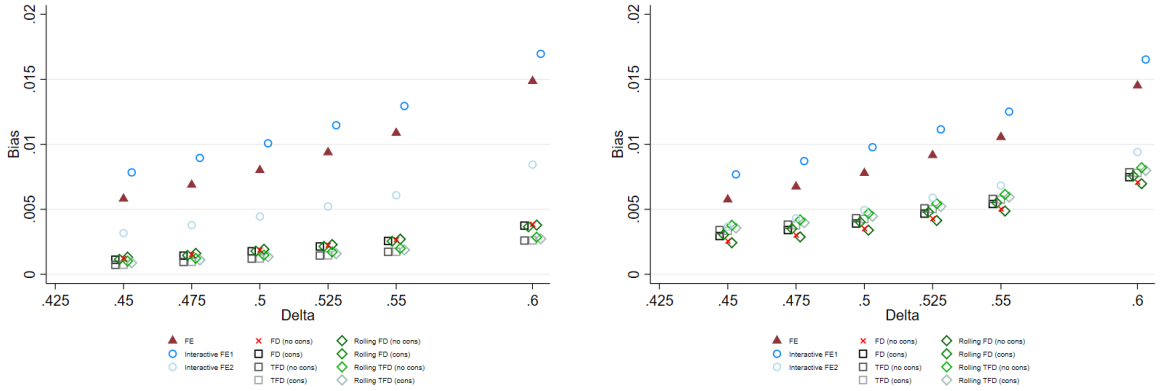
(C) $N = 5000$

FIGURE A.10: Simulation Results: Root Mean Squared Error

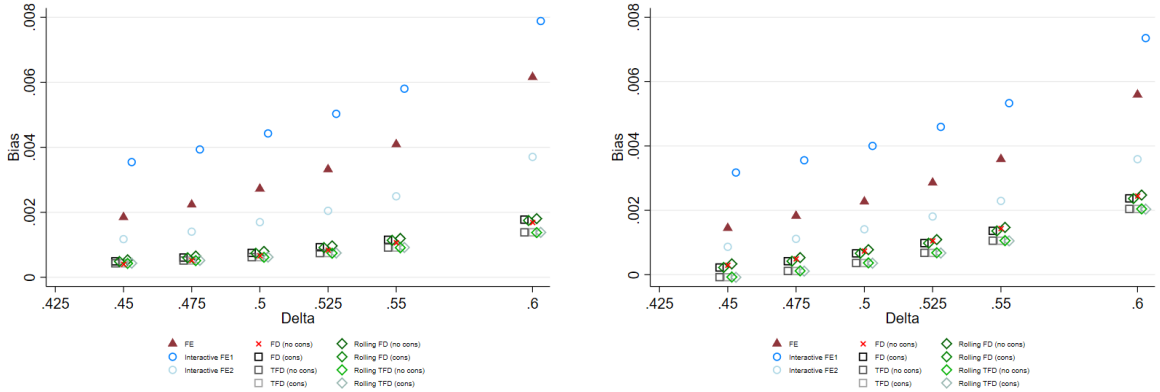
Notes: Delta is the fraction of units with time-varying fixed effects. Performance is relative to the first-difference estimator. DGP5 in left column. DGP6 in right column.



(A) $N = 100$



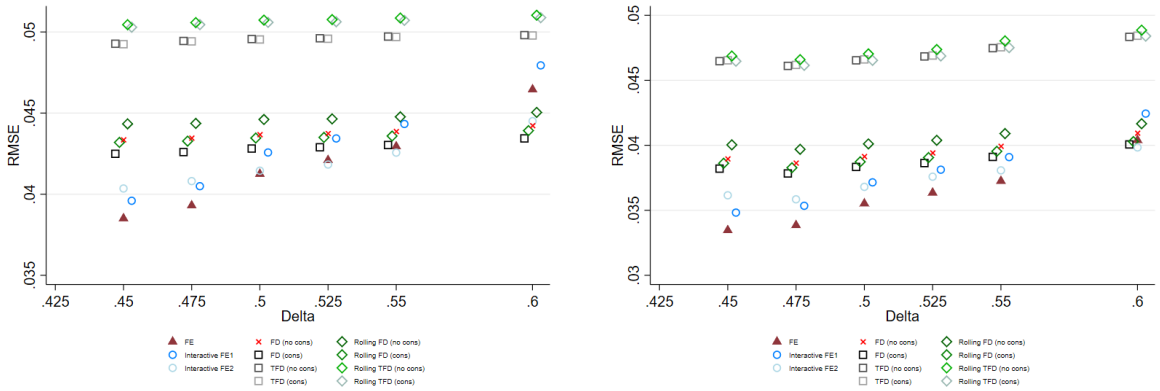
(B) $N = 500$



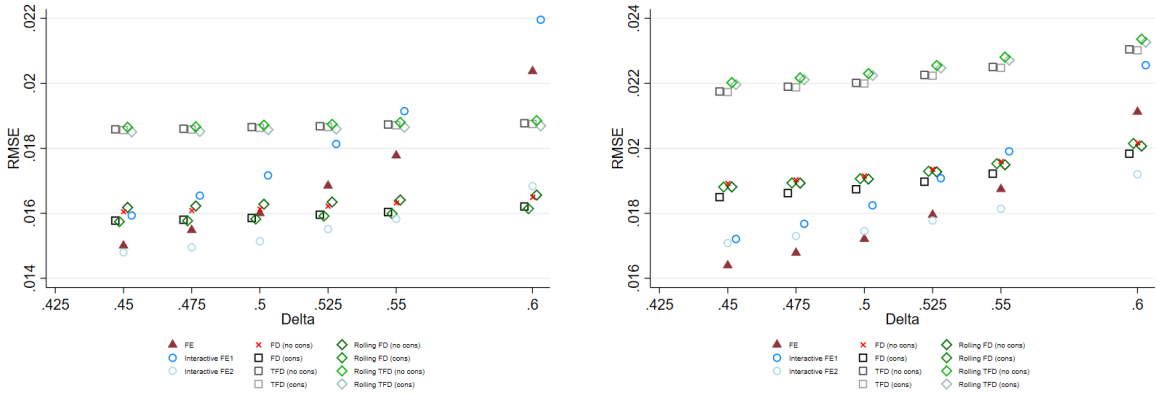
(C) $N = 5000$

FIGURE A.11: Simulation Results: Bias

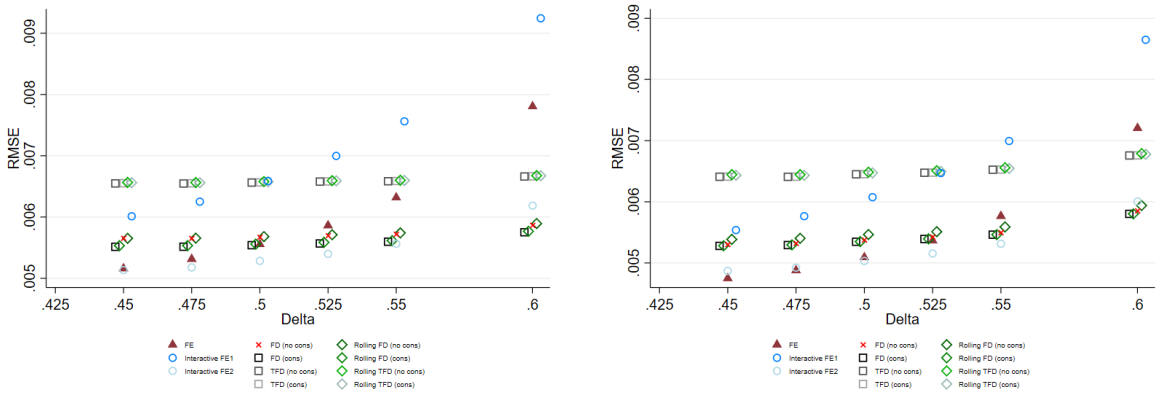
Notes: Delta is the fraction of units with time-varying fixed effects. DGP5 in left column. DGP6 in right column.



(A) $N = 100$



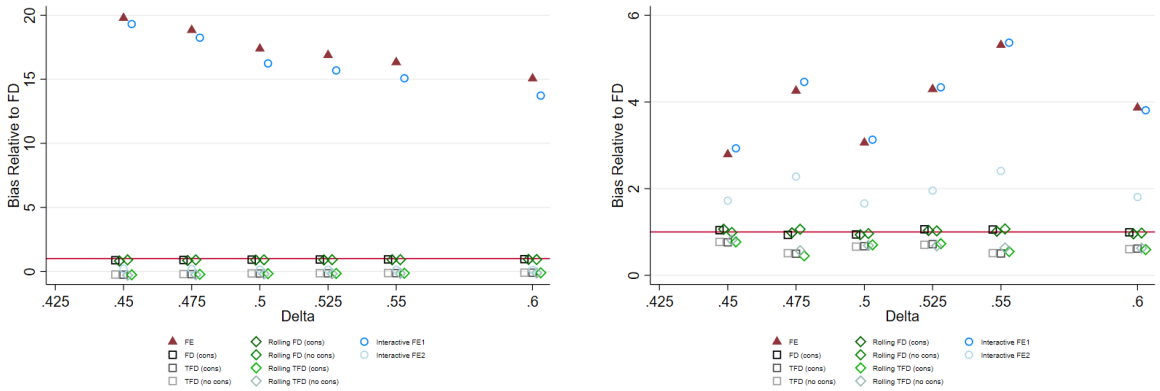
(B) $N = 500$



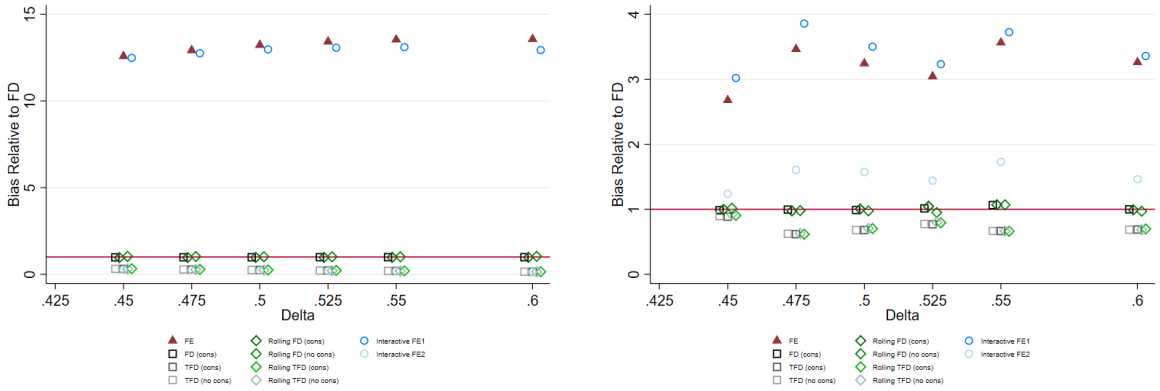
(C) $N = 5000$

FIGURE A.12: Simulation Results: Root Mean Squared Error

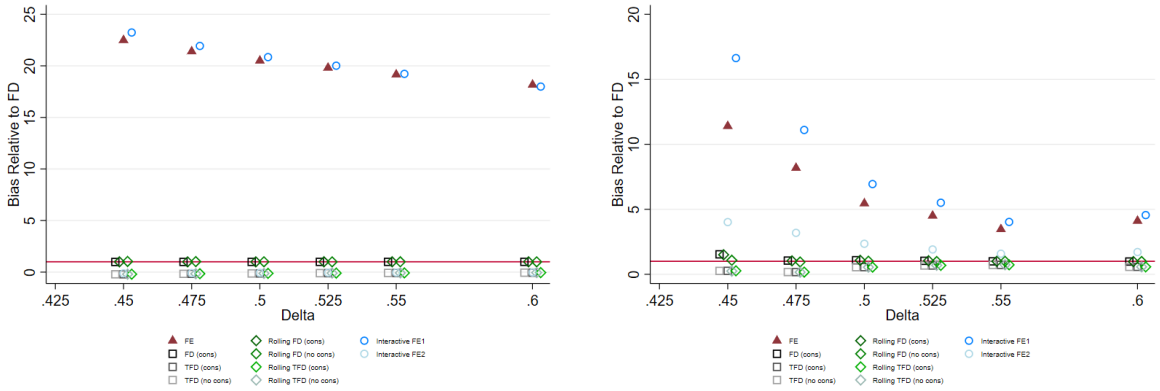
Notes: Delta is the fraction of units with time-varying fixed effects. DGP5 in left column. DGP6 in right column.



(A) $N = 100$



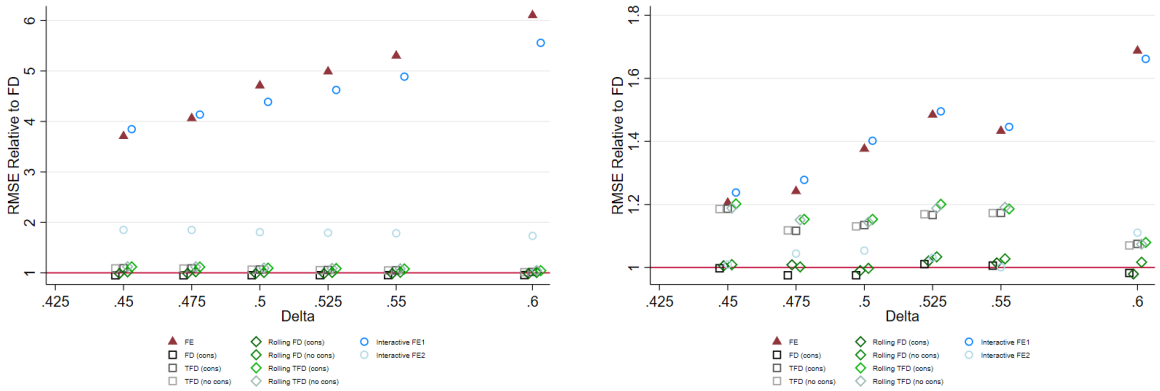
(B) $N = 500$



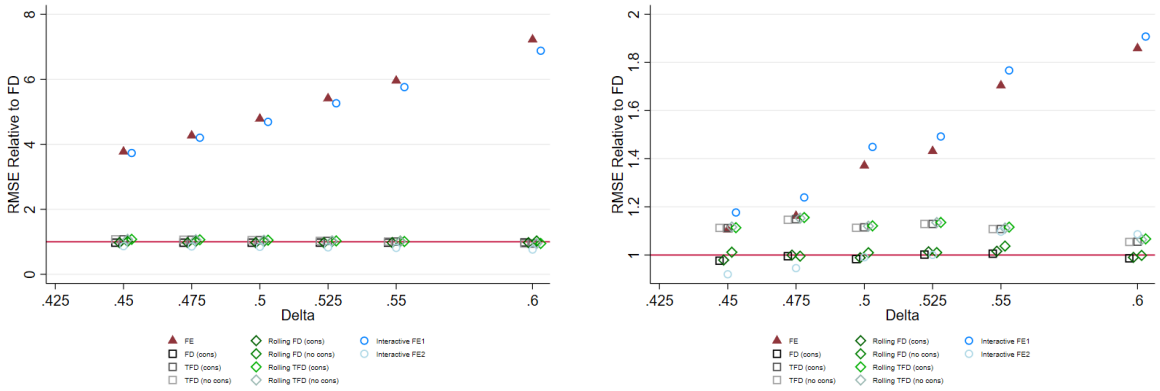
(C) $N = 5000$

FIGURE A.13: Simulation Results: Bias

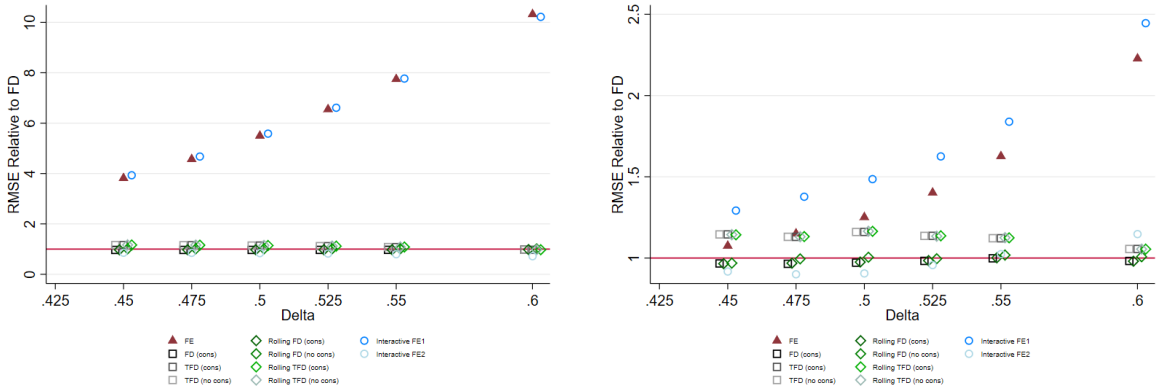
Notes: Delta is the fraction of units with time-varying fixed effects. Performance is relative to the first-difference estimator. DGP7 in left column. DGP8 in right column.



(A) $N = 100$



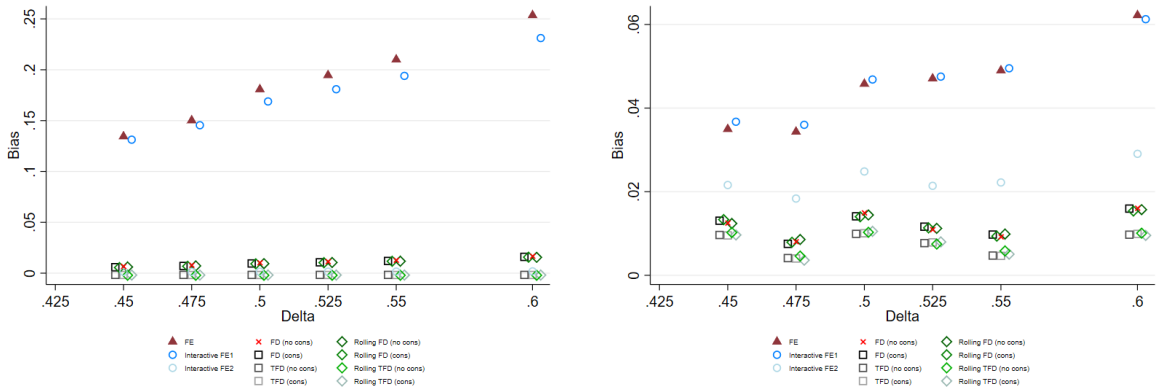
(B) $N = 500$



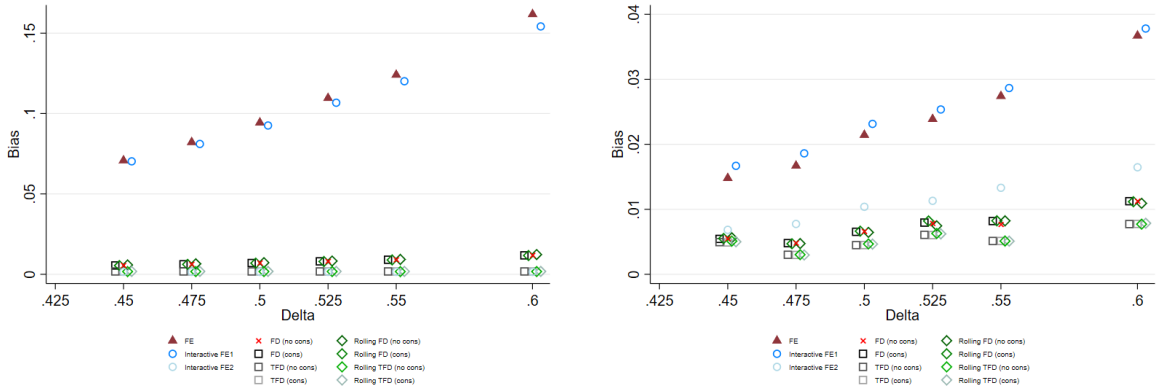
(C) $N = 5000$

FIGURE A.14: Simulation Results: Root Mean Squared Error

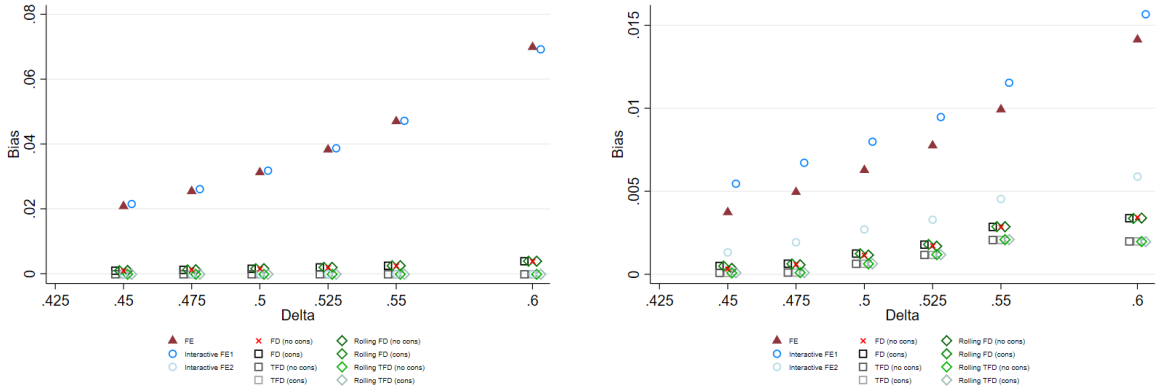
Notes: Delta is the fraction of units with time-varying fixed effects. Performance is relative to the first-difference estimator. DGP7 in left column. DGP8 in right column.



(A) $N = 100$



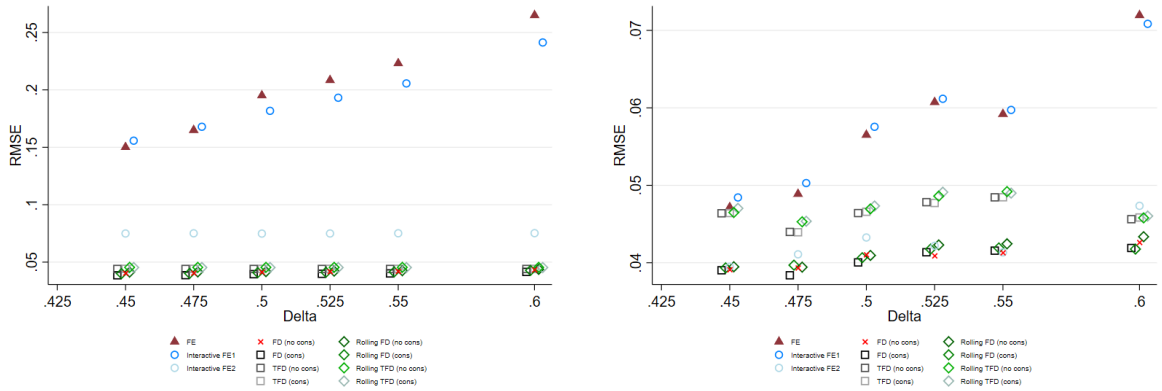
(B) $N = 500$



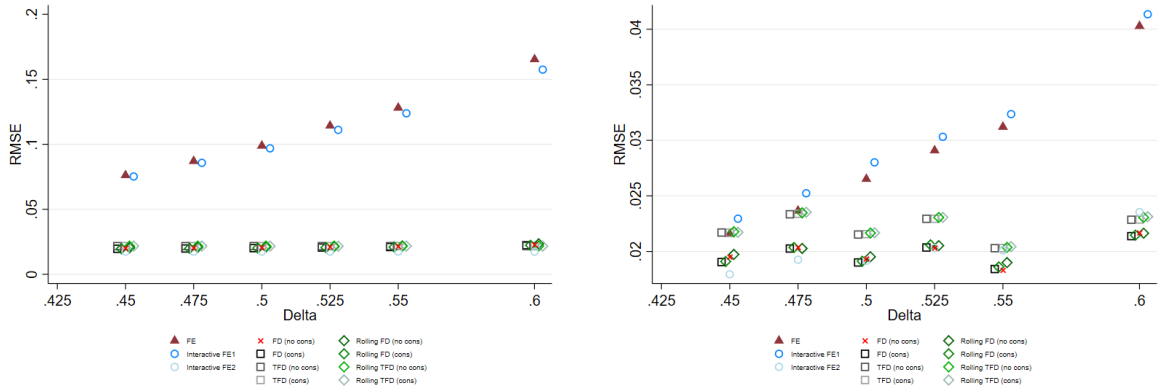
(C) $N = 5000$

FIGURE A.15: Simulation Results: Bias

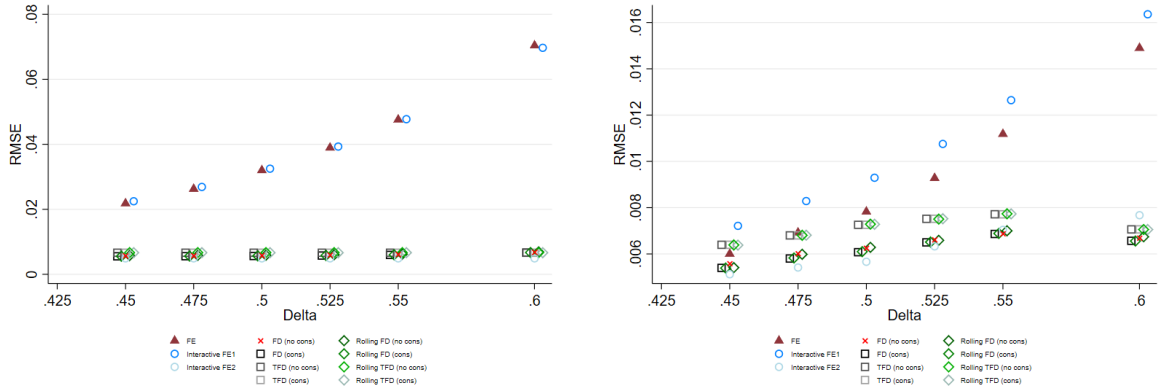
Notes: Delta is the fraction of units with time-varying fixed effects. DGP7 in left column. DGP8 in right column.



(A) $N = 100$



(B) $N = 500$



(C) $N = 5000$

FIGURE A.16: Simulation Results: Root Mean Squared Error

Notes: Delta is the fraction of units with time-varying fixed effects. DGP7 in left column. DGP8 in right column.