What Do We Get From A Two-Way Fixed Effects Estimator? Implications From A General Numerical Equivalence

Shoya Ishimaru*

March 24, 2021

Abstract

This paper shows that a two-way fixed effects (TWFE) estimator is a weighted average of first-difference (FD) estimators with different gaps between periods, generalizing a well-known equivalence theorem in a two-period panel. Exploiting the identity, I clarify required conditions for the causal interpretation of the TWFE estimator. I highlight its several limitations and propose a generalized estimator that overcomes the limitations. An empirical application on the estimates of the minimum wage effects illustrates that recognizing the numerical equivalence and making use of the generalized estimator enable more transparent understanding of what we get from the TWFE estimator.

*Hitotsubashi University, Department of Economics (email: shoya.ishimaru@r.hit-u.ac.jp).

1 Introduction

Increasingly many empirical studies employ a two-way fixed effects (TWFE) estimator to estimate the causal impact of a treatment on an outcome of interest. This estimator is usually motivated by a difference-in-differences (DID) identification logic, and it is in fact often called a DID estimator. In a two-period panel, the TWFE estimator is known to be numerically equivalent to a first-difference (FD) estimator, which more directly implements the DID identification idea. With a binary treatment, recent literature has uncovered that the TWFE estimator aggregates all possible two-period DID comparisons in a multiperiod panel (Strezhnev, 2018; Goodman-Bacon, 2020). However, the absence of a general numerical interpretation that applies to a multiperiod panel with a nonbinary treatment has limited our understanding of what the TWFE estimator generally identifies in the absence of functional form assumptions.

This paper establishes a general numerical property of the TWFE estimator. It is numerically equivalent to: (i) a weighted average of FD estimators with all possible gaps between periods; or alternatively, (ii) a weighted average of two-period TWFE estimators with all possible start and end periods. This result confirms an intuition that the TWFE estimator captures both short-run and long-run associations between changes in the outcome and changes in the treatment, which has been long sensed but has never been formalized.

The equivalence result enables a more transparent causal interpretation of the TWFE estimator by disaggregating the issue into what causal interpretation each constituent estimator has. I provide formal conditions required for the causal interpretation and highlight important limitations of the multiperiod TWFE estimator. First, it requires parallel trends during any duration of time. Second, it can use a limited set of conditioning variables to improve the plausibility of parallel trends assumption. Third, it restricts the conditioning variables' possible relationship with treatment changes.

I propose a generalized TWFE estimator that overcomes these limitations. The proposed estimator aggregates two-period estimators in the same manner as the original estimator but allows for a more flexible specification for each constituent two-period estimator. In particular, it can control for any variable observed before the starting period of each constituent estimator, as in a common identification strategy for a two-period DID (Heckman et al., 1997, 1998; Abadie, 2005). In addition, it can restrict the range of duration that requires parallel trends, excluding "too short" or "too long" duration.

I demonstrate the approach by exploring the TWFE estimates of the minimum wage effect on employment outcomes in the U.S. state–year panel. After confirming the numerical equivalence result, I illustrate the generalized TWFE estimator that I propose. While the standard TWFE estimate suggests no impact of the minimum wage on the net job creation rate, I compute a generalized estimator that restricts the constituent estimators to be based on 1- to 4-year gaps only. This resulting estimate is less likely to be influenced by the potential reverse causality problem that a falling job creation rate or a rising job destruction rate in an early part of the long-run gap causes the slower growth in the minimum wage in the later part. The estimate suggests that a 10% increase in the minimum wage is associated with a 0.24 percentage point decline in the net job creation rate. Another estimate that directly controls for outcome pre-trends, which is not possible for the standard TWFE estimator, indicates that these findings are robust to the difference in outcome pre-trends across states. While my analysis is not meant to provide the "best evidence" on the employment effects of the minimum wage, it contributes to the discussion by clarifying how the TWFE approach maps the panel data into the estimates and demonstrates how empirical researchers may modify this mapping to improve their estimates.

This paper is related to the recent growing literature on the TWFE estimator. The literature focuses almost exclusively on DID and event-study settings with a binary treatment, and intensively investigates properties of the standard TWFE estimator with univariate or multivariate binary regressors.[1] Most importantly, Strezhnev (2018) and Goodman-Bacon (2020) derive the numerical equivalence results similar to mine for a binary treatment. While their results provide important insights specific to the binary treatment setting, my focus is to provide a general result that applies to any TWFE regression and covers a much broader range of empirical settings. In addition, clarifying the required conditions for the causal interpretation of the TWFE estimator in the presence of covariates is a novel contribution of this paper, with a limited analysis in the literature even in the binary treatment setting.

More broadly, this paper is a part of continuing efforts of the econometric literature to improve the transparency of linear regression estimators. The literature has explored how researchers can interpret the regression estimators, the main workhorses of current empirical studies, when linear models that motivate them fail.[2] As in the literature, this paper provides both optimistic and cautionary views on the linear regression estimator: the TWFE estimator may have a general causal interpretation beyond the linear model, but it requires careful consideration of what it identifies under what assumptions.

[1] Athey and Imbens (2018), Callaway and Sant'Anna (2020), and Goodman-Bacon (2020) study staggered adoption settings of DID, while Strezhnev (2018) and de Chaisemartin and D'Haultfoeuille (2020) allow for more general paths of treatment assignments. Borusyak and Jaravel (2018), Schmidheiny and Siegloch (2020), and Sun and Abraham (2020) consider event-study settings.

[2] See Angrist and Krueger (1999) and Angrist and Pischke (2008) for the overview. Mogstad and Wiswall (2010), Løken et al. (2012), Lochner and Moretti (2015), Słoczyński (2020a,b), and Ishimaru (2021) are examples of recent takes.

2 Econometric Results

Section 2.1 shows a general numerical equivalence theorem on the TWFE estimator. Using the equivalence result, Section 2.2 clarifies necessary assumptions for the TWFE estimator to have a valid causal interpretation. Section 2.3 proposes a modification to the TWFE estimator to improve its causal interpretation.

2.1 Numerical Equivalence

I consider a balanced panel with observations (y_{it}, x_{it}) for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where y_{it} is a scalar outcome of interest and x_{it} is a scalar regressor. A TWFE estimator $\hat{\beta}^{FE}$ is the solution to a least-squares problem:

$$\min_{\{\alpha_i\}_{i=1}^{N}, \beta, \{\gamma_t\}_{t=1}^{T-1}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - \alpha_i - \beta x_{it} - \gamma_t\right)^2 .$$

On the other hand, an FD estimator $\hat{\beta}_k^{FD}$ that uses k-period gaps is the solution to

$$\min_{\beta, \{\gamma_t\}_{t=1}^{T-k}} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \left(\Delta_k y_{it} - \beta \Delta_k x_{it} - \gamma_t\right)^2 ,$$

where $\Delta_k v_{it} = v_{i,t+k} - v_{it}$ for $v = x, y$. Letting $k = 1$ gives a standard one-period FD estimator $\hat{\beta}_1^{FD}$, and it is well known that $\hat{\beta}^{FE} = \hat{\beta}_1^{FD}$ with $T = 2$. I establish a general numerical property of the TWFE estimator as follows.

Theorem 1. *A TWFE estimator $\hat{\beta}^{FE}$ is given by*

$$\hat{\beta}^{FE} = \frac{\sum_{s>t} \sum_{i=1}^{N} \left(\tilde{y}_{is} - \tilde{y}_{it}\right)\left(\tilde{x}_{is} - \tilde{x}_{it}\right)}{\sum_{s>t} \sum_{i=1}^{N} \left(\tilde{x}_{is} - \tilde{x}_{it}\right)^2},$$

where $\tilde{v}_{it} = v_{it} - \frac{1}{N} \sum_{j=1}^{N} v_{jt}$ for $v = x, y$.

Proof. See Appendix A.1. ☐

This theorem shows that the TWFE estimator can be obtained by a regression of changes in y_{it} on changes in x_{it} after removing common time effects, pooling all possible start and end periods. As a result, the TWFE estimator associates a difference in y_{it} between any two periods, which can be a short-run or long-run difference, with a difference in x_{it} between the same two periods.

A general numerical relationship between the TWFE estimator and the FD estimators with $T \geq 2$ immediately follows from Theorem 1.

Corollary 1. *$\hat{\beta}^{FE}$ is a convex combination of k-period FD estimators $\left\{\hat{\beta}_k^{FD}\right\}_{k=1}^{T-1}$. In particular,*

$$\hat{\beta}^{FE} = \sum_{k=1}^{T-1} \hat{\omega}_k^{FD} \hat{\beta}_k^{FD},$$

where the weights $\{\hat{\omega}_k^{FD}\}_{k=1}^{T-1}$ are given by

$$\hat{\omega}_k^{FD} = \frac{\sum_{t=1}^{T-k} \sum_{i=1}^{N} \left(\Delta_k \tilde{x}_{it}\right)^2}{\sum_{k'=1}^{T-1} \sum_{t'=1}^{T-k'} \sum_{i=1}^{N} \left(\Delta_{k'} \tilde{x}_{it'}\right)^2} \geq 0,$$

and satisfy $\sum_{k=1}^{T-1} \hat{\omega}_k^{FD} = 1$.

This result confirms that the identity $\hat{\beta}^{FE} = \hat{\beta}_1^{FD}$ does not hold for $T > 2$. When $\hat{\beta}^{FE}$ and $\hat{\beta}_1^{FD}$ substantially differ, it would be helpful to inspect how the k-period FD estimators $\hat{\beta}_k^{FD}$ vary across $k = 1, \ldots, T - K$ and how they are weighted.

Another corollary that follows from Theorem 1 shows that the "global" TWFE estimator $\hat{\beta}^{FE}$ can be written as a weighted average of "local" TWFE estimators.

Corollary 2. *Let $\hat{\beta}_{s,t}^{FE} = \frac{\sum_{i=1}^{N}(\tilde{y}_{is}-\tilde{y}_{it})(\tilde{x}_{is}-\tilde{x}_{it})}{\sum_{i=1}^{N}(\tilde{x}_{is}-\tilde{x}_{it})^2}$ be a TWFE estimator using only two periods $s > t$. Then,*

$$\hat{\beta}^{FE} = \sum_{s>t} \hat{\omega}_{s,t}^{FE} \hat{\beta}_{s,t}^{FE},$$

where the weights $\{\hat{\omega}_{s,t}^{FE}\}_{s>t}$ are given by

$$\hat{\omega}_{s,t}^{FE} = \frac{\sum_{i=1}^{N} \left(\tilde{x}_{is} - \tilde{x}_{it}\right)^2}{\sum_{s'>t'} \sum_{i=1}^{N} \left(\tilde{x}_{is'} - \tilde{x}_{it'}\right)^2} \geq 0,$$

and satisfy $\sum_{s>t} \hat{\omega}_{s,t}^{FE} = 1$.

Therefore, the TWFE estimator can be computed by aggregating all possible two-period TWFE estimators $\hat{\beta}_{s,t}^{FE}$ with the first period t and the second period $s > t$. In the aggregation, each $\hat{\beta}_{s,t}^{FE}$ is weighted by the sample variance of $x_{is} - x_{it}$.

In a special case with a binary x_{it}, Strezhnev (2018) and Goodman-Bacon (2020) derive analogous results to Theorem 1 and its corollaries. However, with a nonbinary x_{it}, this paper is the first (to the best of my knowledge) to establish the numerical relationship of the TWFE estimator with constituent short-run and long-run comparisons. Extensions to even more general settings, such as a multivariate x_{it} and an instrumental variables (IV) regression, are presented in Appendix A.1.

A result with covariates can be derived by replacing (y_{it}, x_{it}) with residualized ones using the Frisch–Waugh–Lovell (FWL) theorem. However, with residuals, the causal interpretation of the result becomes less transparent. Therefore, the next subsection explores the causal interpretation of the TWFE estimator in the presence of covariates.

2.2 Causal Interpretation

A numerical interpretation of the TWFE estimator established by Theorem 1 facilitates its causal interpretation. To consider the population TWFE coefficient with $N \to \infty$ and T being fixed, let $(y_{it}, x_{it}, w_{it})_{t=1}^{T}$ be observations of a scalar outcome y_{it}, a scalar treatment x_{it}, and a vector of covariates w_{it}, i.i.d. across $i = 1, \ldots, N$.[3] A de-meaned value of any variable v_{it} is denoted as $\tilde{v}_{it} = v_{it} - E[v_{it}]$. Let $r_{it} = \tilde{x}_{it} - \tilde{w}_{it}' \delta^{FE}$, where δ^{FE} is the population coefficient from a TWFE regression of x_{it} on w_{it}. The FWL theorem and Theorem 1 imply that the population coefficient of x_{it} from a TWFE regression of y_{it} on x_{it} and w_{it} is

$$\beta^{FE} = \frac{\sum_{k=1}^{T-1} \sum_{t=1}^{T-k} E[\Delta_k y_{it} \Delta_k r_{it}]}{\sum_{k=1}^{T-1} \sum_{t=1}^{T-k} E[\Delta_k x_{it} \Delta_k r_{it}]}. \tag{1}$$

I make the following assumptions to provide the causal interpretation of β^{FE}.

Assumption 1. *The observed outcome is given by $y_{it} = y_{it}(x_{it})$ for each t, where $y_{it}(x)$ is a potential outcome with a treatment x assigned.*

Assumption 2. $Cov(y_{i,t+k}(x_{it}) - y_{it}(x_{it}), \Delta_k x_{it} | \Delta_k w_{it}) = 0$ *for any* (k, t).

Assumption 3. $\sum_{k=1}^{T-1} \sum_{t=1}^{T-k} E[Var(\Delta_k x_{it} | \Delta_k w_{it})] > 0$.

Assumption 4. $E[\Delta_k \tilde{x}_{it} | \Delta_k w_{it}] = \Delta_k \tilde{w}_{it}' \delta_{kt}$ *for any* (k, t).

Assumption 1 rules out dynamic treatment effects by restricting the potential outcome to be a function only of the current treatment status. Assumption 2 is a variant of "parallel trends" assumption, and it means that a change in the treatment status $\Delta_k x_{it}$ is uncorrelated with a potential outcome trend associated with the current treatment status, conditional on changes in covariates $\Delta_k w_{it}$.[4] Assumption 3 requires that $\Delta_k x_{it}$ have some variation conditional on $\Delta_k w_{it}$. Assumption 4 requires that the conditional mean $E(\Delta_k x_{it} | \Delta_k w_{it})$ be

[3] In this analysis, the number of covariates is constant; only the number of fixed effects $\{\alpha_i\}_{i=1}^{N}$ grows with N. It rules out, for example, a specification with a linear or polynomial trend specific to each i. In general, such specification has a valid interpretation only when the specified linear regression model is true.

[4] de Chaisemartin and D'Haultfoeuille (2020) allow for covariates when they extend their analysis of the binary treatment setting. They make a stronger and less intuitive parallel trends assumption that requires $E[y_{i,t+k}(0) - y_{it}(0) | x_{i1}, \ldots, x_{iT}, w_{i1}, \ldots, w_{iT}]$ to be linear in $\Delta_k w_{it}$ and not to depend on the other elements of $(x_{i1}, \ldots, x_{iT}, w_{i1}, \ldots, w_{iT})$.

linear in $\Delta_k w_{it}$. Similar linearity assumptions have been commonly used in interpreting linear regression coefficients in the presence of covariates.[5] This assumption is not as restrictive as it may appear. For example, it mechanically holds if w_{it} consists of disjoint group dummies. Even if covariates include continuous variables, adding higher-order terms and interactions of the original covariates to w_{it} can make this assumption as approximate as possible with sufficient data.

Theorem 2. *Under Assumptions 1–4,*

$$\beta_{FE} = \frac{\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} E\left[\tau_{ikt}\Delta_k x_{it}\Delta_k r_{it} + \left(y_{i,t+k}(x_{it}) - y_{it}(x_{it})\right)\Delta_k \tilde{w}_{it}'\left(\delta_{k,t} - \delta^{FE}\right)\right]}{\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} E\left[\Delta_k x_{it}\Delta_k r_{it}\right]}, \tag{2}$$

where

$$\tau_{ikt} = \begin{cases} E\left[\frac{y_{i,t+k}(x_{i,t+k}) - y_{i,t+k}(x_{it})}{\Delta_k x_{it}} | \Delta_k x_{it}, \Delta_k w_{it}\right] & \Delta_k x_{it} \neq 0 \\ 0 & \Delta_k x_{it} = 0 \end{cases}$$

is the expected per-unit treatment effect of a treatment status change $\Delta_k x_{it}$. The equation (2) can be rewritten as

$$\beta^{FE} = \sum_{k=1}^{T-1}\sum_{t=1}^{T-k} E\left[\tau_{ikt}\omega_{ikt}\right], \quad \omega_{ikt} = \frac{\Delta_k x_{it}\Delta_k r_{it}}{\sum_{k'=1}^{T-1}\sum_{t'=1}^{T-k'} E\left[\Delta_{k'} x_{it'}\Delta_{k'} r_{it'}\right]}, \tag{3}$$

if one of the following conditions holds:

1. $\delta_{kt} = \delta^{FE}$ for any (k, t).

2. w_{it} is given by interacting time-invariant covariates with time dummies, i.e., $w_{it} = (\mathbb{1}_{t=1} w_i' \cdots \mathbb{1}_{t=T-1} w_i')'$.

3. $Cov\left(y_{i,t+k}(x_{it}) - y_{it}(x_{it}), \Delta_k w_{it}\right) = 0$ for any (k, t).

Proof. See Appendix A.2. □

This result implies that the TWFE coefficient β^{FE} can be interpreted as the weighted average of per-unit treatment effects under the required assumptions and conditions.[6] The second term of the equation (2) implies that approximation errors can arise from the fact that

[5] See, for example, Angrist and Krueger (1999), Lochner and Moretti (2015), Słoczyński (2020b), and Ishimaru (2021).

[6] As is common among other studies that investigate the causal interpretation of the TWFE coefficient, the weights ω_{ikt} are not guaranteed to be nonnegative. See, for example, Borusyak and Jaravel (2018), de Chaisemartin and D'Haultfoeuille (2020), Goodman-Bacon (2020), Sun and Abraham (2020), and Ishimaru (2021).

the relationship between treatment changes $\Delta_k x_{it}$ and covariate changes $\Delta_k w_{it}$ may not be homogeneous across time.[7] The failure of any of Assumptions 1–4 can give rise to a further bias. For example, Assumption 2 may not be plausible with large k, if long-run trends are less likely to be parallel. In addition, conditioning on $\Delta_k w_{it}$ may not be sufficient to make parallel trends plausible. Most importantly, $\Delta_k w_{it}$ cannot include information about variables observed before period t in a natural manner, except for time-invariant observables.[8] These concerns motivate modifications to the TWFE estimator, which are proposed below.

2.3 Generalizations of the TWFE Estimator

Numerical representation of the TWFE estimator suggested in Section 2.1 and required set of assumptions for the causal interpretation suggested in Section 2.2 motivate the following natural extensions of the TWFE estimator.

Restriction on k

While the TWFE estimator $\hat{\beta}^{FE}$ is a convex combination of k-period FD estimators $\hat{\beta}_k^{FD}$, using "long" difference as its component may not be desirable if long-run parallel trends assumption is too strong. For example, a reverse causality problem can arise if outcome changes during the early part in the k-period gap induces treatment changes during the later part. Also, "short" first difference estimators can be attenuated due to frictional adjustment in response to the treatment change. These concerns motivate restricting the range of constituent FD estimators as

$$\hat{\beta}^* \equiv \frac{\sum_{k=\underline{K}}^{\overline{K}} \hat{\omega}_k^{FD} \hat{\beta}_k^{FD}}{\sum_{k=\underline{K}}^{\overline{K}} \hat{\omega}_k^{FD}} = \frac{\sum_{k=\underline{K}}^{\overline{K}} \sum_{t=1}^{T-k} \sum_{i=1}^{N} \Delta_k \tilde{y}_{it} \Delta_k \tilde{x}_{it}}{\sum_{k=\underline{K}}^{\overline{K}} \sum_{t=1}^{T-k} \sum_{i=1}^{N} \Delta_k \tilde{x}_{it}^2}, \tag{4}$$

where $1 \leq \underline{K} \leq \overline{K} \leq T-1$. The original TWFE estimator $\hat{\beta}^{FE}$ corresponds to a special case with $\underline{K} = 1$ and $\overline{K} = T-1$.

Additional Controls

As illustrated in Section 2.2, the standard TWFE estimator can only allow for a specific form of conditional parallel trends assumption (i.e., parallel trends conditional on time-invariant

[7] As the third condition indicates, a trivial exception is when potential trends cannot be predicted by covariates.

[8] "Parallel trends conditional on observables before t" is a common identification strategy for a two-period DID (Heckman et al., 1997, 1998; Abadie, 2005). A naive approach would include a lag of some variable, say $v_{i,t-1}$, as a part of w_{it}. This approach would not work because trends would have to be conditioned on $v_{i,t+k-1} - v_{i,t-1}$, which includes information about observables after t.

covariates or changes in time-varying covariates). On the other hand, a two-period DID identification strategy often uses variables observed before the first period as conditioning variables. Given Corollary 2 shows that the TWFE estimator $\hat{\beta}^{FE}$ is the weighted average of two-period estimators $\hat{\beta}^{FE}_{s,t}$, the following procedure would give a natural generalization of the TWFE estimator $\hat{\beta}^{FE}$.

1. For each (s,t) with $s > t$, regress $y_{is} - y_{it}$ on $x_{is} - x_{it}$ with some covariates, such as time-invariant variables, changes in time-dependent variables between periods s and t, and most importantly, time-dependent variables observed until period t. Let $\hat{\beta}^*_{s,t}$ be the coefficient of $x_{is} - x_{it}$. Set the weights $\hat{\omega}^*_{s,t}$ to be proportional to the sum of squared residuals from a regression of $x_{is} - x_{it}$ on control variables.[9] In particular,

$$\hat{\beta}^*_{s,t} = \frac{\sum_{i=1}^N r^x_{ist} r^y_{ist}}{\sum_{i=1}^N \left(r^x_{ist}\right)^2}, \quad \hat{\omega}^*_{s,t} = \frac{\sum_{i=1}^N \left(r^x_{ist}\right)^2}{\sum_{s'>t'} \sum_{i=1}^N \left(r^x_{is't'}\right)^2},$$

   where r^v_{ist} $(v = x, y)$ is a residual from a regression of $v_{is} - v_{it}$ on covariates, separately performed for each (s,t).

2. Construct the modified estimator as

$$\hat{\beta}^* = \sum_{s>t} \hat{\omega}^*_{s,t} \hat{\beta}^*_{s,t} = \frac{\sum_{s>t} \sum_{i=1}^N r^x_{ist} r^y_{ist}}{\sum_{s>t} \sum_{i=1}^N \left(r^x_{ist}\right)^2}, \tag{5}$$

The resulting estimator $\hat{\beta}^*$ has a valid causal interpretation if each $\hat{\beta}^*_{s,t}$ satisfies Assumptions 1–4. Note that each $\hat{\beta}^*_{s,t}$, being a two-period TWFE estimator, is free from the approximation error between (2) and (3) associated with time homogeneity conditions. This estimator can be combined with the restriction on the gap duration: $\underline{\mathrm{K}} \leq s - t \leq \overline{K}$.

3 Empirical Illustration

This section illustrates the econometric results and approaches in Section 2 using the TWFE estimates of the impact of minimum wages on employment. As Manning (2021) remarks, the analysis based on state-level panel data in the United States has attracted much attention in the literature for decades. I use this common empirical setting for illustrating the points in Section 2 in an intuitive manner, but not for advocating the "best evidence" in this contentious topic.

[9] Alternatively, the original weights $\hat{\omega}^{FE}_{s,t}$, which is proportional to the sum of squared residuals from a regression of $x_{is} - x_{it}$ on a constant, can be used to aggregate $\hat{\beta}^*_{s,t}$. It would make the new estimator $\hat{\beta}^*$ subject to exactly the same weights as the original estimator $\hat{\beta}^{FE}$.

To present the estimates, I use a state–year panel of employment outcomes and minimum wage laws in 50 U.S. states and D.C. in 1990–2018. The outcomes are the log employment rates of teens (age 16–19), young workers (age 20–34), and experienced workers (age 35–54), the gross job creation and destruction rates, and the net job creation rate.[10] The employment rates are given by the ratios of the civilian employment to the civilian population, aggregating the basic monthly Current Population Survey (CPS). The job creation and destruction rates are from the Business Dynamics Statistics (BDS). The minimum wage data is from Vaghul and Zipperer (2019).

I start by verifying the numerical equivalence result established in Section 2.1. Figure 1 illustrates Corollary 1 by presenting FD estimates and how they are weighted, from a regression of the log employment rate of teens on the log minimum wage. The FD estimate is -0.025 with a 1-year gap and decreases as the gap gets longer. The weighted average of all FD estimates is -0.133, which exactly matches the TWFE estimate. I also confirm the numerical equivalence using other outcomes.

Table 1 confirms Corollary 2. The first column of the table shows the standard TWFE estimates $\hat{\beta}^{FE}$, while the subsequent columns describe the distribution of all possible two-period TWFE estimates $\hat{\beta}^{FE}_{s,t}$. Each $\hat{\beta}^{FE}_{s,t}$ is treated as one data point in the distribution and weighted by $\hat{\omega}^{FE}_{s,t}$. Because each two-period estimate $\hat{\beta}^{FE}_{s,t}$ is computed from 51 observations, the estimates have substantial variation. Nevertheless, the weighted average exactly matches the standard TWFE estimate $\hat{\beta}^{FE}$.

Then, I explore generalized estimators proposed in Section 2.3. Table 2 presents the estimator associated with each employment outcome restricting the gap duration k as in (4). As already illustrated in Figure 1, the estimated impact on the log teen employment rate is more sizable with "long" differences. Possible interpretations of this finding are: (i) sluggish short-run response due to the fact that employment rate is "stock" rather than "flow" (Dube et al., 2016); or (ii) failure of parallel trends of potential employment rate in the long run. Patterns of estimates for employment rates for older workers can provide an informal placebo test to evaluate the second interpretation. These estimates are consistently much closer to 0 compared with the estimates for teen employment, throughout possible ranges of the gap duration k. These pieces of evidence are at odds with the second interpretation, although potential trends of teen employment may not be directly comparable with those of older workers.

Estimates for the job creation and destruction rates also exhibit interesting patterns.

[10] I use the log employment rates following a common specification in the literature. The job creation and destruction rates are in percentage points without the log transformation because the net job creation rates are often negative.

The short-run estimates suggest the minimum wage increase is associated with the decline in the job creation rate and the increase in the job destruction rate. On the other hand, the long-run estimates indicate the opposite relationships. These give rise to a statistically insignificant effect of the minimum wage on the net job creation rate implied by the standard TWFE estimator, with a positive point estimate.

The sluggish short-run response is less likely for the job creation and destruction rates because they are "flow" variables. Moreover, the positive long-run "effects" on the net job creation are difficult to interpret as causal because they mostly come from the decline in the gross job destruction rates. If a higher minimum wage increases the net job creation rate through the increase in labor supply, the effects should rather appear as the increase in the gross job creation rates. Therefore, a more likely explanation for the inconsistency between short-run and long-run estimates is the failure of long-run parallel trends due to reverse causality. In particular, suppose that the decline in the job creation rate or the increase in the job destruction rate in the first half of the long-run gap discourages state legislature from raising the minimum wage in the second half. Then, it would induce a positive bias for the long-run "effect" on the job creation and a negative bias for the one on the job destruction. Therefore, the short-run estimates are expected to reflect the causal effects of the minimum wage increase better.

Even with the focus on the short-run estimates, the estimated minimum wage impact on the job destruction is statistically insignificant and quantitatively weaker compared with the impact on the job creation. For example, the estimates based on 1- to 4-year gaps imply a 0.19 percentage point decline in the job creation rate but a 0.05 percentage point increase in the job destruction rate associated with a 10% increase in the minimum wage.[11] The difference in the magnitude of the impacts can be explained by the labor market friction.[12] In a frictional labor market, firms get a positive marginal surplus from each worker unless they are currently hit by substantial negative productivity shocks. Therefore, they have no incentive to reduce employment unless the minimum wage increase is so substantial that the marginal surplus of the firms becomes negative. On the other hand, a smaller marginal surplus immediately reduces their incentive for job creation.[13] These give rise to asymmetric responses of the job creation and destruction to the minimum wage increase.

The estimates presented thus far do not use covariates, and an unconditional parallel

[11] Combining two channels, a 10% minimum wage increase is predicted to reduce the net job creation rate by a 0.24 percentage point. It is not a trivial impact, given that the average annual net job creation rate is 1.33% during the same period and that not all newly employed workers are near-minimum wage workers.

[12] Van den Berg and Ridder (1998), Acemoglu (2001), Flinn (2006), and Ahn et al. (2011) are examples of studies that explore the impact of the minimum wage in a frictional labor market.

[13] Of course, the impact on the equilibrium job creation is theoretically ambiguous due to labor supply response such as the increase in search efforts.

trends assumption can be too strong even if the range of the gap duration k is appropriately restricted. One of the potential concerns is that trends of potential outcomes in states with faster minimum wage growths may be different from states with slower minimum wage increases. While "controlling" for state-specific linear time trends has been a common practice in the minimum wage literature, Meer and West (2016) point out that such specification may induce bias by fitting post-treatment dynamics as well as pre-treatment trends. Generalized estimators I propose, on the other hand, can directly control for pre-treatment trends without being influenced by post-treatment dynamics. In particular, for each state i and year t, I perform a regression of employment outcomes $y_{i,t-12}, \ldots, y_{i,t-3}$ on calendar years $t-12, \ldots, t-3$.[14] I exclude preceding 3-year periods not to introduce bias associated with the mechanical correlation of them with an outcome y_{it}. Then, I use the predicted pre-trend as a covariate for each two-period TWFE estimate $\hat{\beta}^*_{s,t}$ that constitutes a generalized TWFE estimate as defined in (5). Table 3 presents the estimates with and without restrictions on the gap duration. Each estimate is fairly comparable to the corresponding estimate in Table 2, suggesting that empirical findings from Table 2 are robust to the difference in pre-treatment trends across states.

4 Conclusion

Economists have known little about the properties of the TWFE estimator outside the linear regression model in a textbook, despite its important role in empirical studies. This paper establishes a general numerical equivalence theorem that the multiperiod TWFE estimator is the weighted average of simpler estimators, FD or two-period TWFE estimators. This finding gives much more transparency to the process of how the TWFE estimator maps the data into the estimated coefficient, enabling future empirical and econometric research to understand what we get from the TWFE estimator better.

The increased transparency is exemplified by this paper's own analysis of the causal interpretation of the TWFE estimator using the equivalence result. Required conditions for the causal interpretation under the potential framework clarified in the analysis reveal several important limitations of the standard multiperiod TWFE estimator: requiring parallel trends during any duration of time; accommodating a limited set of conditioning variables; and requiring the relationship of covariates with the treatment to be homogeneous across time.

I propose a natural generalization of the TWFE estimator that overcomes these limi-

[14] I use the data in 1978–1989, which are not used in the main analysis, to estimate state-specific pre-trends for earlier years in the panel. For example, the pre-trends for 1990 in the teen employment regression are based on the teen employment data in 1978–1987. I also try shorter pre-trends as controls and confirm that estimates are not sensitive to the choice of pre-trend lengths.

tations, by making modifications to each two-period TWFE estimator that comprises the multiperiod TWFE estimator. Application of the estimator to study the effect of the minimum wage finds that the minimum wage increase is, even after controlling for pre-trends, associated with the decline in the net job creation rate disproportionally through the decline in the gross job creation, which would have been not evident only with the standard TWFE estimator. These results suggest that empirical researchers should carefully think about how the TWFE estimator summarizes the panel data and consider using the generalized estimator to improve the robustness of their empirical results.

References

Abadie, A. (2005): "Semiparametric difference-in-differences estimators," *The Review of Economic Studies*, 72, 1–19.

Acemoglu, D. (2001): "Good jobs versus bad jobs," *Journal of Labor Economics*, 19, 1–21.

Ahn, T., P. Arcidiacono, and W. Wessels (2011): "The distributional impacts of minimum wage increases when both labor supply and labor demand are endogenous," *Journal of Business & Economic Statistics*, 29, 12–23.

Angrist, J. D., and A. B. Krueger (1999): "Empirical strategies in labor economics," *Handbook of Labor Economics*, 3, 1277–1366.

Angrist, J. D., and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*: Princeton University Press.

Athey, S., and G. W. Imbens (2018): "Design-based analysis in difference-in-differences settings with staggered adoption," *Working Paper*.

Van den Berg, G. J., and G. Ridder (1998): "An empirical equilibrium search model of the labor market," *Econometrica*, 1183–1221.

Borusyak, K., and X. Jaravel (2018): "Revisiting Event Study Designs," *Working Paper*.

Callaway, B., and P. H. Sant'Anna (2020): "Difference-in-differences with multiple time periods," *Journal of Econometrics*.

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

Dube, A., T. W. Lester, and M. Reich (2016): "Minimum wage shocks, employment flows, and labor market frictions," *Journal of Labor Economics*, 34, 663–704.

Flinn, C. J. (2006): "Minimum wage effects on labor market outcomes under search, matching, and endogenous contact rates," *Econometrica*, 74, 1013–1062.

Goodman-Bacon, A. (2020): "Difference-in-differences with variation in treatment timing," *Working Paper*.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an econometric evaluation estimator," *The Review of Economic Studies*, 65, 261–294.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," *The Review of Economic Studies*, 64, 605–654.

Ishimaru, S. (2021): "Empirical Decomposition of the IV-OLS Gap with Heterogeneous and Nonlinear Effects," *Working Paper*.

Lee, A. J. (1990): *U-statistics: Theory and Practice*: Marcel Dekker Inc.

Lochner, L., and E. Moretti (2015): "Estimating and testing models with many treatment levels and limited instruments," *Review of Economics and Statistics*, 97, 387–397.

Løken, K. V., M. Mogstad, and M. Wiswall (2012): "What Linear Estimators Miss: The Effects of Family Income on Child Outcomes," *American Economic Journal: Applied Economics*, 4, 1–35.

Manning, A. (2021): "The Elusive Employment Effect of the Minimum Wage," *Journal of Economic Perspectives*, 35, 3–26.

Meer, J., and J. West (2016): "Effects of the minimum wage on employment dynamics," *Journal of Human Resources*, 51, 500–522.

Mogstad, M., and M. Wiswall (2010): "Linearity in instrumental variables estimation: Problems and solutions," *IZA Discussion Paper No. 5216*.

Schmidheiny, K., and S. Siegloch (2020): "On event studies and distributed-lags in two-way fixed effects models: Identification, equivalence, and generalization," *ZEW-Centre for European Economic Research Discussion Paper*.

SŁOCZYŃSKI, T. (2020a): "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *Review of Economics and Statistics*, 1–27.

——— (2020b): "When Should We (Not) Interpret Linear IV Estimands as LATE?" *Working Paper*.

STREZHNEV, A. (2018): "Semiparametric weighting estimators for multi-period difference-in-differences designs," *Working Paper*.

SUN, L., AND S. ABRAHAM (2020): "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*.

VAGHUL, K., AND B. ZIPPERER (2019): "Historical State and Sub-state Minimum Wages," https://github.com/benzipperer/historicalminwage/releases/tag/v1.2.0, Version 1.2.0.

Figures and Tables

Figure 1: The TWFE Coefficient as the Weighted Average of the FD Coefficients

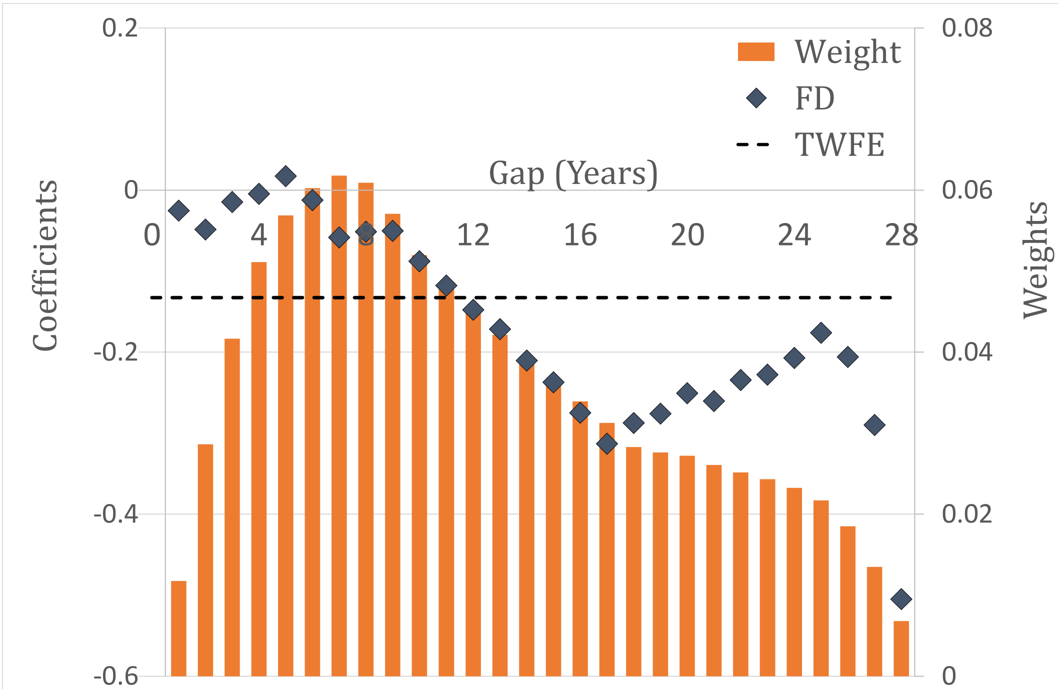

Table 1: The Distribution of Two-Period TWFE Estimates

	Standard TWFE	Mean	S.D.	P5	P25	Median	P75	P95
Log Employment, Age 16–19	–0.133	–0.133	0.236	–0.504	–0.235	–0.131	0.005	0.216
Log Employment, Age 20–34	0.014	0.014	0.066	–0.091	–0.010	0.023	0.051	0.090
Log Employment, Age 35–54	–0.003	–0.003	0.044	–0.090	–0.022	0.003	0.028	0.050
Job Creation Rate (%)	–1.06	–1.06	2.59	–5.13	–2.51	–0.77	0.67	2.29
Job Destruction Rate (%)	–1.39	–1.39	3.30	–6.93	–3.60	–1.16	0.61	3.08
Net Job Creation Rate (%)	0.33	0.33	4.99	–7.85	–2.73	0.84	3.45	7.39

Notes: The table presents the distribution of two-period TWFE estimates $\hat{\beta}^{FE}_{s,t}$ weighted by $\hat{\omega}^{FE}_{s,t}$ defined in Corollary 2, along with the standard TWFE estimate $\hat{\beta}^{FE}$. There are $\frac{1}{2} \times 28 \times 27 = 406$ combinations of the first period t and the second period $s > t$.

Table 2: Generalized TWFE, Restricting Gap Duration

	Restrictions on Gap (Years)						
	None	1–4	5–8	9–12	13–16	17–20	21–28
Log Employment, Age 16–19	–0.133	–0.019	–0.027	–0.098	–0.221	–0.283	–0.239
	(0.062)	(0.061)	(0.072)	(0.057)	(0.085)	(0.133)	(0.111)
Log Employment, Age 20–34	0.014	0.005	–0.001	0.015	–0.002	0.014	0.057
	(0.014)	(0.017)	(0.018)	(0.017)	(0.020)	(0.025)	(0.035)
Log Employment, Age 35–54	–0.003	–0.012	–0.017	–0.009	–0.008	0.019	0.027
	(0.013)	(0.011)	(0.013)	(0.014)	(0.017)	(0.021)	(0.020)
Job Creation Rate (%)	–1.06	–1.86	–1.42	–1.17	–2.07	–0.78	1.05
	(0.59)	(0.51)	(0.63)	(0.71)	(0.92)	(0.96)	(1.14)
Job Destruction Rate (%)	–1.39	0.53	0.26	–2.24	–2.45	–0.69	–3.90
	(0.64)	(0.55)	(0.54)	(0.60)	(0.98)	(1.32)	(1.13)
Net Job Creation Rate (%)	0.33	–2.39	–1.68	1.07	0.37	–0.10	4.95
	(0.90)	(0.84)	(0.86)	(0.99)	(1.47)	(1.57)	(1.55)
No. of (s, t) pairs	406	106	90	74	58	42	36

Notes: The table presents the estimates of the minimum wage effect based on the generalized estimator defined in (4). By construction, the first column corresponds to the standard TWFE estimate $\hat{\beta}^{FE}$. Standard errors are in parentheses and they are robust to heteroskedasticity and correlation across observations on the same state.

Table 3: Generalized TWFE, Restricting Gap Duration and Controlling for Pre-trends

	Restrictions on Gap (Years)						
	None	1–4	5–8	9–12	13–16	17–20	21–28
Log Employment, Age 16–19	–0.122	–0.015	–0.021	–0.103	–0.194	–0.252	–0.233
	(0.066)	(0.056)	(0.076)	(0.058)	(0.091)	(0.146)	(0.117)
Log Employment, Age 20–34	0.018	0.011	0.002	0.018	0.004	0.019	0.060
	(0.015)	(0.014)	(0.018)	(0.018)	(0.021)	(0.028)	(0.034)
Log Employment, Age 35–54	–0.002	–0.008	–0.015	–0.007	–0.003	0.004	0.026
	(0.013)	(0.011)	(0.014)	(0.013)	(0.017)	(0.021)	(0.021)
Job Creation Rate (%)	–1.29	–2.05	–1.62	–1.22	–2.40	–1.29	0.77
	(0.61)	(0.52)	(0.54)	(0.76)	(0.93)	(0.98)	(1.07)
Job Destruction Rate (%)	–1.44	0.28	0.17	–2.48	–2.60	–0.46	–3.60
	(0.67)	(0.59)	(0.61)	(0.63)	(1.00)	(1.34)	(1.11)
Net Job Creation Rate (%)	0.32	–2.09	–1.66	1.77	0.23	–0.89	4.34
	(0.95)	(0.93)	(0.83)	(1.10)	(1.48)	(1.69)	(1.46)
No. of (s, t) pairs	406	106	90	74	58	42	36

Notes: The table presents the estimates of the minimum wage effect based on the generalized estimator defined in (5). For each constituent two-period estimator $\hat{\beta}^{*}_{s,t}$ with the first period t and the second period $s > t$, pre-trend of the outcome during years $t - 12$ through $t - 3$ is controlled for. Standard errors are in parentheses and they are robust to heteroskedasticity and correlation across observations on the same state.

A Proofs

A.1 Proof of Theorem 1

The following lemma is a key component of the proof.

Lemma 1. *Let $(x_t)_{t=1}^T$ and $(y_t)_{t=1}^T$ be any two sequences of real numbers with means $\overline{x} = \frac{1}{T}\sum_{t=1}^T x_t$ and $\overline{y} = \frac{1}{T}\sum_{t=1}^T y_t$. Then,*

$$\sum_{t=1}^{T}(x_t-\overline{x})(y_t-\overline{y}) = \frac{1}{T}\sum_{s>t}(x_s-x_t)(y_s-y_t).$$

While this lemma closely relates to a well-known property of U-statistics (see Lee (1990), for example), I present the proof below for completeness.

Proof. The proof follows from a simple algebraic identity:

$$\begin{aligned}
\sum_{s>t}(x_s-x_t)(y_s-y_t) &= \frac{1}{2}\sum_{s=1}^{T}\sum_{t=1}^{T}(x_s-x_t)(y_s-y_t) \\
&= \frac{1}{2}\sum_{s=1}^{T}\sum_{t=1}^{T}(x_s-\overline{x}+\overline{x}-x_t)(y_s-\overline{y}+\overline{y}-y_t) \\
&= \frac{1}{2}\sum_{s=1}^{T}\sum_{t=1}^{T}(x_s-\overline{x})(y_s-\overline{y}) + \frac{1}{2}\sum_{s=1}^{T}\sum_{t=1}^{T}(x_t-\overline{x})(y_t-\overline{y}) \\
&+ \frac{1}{2}\sum_{s=1}^{T}\sum_{t=1}^{T}(x_s-\overline{x})(y_t-\overline{y}) + \sum_{s=1}^{T}\sum_{t=1}^{T}(x_t-\overline{x})(y_s-\overline{y}) \\
&= T\sum_{t=1}^{T}(x_t-\overline{x})(y_t-\overline{y}).
\end{aligned}$$

□

Now I present the main proof of Theorem 1.

Proof. Let $\tilde{v}_{it} = v_{it} - \frac{1}{N}\sum_{i=1}^N v_{it}$ and $\overline{\tilde{v}}_i = \frac{1}{T}\sum_{t=1}^T \tilde{v}_{it}$ for any variable v. The FWL theorem implies that a TWFE estimator $\hat{\beta}_{FE}$ is given by

$$\hat{\beta}_{FE} = \frac{\sum_{i=1}^N\sum_{t=1}^T\left(\tilde{y}_{it}-\overline{\tilde{y}}_i\right)\left(\tilde{x}_{it}-\overline{\tilde{x}}_i\right)}{\sum_{i=1}^N\sum_{t=1}^T\left(\tilde{x}_{it}-\overline{\tilde{x}}_i\right)^2}.$$

Applying Lemma 1 to $(\tilde{x}_{it}, \tilde{y}_{it})_{t=1}^{T}$ for each $i = 1, \dots, N$ yields

$$\hat{\beta}_{FE} = \frac{\sum_{i=1}^{N} \sum_{s>t} (\tilde{y}_{is} - \tilde{y}_{it}) (\tilde{x}_{is} - \tilde{x}_{it})}{\sum_{i=1}^{N} \sum_{s>t} (\tilde{x}_{is} - \tilde{x}_{it})^2}.$$

This completes the proof. $\square$

Extending Theorem 1 to a multivariate regression or an IV regression is straightforward. In fact, with a vector regressor x_{it}, a TWFE estimator is given by

$$\hat{\beta}_{FE} = \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{x}_{it} - \overline{\tilde{x}}_i \right) \left(\tilde{x}_{it} - \overline{\tilde{x}}_i \right)' \right\}^{-1} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \overline{\tilde{y}}_i \right) \left(\tilde{x}_{it} - \overline{\tilde{x}}_i \right) \right\},$$

and applying Lemma 1 to each element of vectors and matrices yields

$$\hat{\beta}_{FE} = \left\{ \sum_{i=1}^{N} \sum_{s>t} (\tilde{x}_{is} - \tilde{x}_{it}) (\tilde{x}_{is} - \tilde{x}_{it})' \right\}^{-1} \left\{ \sum_{i=1}^{N} \sum_{s>t} (\tilde{y}_{is} - \tilde{y}_{it}) (\tilde{x}_{is} - \tilde{x}_{it}) \right\}.$$

With a scalar endogenous regressor x_{it} and a scalar instrument z_{it}, a TWFE estimator is given by

$$\begin{aligned} \hat{\beta}_{FE} &= \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \overline{\tilde{y}}_i \right) \left(\tilde{z}_{it} - \overline{\tilde{z}}_i \right)}{\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{x}_{it} - \overline{\tilde{x}}_i \right) \left(\tilde{z}_{it} - \overline{\tilde{z}}_i \right)} \\ &= \frac{\sum_{i=1}^{N} \sum_{s>t} (\tilde{y}_{is} - \tilde{y}_{it}) (\tilde{z}_{is} - \tilde{z}_{it})}{\sum_{i=1}^{N} \sum_{s>t} (\tilde{x}_{is} - \tilde{x}_{it}) (\tilde{z}_{is} - \tilde{z}_{it})}. \end{aligned}$$

A.2 Proof of Theorem 2

Proof. Each element of the numerator in the equation (1) is given by

$$\begin{aligned} E\left[\Delta_k y_{it} \Delta_k r_{it}\right] &= E\left[\left(y_{i,t+k}(x_{i,t+k}) - y_{i,t+k}(x_{it})\right) \Delta_k r_{it}\right] + E\left[\left(y_{i,t+k}(x_{it}) - y_{it}(x_{it})\right) \Delta_k r_{it}\right] \\ &= E\left[\tau_{ikt} \Delta_k x_{it} \Delta_k r_{it}\right] + E\left[\left(y_{i,t+k}(x_{it}) - y_{it}(x_{it})\right) \left(\Delta_k r_{it} - E\left[\Delta_k r_{it} | \Delta_k w_{it}\right]\right)\right] \\ &\quad + E\left[\left(y_{i,t+k}(x_{it}) - y_{it}(x_{it})\right) E\left[\Delta_k r_{it} | \Delta_k w_{it}\right]\right] \\ &= E\left[\tau_{ikt} \Delta_k x_{it} \Delta_k r_{it}\right] + E\left[\left(y_{i,t+k}(x_{it}) - y_{it}(x_{it})\right) \Delta_k \tilde{w}_{it}' \left(\delta_{kt} - \delta^{FE}\right)\right]. \end{aligned} \quad (6)$$

Note that the second term is eliminated at the last equality using Assumption 2, because

$$E\left[\left(y_{i,t+k}(x_{it}) - y_{it}(x_{it})\right) \left(\Delta_k r_{it} - E\left[\Delta_k r_{it} | \Delta_k w_{it}\right]\right)\right] = E\left[Cov\left(y_{i,t+k}(x_{it}) - y_{it}(x_{it}), \Delta_k x_{it} | \Delta_k w_{it}\right)\right].$$

This completes the derivation of the equation (2). In addition, letting $\delta_{kt} = \delta^{FE}$ or $Cov(y_{i,t+k}(x_{it}) - y_{it}(x_{it}), \Delta_k w_{it}) = 0$ in the above gives the equation (3).

Finally, suppose $w_{it} = (\mathbb{1}_{t=1} w_i' \ \cdots \ \mathbb{1}_{t=T-1} w_i')'$. Assumption 4 implies $E\left[\tilde{x}_{it}|w_i\right] = \tilde{w}_i' \delta_t^*$ with some vector δ_t^*, and $r_{it} = \tilde{x}_{it} - \tilde{w}_i' \delta_t^*$. Plugging this into (6) yields

$$\begin{aligned} E\left[\Delta_k y_{it} \Delta_k r_{it}\right] &= E\left[\tau_{ikt} x_{it} \Delta_k r_{it}\right] + E\left[\left(y_{i,t+k}(x_{it}) - y_{it}(x_{it})\right)\left(\Delta_k r_{it} - E\left[\Delta_k r_{it}|w_i\right]\right)\right] \\ &\quad + E\left[\left(y_{i,t+k}(x_{it}) - y_{it}(x_{it})\right) E\left[\Delta_k r_{it}|w_i\right]\right]. \end{aligned}$$

Again, the second term is eliminated by Assumption 2. The last term cancels because

$$E\left[\Delta_k r_{it}|w_i\right] = E\left[\Delta_k \tilde{x}_{it}|w_i\right] - \tilde{w}_i'(\delta_{t+k}^* - \delta_t) = 0.$$

This completes the proof. $\square$