Charles E. Gibbons[1] / Juan Carlos Suárez Serrato[2] / Michael B. Urbancic[3]

Broken or Fixed Effects?

[1] The Brattle Group, 201 Mission Street Suite 2800, San Francisco, CA 94105, USA, E-mail: charlie.gibbons@brattle.com
[2] Department of Economics, Duke University and NBER, Durham, NC, USA, E-mail: jc@jcsuarez.com
[3] Department of Economics, University of Oregon, Eugene, OR, USA

Abstract:
We replicate eight influential papers to provide empirical evidence that, in the presence of heterogeneous treatment effects, OLS with fixed effects (FE) is generally not a consistent estimator of the average treatment effect (ATE). We propose two alternative estimators that recover the ATE in the presence of group-specific heterogeneity. We document that heterogeneous treatment effects are common and the ATE is often statistically and economically different from the FE estimate. In all but one of our replications, there is statistically significant treatment effect heterogeneity and, in six, the ATEs are either economically or statistically different from the FE estimates.

Keywords: average treatment effects, fixed effects models, heterogeneous treatment effects

JEL classification: C21, C18, C52

DOI: 10.1515/jem-2017-0002

Fixed effects are a common means to "control for" unobservable differences among observations based upon observable characteristics; examples include age, year, or location in cross-sectional studies or individual or firm effects in panel data. While fixed effects permit different mean outcomes among groups, the estimates of treatment effects are typically required to be the same; in more colloquial terms, the intercepts of the conditional expectation functions may differ, but not the slopes.

Our main contribution is considering the empirical importance of heterogeneity in these slopes (*i.e.*, treatment effects) across fixed effects groups. In particular, we compare treatment effect estimates using a fixed effects estimator (FE) to the average treatment effect (ATE) in replications of eight influential papers from the *American Economic Review* published between 2004 and 2009.[1] We first consider a randomized experiment as a case study in Section 1 and, in Section 3, we show generally that heterogeneous treatment effects are common and that the FE and ATE are often statistically and economically different. In all but one paper, there is at least one statistically significant source of treatment effect heterogeneity. In five papers, this heterogeneity induces the ATE to be statistically different from the FE estimate at the 5% level (7 of 8 are statistically different at the 10% level). Five of these differences are economically significant, which we define as an absolute difference exceeding 10%. Based upon this result, we conclude that methods that consistently estimate the ATE offer more interpretable results than standard FE models.

In Section 2, we provide a formal framework to establish the theoretical bias of the FE estimator in the presence of heterogenous treatment effects. We derive the probability limit of the FE under heterogeneous treatment effects and provide an interpretation as a weighted average of group-specific effects. We propose two alternative estimators that are able to consistently estimate the ATE under group-specific heterogeneity and derive the joint asymptotic distributions of these estimators with the FE.

One approach to incorporate heterogeneous marginal effects into a regression framework is the correlated random coefficients model (CRC). Our paper explores the empirical relevance of CRC models by considering a simplified version: a fixed effects regression that includes group-specific marginal effects. This assumption corresponds to the following data-generating process:

$$y_i = x_i \beta_{g(i)} + \mathbf{z}_i' \gamma + \epsilon_i, \tag{1}$$

where y_i is the outcome for observation i among N, x_i is treatment or another variable of interest, and $\mathbf{z}_i$ contains control variables, including group-specific fixed effects. The treatment effects are group-specific for each of the $g = 1, \ldots, G$ groups, where group membership is known for each observation. Lastly, ϵ_i is mean 0 with variance-covariance matrix Ω. Our analysis of this model can be viewed as a special case of the results in Chernozhukov et al. (2013).

There is a long tradition in the econometrics literature considering average partial effects (see, *e.g.*, Blundell and Powell 2003; Chamberlain 1980, 1982; 1984; 1992; Chernozhukov et al. 2013; Graham and Powell 2012; Wooldridge 1997, 2005).[2]

Charles E. Gibbons is the corresponding author.

Definition 1: Average treatment effect (ATE)

The average treatment effect (ATE) for Equation 1 is defined as

$$\beta^{\text{ATE}} \equiv \sum_g \pi_g \beta_g,$$

where π_g is the population frequency of group g.

An established result is that fixed effects regressions average the group-specific slopes proportional to both the sample frequency of the group and the conditional variance of treatment, an average that generally does not coincide with the average treatment effect.[3] Though this theoretical result is well established, there has been little guidance for the applied researcher regarding the empirical importance of the difference. We find that the difference can be large.

Comparison to the literature. Our approach is similar to the CRC model of Chamberlain (1982) (see also Chamberlain (1984, 1992)). The primary differences between our setting and that of the CRC is that (i) we focus on cross-sectional data, whereas the CRC is based on panel data; and (ii) we employ fixed, rather than random effects. Because of the general similarities, our approach is related to the large literature analyzing non-separable correlated heterogeneity in panel data contexts. Closest to our derivation, Wooldridge (2005) shows conditions under which the FE provides consistent estimates of the average partial effect. Our analysis builds upon this derivation for the case of fixed coefficients and offers a different interpretation of the necessary conditions for this result. Graham and Powell (2012) study the identification and estimation of average partial effects under "irregularity" conditions where the information bound may be singular and Arellano and Bonhomme (2012) study the identification and estimation of distributions of coefficients in CRC models.

Another important example is Chernozhukov et al. (2013), who study average and quantile treatment effects and derive results that nest our approach. In particular, while we focus on cross-sectional settings, our models are relevant for panel models with discrete regressors, as in Chernozhukov et al. (2013). Ghanem (2017) studies testable implications of the assumptions made in these non-separable panel data models. Finally, Imai and Kim (2016) study the linear fixed effect model from a matching perspective, reformulate our result from this perspective, and study dynamic extensions. While these papers provide a strong theoretical reason to believe that FE does not provide sample-weighted estimates, we illustrate the empirical importance of this distinction using a broad array of microeconometric questions.

In the presence of heterogeneous treatment effects, the FE gives a weighted average of these effects. The weights depend not only on the frequency of the groups, but also upon sample variances within the groups. Angrist and Krueger (1999) compare the results from regression and matching estimators to demonstrate that the effects of a dichotomous treatment are averaged using different weights under each procedure. Many empirical studies, including many of those that we replicate in this paper, run separate regressions by group out of concern for the presence of treatment effect heterogeneity. Less common are the more parsimonious interacted model or weighted regression approaches that we propose, but which assume that there is no heterogeneity in coefficients for other predictors. A related approach is the random growth model, which uses individual-specific time trends to control for differing growth rates (see, *e.g.*, Heckman and Hotz 1989; Papke 1994; Friedberg 1998). This heterogeneity is used to control for omitted variables, rather than to model the treatment effect of interest itself, however. Solon, Haider, and Wooldridge (2015) declare that the FE may be biased in the presence of heterogeneous treatment effects and note that weighted least squares can be used to recover the average partial effect. We build upon their discussion by deriving the necessary weights and providing applications to illustrate empirically the importance of the difference between weighted and FE estimates.

1 A Case Study: Karlan and Zinman (2008)

Even if an experiment ensures that treatment is independent of any other covariates, the FE might not be a consistent estimator of the ATE. Among our *AER* replications, there is one experiment that can be used to illustrate this point: Karlan and Zinman (2008). In this paper, the authors randomize the interest rate offered for a microloan across a population of South Africans and estimate the credit elasticity. One set of fixed effects that the authors use is the "pre-approved risk category" of the borrower (low, medium, or high). To offer interest rates commensurate with prevailing market rates, the authors charge higher rates to higher risk individuals. As we will show, however, that differing means in treatment do not drive the difference between the FE and ATE estimates, but rather differences in variances. To this point, the authors offer not only higher rates to riskier borrowers, but also offer a greater range of rates to this group and, as a result, the variance of treatment differs across the groups. Thus, the FE estimate will not be equal to the ATE if the responsiveness to interest rates varies across risk groups.

The FE weights are given in column 3 of Table 1. These are the relative variances of treatment by group multiplied by the sample frequency of that group (see Proposition 1). Using these weights and the group effect estimated using an interacted model (given in column 2 of Table 1), we calculate the FE estimate in the bottom row of the table in the "FE weight" column. Compare the weights from the FE model to the sample frequencies used to calculate the ATE. Note that high risk individuals are over-weighted in the FE model due to their relatively high variance in treatment and the low and medium risk individuals are under-weighted.

Table 1: Karlan and Zinman (2008) treatment effect weighting.

		Weight	
Risk group	**Effect**	**FE**	**Sample**
Low	−32.4	0.044	0.125
Medium	−9.9	0.058	0.092
High	−2.7	0.898	0.783
Average		−4.393	−7.047
Std. error		(1.129)	(1.917)

The ATE estimated is the IWE estimator. The FE estimate here, −4.40, does not precisely equal the FE estimate of −4.37 reported in the paper due to slight correlation between mailer wave fixed effects, excluded from this simplified exposition, and the interest rate. Subsequent replication results in our paper do recover the actual values reported in the replicated papers, including this one, unless otherwise noted.

We find that high-risk borrowers are much less responsive to the interest rate than low-risk borrowers. Because high-risk individuals are over-weighted and have a smaller (in absolute value) treatment effect, the FE estimate underestimates the sample-weighted responsiveness of individuals to the interest rate by over 60%.

2 Estimating the Average Treatment Effect

In this section, we first derivate the bias of the FE estimator under treatment effect heterogeneity. Based upon this result, we provide two alternative estimators that eliminate this bias. We also discuss testing procedures related to our proposed estimators.

2.1 Bias of the Fixed Effects Estimator

One way to parameterize the treatment effect heterogeneity in Equation 1 is by interacting the fixed effects with treatment; call this vector $\mathbf{a}_i$.[4] Then, the data-generating process can be rewritten as:

$$y_i = \mathbf{a}_i'\beta + \mathbf{z}_i'\gamma + \epsilon_i, \tag{2}$$

where β is now a vector of coefficients. Further define the $N \times 1$ column vector forms $\mathbf{Y}$, $\mathbf{X}$, and $\boldsymbol{\epsilon}$ as vectors across the N observations and $\mathbf{A}$ and $\mathbf{Z}$ as matrices across observations. Define $\mathbf{M} = \mathbf{I}_N - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$ as the annihilator matrix for $\mathbf{Z}$; $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{A}}$ are annihilated versions. Notably, $\tilde{x}_i$ is a value in the $\tilde{\mathbf{X}}$ vector.

As a baseline case, consider an OLS model with fixed effects that does not account for treatment effect heterogeneity, which we call the *fixed effects estimator*.

Definition 2: Fixed effects estimator (FE)
Define the standard fixed effect estimator (FE) as:

$$\hat{b}^{\text{FE}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}.$$

In general, the FE is a biased and inconsistent estimator of the ATE.

Proposition 1: Bias and inconsistency of FE
Under the usual assumptions for Equation 1 (see Online Appendix A), the expected value of the FE is:

$$\mathbb{E}\left[\hat{b}^{\mathrm{FE}}\middle|\mathbf{X},\mathbf{Z},\mathbf{A}\right]=\left[\sum_i \tilde{x}_i^2\right]^{-1}\sum_i \tilde{x}_i\tilde{\mathbf{a}}_i'\beta=\beta^{\mathrm{ATE}}+\sum_g \frac{N_g}{N}\beta_g\left[\frac{\widehat{\mathrm{Var}}\left(\tilde{x}_i \mid g(i)=g\right)}{\widehat{\mathrm{Var}}\left(\tilde{x}_i\right)}-1\right]+o_p(1),$$

where $\widehat{\mathrm{Var}}(\cdot)$ is the sample variance and N_g is the number of observations in group g. Further, the FE converges in probability to:

$$\hat{b}^{\mathrm{FE}} \xrightarrow[n\to\infty]{p} \beta^{\mathrm{ATE}}+\sum_g \pi_g\beta_g\left[\frac{\mathrm{Var}\left(\tilde{x}_i \mid g(i)=g\right)}{\mathrm{Var}\left(\tilde{x}_i\right)}-1\right].$$

Hence, if the variance of x_i conditional on $\mathbf{z}_i$ varies across groups and treatment effects also vary across groups, then the FE is a biased and inconsistent estimator for the ATE.

Proposition 1 reveals that, while the FE is an average of the group-specific effects, the weights generally do not coincide with sample frequencies. Instead, FE upweights groups with high variance in treatment conditional upon other covariates and downweights groups with low variance in treatment. This is an efficient approach if the treatment effect is the same for all groups, but leads to biased and inconsistent estimates of the ATE when the treatment effect varies across groups.

An example where FE would give unbiased results is a regression using data from a perfectly randomized experiment where treatment has the same variance across groups. Such perfection is likely unattainable in observational or experimental settings, however. Indeed, in Section 1, we replicated a randomized experiment from Karlan and Zinman (2008) as a case study. In that experiment, treatment is randomized within different fixed effects groups, but the variances of treatment are not the same across groups. There, we found that the ATE differs from the FE estimate by over 60%.

2.2 Alternative Estimators

We offer two alternative estimators for the ATE that, unlike the FE, are unbiased and consistent. For the first estimator, Equation 2 hints that an interacted model could be used to estimate the treatment effect for each group; the resulting group-specific estimates are averaged to provide the ATE. This is the *interaction-weighted estimator*.

Definition 3: Interaction-weighted estimator (IWE)

The interaction-weighted estimator is found by estimating β from Equation 2 using an interacted model, then using these estimates to calculate the ATE. Thus, the IWE is given by:

$$\hat{b}^{\mathrm{IWE}}=\hat{\mathbf{f}}\left(\tilde{\mathbf{A}}'\tilde{\mathbf{A}}\right)^{-1}\tilde{\mathbf{A}}'\tilde{\mathbf{Y}},$$

where [5]

$$\hat{\mathbf{f}}=\frac{1}{N}\begin{bmatrix}N & N_1 & \cdots & N_{G-1}\end{bmatrix}.$$

Proposition 1 shows that, while FE provides a weighted average of the treatment effects, these weights do not equal sample frequencies. The *regression-weighted estimator* re-weights each observation to undo the FE weighting and applies the frequency weighting of the ATE. A potential advantage of this approach is that it does not require estimating each group's treatment effect.

Definition 4: Regression-weighted estimator (RWE)

The regression-weighted estimator re-weights each observation according to

$$\hat{w}_i=\left[\widehat{\mathrm{Var}}\left(\tilde{x}_j \mid g(j)=g(i)\right)\right]^{-1/2}; \tag{3}$$

that is, inversely proportional to the standard deviation of the conditional treatment values within its group. Let $\hat{\mathbf{W}}$ be a diagonal matrix of these values squared. Then, the RWE is given by:

$$\hat{b}^{\mathrm{RWE}}=\left(\tilde{\mathbf{X}}'\hat{\mathbf{W}}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\hat{\mathbf{W}}\tilde{\mathbf{Y}}.$$

To calculate the RWE, first estimate the annihilator matrix **M**. Then, calculate the weights according to Equation 3. Then, perform weighted least squares using the annihilated data. Note that the RWE can be re-written as:

$$\begin{aligned}\hat{b}^{\text{RWE}} &= \left(\sum_i \frac{\tilde{x}_i^2}{\widehat{\text{Var}}\left(\tilde{x}_i \mid g(j) = g(i)\right)}\right)^{-1} \sum_i \frac{\tilde{x}_i \tilde{y}_i}{\widehat{\text{Var}}\left(\tilde{x}_j \mid g(j) = g(i)\right)} \\ &= \frac{1}{N} \sum_g N_g \frac{\widehat{\text{Cov}}\left(\tilde{x}_i, \tilde{y}_i \mid g(i) = g\right)}{\widehat{\text{Var}}\left(\tilde{x}_i \mid g(i) = g\right)}.\end{aligned}$$

The IWE and RWE can be compared to the FE. First, it should be noted that, unlike the FE, both the IWE and the RWE are unbiased estimators of the ATE (see Online Appendix A). Furthermore, they are consistent, which we illustrate by deriving the joint asymptotic distribution of the three estimators.[6] To do so, we first define $\hat{\Omega}$ to be the variance-covariance matrix of ϵ, which may be defined following standard heteroskedastic- or cluster-robust approaches.

Proposition 2: Asymptotic distribution of the estimators

Under standard assumptions for the data-generating process given by Equation 1 (see Online Appendix A and, e.g., Wooldridge (2001)), the asymptotic distribution of the estimators is

$$\sqrt{N}\begin{bmatrix} \hat{b}^{\text{FE}} - \beta^{\text{FE}} \\ \hat{b}^{\text{IWE}} - \beta^{\text{ATE}} \\ \hat{b}^{\text{RWE}} - \beta^{\text{ATE}} \end{bmatrix} \xrightarrow{d} N\left(\mathbf{0}, \begin{bmatrix} \Sigma_{FE} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{12}' & \Sigma_{IWE} & \Sigma_{23} \\ \Sigma_{13}' & \Sigma_{23}' & \Sigma_{RWE} \end{bmatrix}\right),$$

where

$$\begin{aligned}
\mathbf{V}_{\tilde{\mathbf{X}}} &= \mathbb{E}\left[\tilde{x}_i^2\right] & \mathbf{V}_{\tilde{\mathbf{A}}} &= \mathbb{E}\left[\tilde{\mathbf{a}}_i'\tilde{\mathbf{a}}_i\right] \\
\mathbf{V}_{\tilde{\mathbf{X}}}^{W} &= \mathbb{E}\left[w_i^2\tilde{x}_i^2\right] = 1 & \mathbf{f} &= \left[1 \ \pi_1 \ \dots \ \pi_{G-1}\right] \\
\Sigma_{FE} &= \mathbf{V}_{\tilde{\mathbf{X}}}^{-1}\left[\text{plim}\frac{\tilde{\mathbf{X}}'\hat{\Omega}\tilde{\mathbf{X}}}{N}\right]\mathbf{V}_{\tilde{\mathbf{X}}}^{-1} & \Sigma_{12} &= \mathbf{V}_{\tilde{\mathbf{X}}}^{-1}\left[\text{plim}\frac{\tilde{\mathbf{X}}'\hat{\Omega}\tilde{\mathbf{A}}}{N}\right]\mathbf{V}_{\tilde{\mathbf{A}}}^{-1}\mathbf{f}' \\
\Sigma_{IWE} &= \mathbf{f}\,\mathbf{V}_{\tilde{\mathbf{A}}}^{-1}\left[\text{plim}\frac{\tilde{\mathbf{A}}'\hat{\Omega}\tilde{\mathbf{A}}}{N}\right]\mathbf{V}_{\tilde{\mathbf{A}}}\mathbf{f}' & \Sigma_{13} &= \mathbf{V}_{\tilde{\mathbf{X}}}^{-1}\left[\text{plim}\frac{\tilde{\mathbf{X}}'\hat{\Omega}\mathbf{W}\tilde{\mathbf{X}}}{N}\right]\left[\mathbf{V}_{\tilde{\mathbf{X}}}^{W}\right]^{-1} \\
\Sigma_{RWE} &= \left[\mathbf{V}_{\tilde{\mathbf{X}}}^{W}\right]^{-1}\left[\text{plim}\frac{\tilde{\mathbf{X}}'\mathbf{W}\hat{\Omega}\mathbf{W}\tilde{\mathbf{X}}}{N}\right]\left[\mathbf{V}_{\tilde{\mathbf{X}}}^{W}\right]^{-1} & \Sigma_{23} &= \mathbf{f}\mathbf{V}_{\tilde{\mathbf{A}}}^{-1}\left[\text{plim}\frac{\tilde{\mathbf{A}}'\hat{\Omega}\mathbf{W}\tilde{\mathbf{X}}}{N}\right]\left[\mathbf{V}_{\tilde{\mathbf{X}}}^{W}\right]^{-1}.
\end{aligned}$$

Remarks.

1. Identification is achieved if the FE model is identified and $\text{Var}(\tilde{x}_i \mid g(i) = g) > 0 \ \forall \ g$, that is, if there is variation in treatment (either in level or assignment status) within each group.

2. The IWE estimates the treatment effect for each group, allowing the researcher to examine the various treatment effects, which themselves may be of interest. The RWE does not estimate the group-level effects, which is an advantage if the sample size is relatively small. The effective sample size is often small when clustered standard errors are employed and the RWE may be more successful in this situation. This is particularly true if the level of heterogeneity and the level of clustering are the same or colinear.[7]

3. In the presence of heterogeneous treatment effects, the IWE may reduce standard errors by modeling the effects directly. The IWE may also be more robust to model misspecification.

4. We only consider heterogeneity in β and assume constant γ coefficients across groups. Under this assumption, the IWE estimator is a more parsimonious version of a fully saturated model estimated separately for each group. The econometrician must decide whether this assumption is acceptable for his or her particular application.

5. When the IWE is estimated, a standard Wald test can be used to test for the presence of heterogeneous treatment effects. When the IWE and its associated interactions are not estimated, a score test based on the FE can be used instead (see the next subsection).

6. Given the asymptotic result in Proposition 2, it is straightforward to perform a test of equality between either estimate of the ATE and the FE estimate.
7. These results can be confirmed using a Monte Carlo simulation; see Online Appendix B.
8.

2.3 Testing for Heterogeneous Treatment Effects

Armed with two estimators of the ATE, we next consider testing. First, we derive tests for the presence of heterogeneous treatment effects using both Wald and score tests. Then, we offer a specification test for equality between the ATE and the FE. These tests are implemented by Stata commands and an R package available from the authors, as discussed in Online Appendix C.

2.3.1 Wald Test for Modeled Heterogeneity

If the IWE is estimated following Equation 2, then testing for the presence of heterogeneous treatment effects is straightforward. Standard or robust methods can be used to test for the joint significance of the interaction terms.

Proposition 3: Wald test for modeled heterogeneity
The Wald test statistic for heterogeneous treatment effects is calculated according to

$$T_W = \mathbf{p}\mathbf{V}^{\text{INT}}\mathbf{p}',$$

where

$$\mathbf{V}^{\text{INT}} = \left(\tilde{\mathbf{A}}'\tilde{\mathbf{A}}\right)^{-1}\tilde{\mathbf{A}}'\hat{\Omega}\tilde{\mathbf{A}}\left(\tilde{\mathbf{A}}'\tilde{\mathbf{A}}\right)^{-1}$$

and the $(G-1)\times G$ matrix

$$\mathbf{p} = \begin{bmatrix}\mathbf{0} & \mathbf{1}_{G-1}\end{bmatrix}.$$

Asymptotically, this test statistic has a χ^2_{G-1} distribution under the null hypothesis.

2.3.2 Score Test for Unmodeled Heterogeneity

If the RWE is estimated, the researcher may not be interested in or able to estimate the treatment effects by group. Nonetheless, the presence of heterogeneous treatment of the form modeled by the IWE can be tested.

This procedure begins by obtaining the residual from the FE model for each observation e_i.[8] The score is calculated according to

$$\mathbf{s}\left(y_i; \mathbf{a}_i, \mathbf{z}_i, \hat{b}^{\text{FE}}\right) = e_i\begin{bmatrix}\mathbf{z}_i\\ \mathbf{a}_i\end{bmatrix}.$$

Proposition 4: Score test for unmodeled heterogeneity
A score test statistic for the presence of heterogeneous treatment effects has the form[9]

$$T_S = N\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{s}\left(y_i; \mathbf{a}_i, \mathbf{z}_i, \hat{b}^{\text{FE}}\right)\right)'\mathbf{S}_0^{-1}\mathbf{C}'\left(\mathbf{C}\mathbf{S}_0^{-1}\mathbf{C}'\right)^{-1}\mathbf{C}\mathbf{S}_0^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{s}\left(y_i; \mathbf{a}_i, \mathbf{z}_i, \hat{b}^{\text{FE}}\right)\right),$$

where

$$\mathbf{S}_0 = \frac{1}{N}\sum_{i=1}^{N} \mathbf{s}\left(y_i; \mathbf{a}_i, \mathbf{z}_i, \hat{b}^{\mathrm{FE}}\right) \mathbf{s}\left(y_i; \mathbf{a}_i, \mathbf{z}_i, \hat{b}^{\mathrm{FE}}\right)'$$

and

$$\mathbf{C} = \left[\mathbf{0}_{(G-1)\times(K+1)} \quad \mathbf{I}_{G-1}\right]$$

(see, e.g., Wooldridge 2001). If clustering is desired, with C clusters and N_c observations in cluster c, then instead we have

$$\mathbf{S}_0 = \frac{1}{C}\sum_{c=1}^{C}\sum_{j=1}^{N_c}\sum_{i=1}^{N_c} \mathbf{s}\left(y_i; \mathbf{a}_i, \mathbf{z}_i, \hat{b}^{\mathrm{FE}}\right) \mathbf{s}\left(y_i; \mathbf{a}_i, \mathbf{z}_i, \hat{b}^{\mathrm{FE}}\right)'.$$

Like the Wald test above, this test statistic has an asymptotic χ^2_{G-1} distribution under the null hypothesis.[10]

2.3.3 Test for Equality Between the ATE and FE Estimates

Even if heterogeneous treatment effects are present, the ATE and FE may be equal or at least statistically indistinguishable. In this subsection, we derive a test that is able to distinguish between the two estimates. The same approach can be applied for either estimator of the ATE (*i.e.*, RWE or IWE) and we refer to the chosen estimator as $\hat{b}^{\mathrm{ATE}}$.

Proposition 5: Specification test of the differences between the FE and ATE estimates
The test of the following null hypothesis

$$H_0 : \beta^{\mathrm{ATE}} - \beta^{\mathrm{FE}} = 0$$
$$H_a : \beta^{\mathrm{ATE}} - \beta^{\mathrm{FE}} \neq 0$$

can be conducted using a Hausman-style test. Note that the Wald test statistic

$$T_E = \frac{\left(\hat{b}^{\mathrm{ATE}} - \hat{b}^{\mathrm{FE}}\right)^2}{\mathrm{Var}\left[\hat{b}^{\mathrm{ATE}} - \hat{b}^{\mathrm{FE}}\right]}$$

has an asymptotic $\chi^2(1)$ distribution under H_0. The variance term is easily computed using the joint asymptotic distribution given in Proposition 2.

3 Comparing FE and ATE Estimates: An *AER* Investigation

To consider the empirical relevance of the distinction between the FE and ATE estimators, we turn to highly-cited papers published in the *American Economic Review* between 2004 and 2009. The papers that we choose are well known in their respective fields and rightfully serve as prime examples of respected empirical work. We find the eight most-cited papers that use fixed effects in an OLS model as part of their primary specification and meet additional requirements that serve to limit our scope to papers in applied microeconomics with a clear effect of interest. These papers are listed in Table 2 along with the outcomes, effects of interest, fixed effects considered, and models replicated as identified by the table and column number of appearance in the original paper. A complete description of the process that we follow to identify these papers can be found in Online Appendix D.1.

Table 2: Papers from the *AER* used in the meta-analysis.

Citation	Outcome	Effect of interest	Fixed effects	Table	Column

Banerjee and Iyer (2005)	Fertilizer use	Proportion non-landlord land	Coastal dummy, year	3	1
	Proportion irrigated Proportion other cereals Proportion rice Proportion wheat Proportion white rice Rice yield (log) Wheat yield (log)				
Bedard and Deschênes (2006)	Smoking dummy	Veteran status	Age, education, race, region	5	1
Card, Dobkin, and Maestas (2008)	Saw doctor dummy	Age over 65 dummy	Ethnicity, gender, region, year, education level	3	6, 8
	Was hospitalized dummy				
Karlan and Zinman (2008)	Loan size	Interest rate (log)	Mailer wave, risk category	4	1
Lochner and Moretti (2004)	Imprisonment	Education	Race, age, year	3	1
Meghir and Marten (2005)	Wage (log; change in)	Education reform	High ability dummy, high father's education dummy, sex, year	2	1 (row 1)
Oreopoulos (2006)	Wage (log)	Education	Age, Northern Ireland dummy	2	3
Pérez-González (2006)	Market-book ratio Operating returns	CEO heir inheritance	High family ownership dummy, year	9	1, 6

Additional details on our replications are found in Online Appendix D.

To consider whether the difference between the FE and ATE estimators is empirically important, we test for heterogeneous treatment effects and for a difference between the FE and ATE estimates.[11] Our results are summarized in Table 3. For each paper, we list the groups that we consider as potential dimensions of treatment effect heterogeneity along with a test for the presence of heterogeneity, a specification test comparing the ATE and FE estimates, and the percent difference in the two estimates. In the final column, we indicate whether the author considers treatment effect heterogeneity among the groups. These statistics all use the RWE and we compute standard errors following the level of clustering used by the original author.[12] The results for the IWE are generally very similar, as we would expect, and these results are included in the detailed tables of Online Appendix D.3.

Table 3: *AER* replication results.

Citation	Fixed effect	Joint test	Diff. test	Percent	In paper
		(p-value)	(p-value)	diff.	
(1)	(2)	(3)	(4)	(5)	(6)
Banerjee and Iyer (2005) (Proportion irrigated)	Coastal	0.065*	0.013**	−31.7†	
	Year	0.000***	0.896	0.0	
Bedard and Deschênes (2006)	Age	0.942	0.830	−0.2	
	Education	0.002***	0.875	−0.1	
	Race	0.080 *	0.084 *	0.5	
	Region	0.697	0.392	0.1	
Card, Dobkin, and Maestas (2008)	Ethnicity (*outcome: saw doctor*)	0.000***	0.211	−0.5	X
	Gender	0.000***	0.582	−0.4	
	Region	0.028 **	0.258	0.3	
	Year	0.229	0.603	0.8	
	Education (Whites only)	0.028**	0.278	−2.0	X
	Education (non-Whites only)	0.967	0.798	−0.4	X
	Ethnicity (*outcome: hospitalized*)	0.001***	0.614	−0.1	X
	Gender	0.000***	0.068*	−0.5	

	Region	0.004***	0.301	0.2	
	Year	0.383	0.436	−1.3	
	Education (Whites only)	0.096*	0.431	1.0	X
	Education (non-Whites only)	0.743	0.296	3.3	X
Karlan and Zinman (2008)	Mailer wave	0.234	0.782	0.2	
	Risk category	0.005***	0.003***	69.7†	
Lochner and Moretti (2004)	Race (all)	0.000***	0.000***	−1.7	X
	Age (Blacks only)	0.000***	0.000***	32.6†	
	Year (Blacks only)	0.000***	0.000***	1.6	
	Age (Whites only)	0.000***	0.000***	29.0†	
	Year (Whites only)	0.005***	0.095*	−0.2	
Meghir and Marten (2005)	High father's education	0.000***	0.000***	15.5†	X
	Gender	0.344	0.514	0.3	X
	Year	0.000***	0.337	0.1	
Oreopoulos (2006)	N.Ireland	0.000***	0.001***	0.8	X
	Age (Great Britain)	0.242	0.006***	1.8	
	Age (N. Ireland)	0.590	0.275	0.8	
	Age (N. Ireland & Great Britain)	0.005***	0.053*	1.2	
Pérez-González (2006)	Year (*outcome: MB*)	0.143	0.327	−11.3†	
	High family ownership	0.135	0.510	9.2	
	Year (*outcome: OR*)	0.111	0.491	−7.5	
	High family ownership	0.423	0.503	9.4	

All results are using the RWE estimator. Column 3 gives the *p*-value for the test of the joint significance of the interaction terms using a score test. Column 4 gives the *p*-value for a *t* test of the difference between the ATE and FE estimates. Column 5 gives the percent difference between these two estimates. The last column indicates whether the author considers heterogeneity among these groups. A single star indicates significance at the 10 percent level, two stars indicate significance at the 5 percent level, and three stars indicate significance at the 1 percent level. A dagger indicates a difference of more than 10 percent between the two estimates.

Column (3) shows that all but one paper has at least one set of fixed effects groups that exhibit treatment effect heterogeneity. This heterogeneity translates into significant differences between the ATE and FE estimates for five papers at the 5% level and seven papers at the 10% level, as seen in Column (4). Defining a difference to be "economically significant" if it exceeds 10%, Column (5) shows that five papers have economically significant differences between the ATE and FE estimates. The average of the largest deviation for each paper that we consider is 21%. As a comparison, Graham and Powell (2012) find a 25% difference between their CRC and FE estimates.

The weighting scheme employed by FE yields a more efficient estimator in the absence of heterogeneous treatment effects. This suggests that FE may be more efficient if heterogeneity is relatively unimportant. As we have shown, however, the FE is generally an inconsistent estimator of the ATE. This presents a bias-variance trade-off. Figure 1 shows the relationship between the largest absolute difference between the FE and RWE estimates for each paper and compares that to the percent difference in the standard errors of the two estimators.[13] The ATE estimator exhibits standard errors that are less than ten percent larger than those for the FE in six of eight cases.[14] Overall, the results indicate that there is not generally a strong bias-variance trade-off unless the differences between the estimates are great. But, if the difference between the estimates is great (*i.e.*, the bias is high), then the ATE should be preferred for policy and interpretablity reasons.

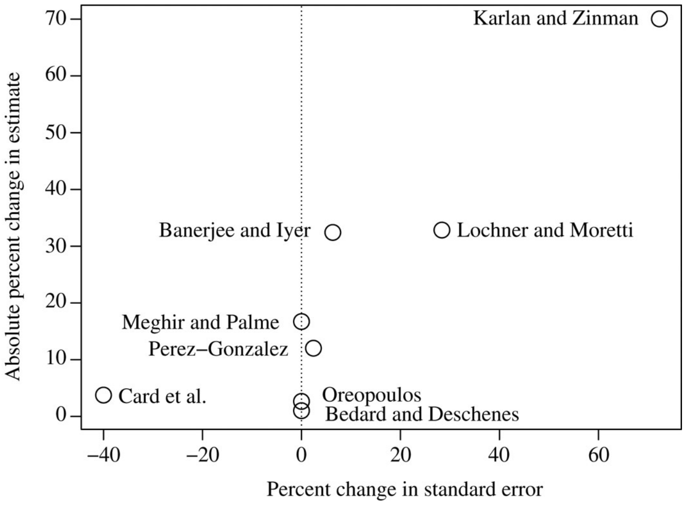

Figure 1: The relationship between the difference in the estimates and the change in variance among the *AER* replications
Notes: Figure is based on the full results presented in Online Appendix D.3. Figure plots estimates from the RWE and corresponding standard errors at the level of clustering used by the original authors, where applicable.

4 Conclusion

We show that, in the presence of heterogeneous treatment effects, OLS with group fixed effects generally offers a biased estimator of the average treatment effect, a result that has relevance for a variety of fields, including labor, development, health, public finance, and corporate finance. Based on this evidence, we suggest that researchers explore the impact that heterogeneous treatment effects may have on their estimates by considering interaction-weighted or regression-weighted estimators or by analyzing the group-specific weights implied by OLS with fixed effects. We believe that reporting average treatment effects will make estimates more interpretable for individual papers and, perhaps more importantly, across academic studies without increasing the variance of the estimates.

The methods employed in this paper, however, are subject to three notable limitations. First, when clustered standard errors are used, small-sample issues may arise when the number of groups grows close to the number of clusters. When this situation arises, researchers must choose between estimating conservative standard errors and providing a treatment effect that is representative of the whole sample. The optimal solution is inherently application specific.

Second, our discussion has been limited to the case of OLS and we have ignored issues of endogeneity. In cases where the treatment of interest can be assumed to be “as-good-as-random,” as in the cases of a randomized or natural experiment, regression discontinuity, or difference-in-differences identification strategies, our methods may be applied directly. When instrumental variables are used, however, our methods will be complicated by the weights inherent in local average treatment effect estimation (Abadie 2002; Kling 2001); in particular, see Wooldridge (1997) for an analysis of CRC models in the context of instrumental variables estimation.

Finally, our focus in this paper has been to analyze heterogeneity in treatment effects across observable groups. Heterogeneity may also arise along unobservable margins (see, *e.g.*, Bitler, Gelbach, and Hoynes 2014).

Acknowledgement

We are grateful for comments from Michael Anderson, Alan Auerbach, Rodney Andrews, Joshua Angrist, Marianne Bitler, Henning Bohn, Moshe Buchinsky, Federico Bugni, Colin Cameron, Carlos Dobkin, Shakeeb Khan, Maximilian Kasy, Patrick Kline, Yolanda Kodrzycki, Trevon Logan, Fernando Lozano, Matt Masten, Arnaud Maurel, Doug Miller, Juan Carlos Montoy, Enrico Moretti, Ron Oaxaca, Steve Raphael, Adam Rosen, Jesse Shapiro, Jasjeet Sekhon, Todd Sorensen, Doug Steigerwald, Rocio Titiunik, and Philippe Wingender and for the comments and suggestions of seminar participants at UC Berkeley, the 2008 AEA Pipeline Conference at UCSB, and the 2009 All UC Labor Conference. Any remaining errors are the fault of the authors.

Notes

[1] See Murphy and Topel (1985), Gentzkow and Shapiro (2013), and Oster (2014) for other examples of papers that replicate published studies to elucidate a methodological point. We only analyze the data that the authors openly provide on the EconLit website. Though some of these papers include both OLS and instrumental variables approaches, we consider the implications of heterogeneous treatment effects for the OLS specifications only to focus on the weighting scheme applied by this common procedure.

[2] We assume that the sample is representative of the population of interest for the ATE; specifically, $N_g/N \to \pi_g$.

[3] See, *e.g.*, Angrist and Krueger (1999), Wooldridge (2005), and Angrist and Pischke (2009).

[4] Consider $\mathbf{a}_i$ having first x_i, followed by x_i interacted with $G - 1$ fixed effects.

[5] These weights are designed to align with the definition of $\mathbf{a}_i$; see footnote 4.

[6] The fixed effects that we consider denote group membership and the sizes of these groups grow with overall sample size – *i.e.*, $N_g \to \infty$; $\forall\ g \in 1, \ldots, G$, G fixed. This is somewhat opposite of the typical configuration in panel data problems.

[7] The RWE estimator is identified in this situation because the model form is the same as the FE model, which is identified and the clustered variance-covariance matrix is well-defined, but observations are differentially weighted based on covariates, rather than features of the error structure.

[8] $\mathbf{e} = \mathbf{MY} - \mathbf{MX}\hat{b}^{\mathrm{FE}}$.

[9] This form assumes that the information matrix equality holds, which is true under standard regularity conditions and correct specification under the null (see Cameron and Trivedi 2005).

[10] This test may outperform the Wald test when a clustered variance-covariance matrix is used (Kline and Santos 2012).

[11] We develop a Stata command and R package to perform these analyses. See Online Appendix C. We have posted these resources online for researchers interested in implementing these tests.

[12] In Online Appendix D.3, we provide both the clustered and non-clustered heteroskedasticity-robust results. If the fixed effects groups are colinear with the clustering term, we are not able to cluster the IWE estimator. This is the case for the coastal interaction in Banerjee and Iyer (2005) and in the models of Oreopoulos (2006). Because the RWE estimator does not require estimating the interactions, clustering is possible in these cases. We choose to present the RWE results in Table 3 for this reason.

[13] If the difference in the standard errors is positive, the RWE has a larger standard error.

[14] It is perhaps not surprising that the standard errors for Karlan and Zinman (2008) increase substantially given the large change in the estimate (over 60% for the RWE). But the t-statistics are similar: −4.00 using FE and −3.94 using the RWE.

References

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models." *Journal of the American Statistical Association* 97 (457): 284–292.

Angrist, Joshua D. and Alan B. Krueger. 1999. Empirical Strategies in Labor Economics. In *Handbook of Labor Economics*, ed. Orley Ashenfelter and David Card. Vol. 3. Amsterdam: Elsevier.

Angrist, Joshua and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Arellano, Manuel and Stéphane Bonhomme. 2012. "Identifying Distributional Characteristics in Random Coefficients Panel Data Models." *The Review of Economic Studies* 79 (3): 987–1020.

Banerjee, Abhijit and Lakshmi Iyer. 2005. "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *American Economic Review* 95 (4): 1190–1213.

Bedard, Kelly and Olivier Deschênes. 2006. "The Long-Term Impact of Military Service on Health: Evidence from World War II and Korean War Veterans." *American Economic Review* 96 (1): 176–194.

Bitler, Marianne P., Jonah B. Gelbach and Hilary W. Hoynes. 2014. Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment. Working Paper 20142 National Bureau of Economic Research.

Blundell, R. W. and James L. Powell. 2003. Endogeneity in Nonparametric and Semiparametric Regression Models. In *Advances in Economics and Econometrics: Theory and Applications*, ed. M. Dewatripont, L. P. Hansen and S. J. Turnovsky. Vol. II. Cambrige: Cambridge University Press.

Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics*. Cambridge: Cambridge University Press.

Card, David, Carlos Dobkin and Nicole Maestas. 2008. "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare." *American Economic Review* 98 (5): 2242–2258.

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47: 225–238.

Chamberlain, Gary. 1982. "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18: 5–46.

Chamberlain, Gary. 1984. Chapter 22 Panel data. In *Handbook of Econometrics*, 1247–1318. Vol. 2. Amsterdam: Elsevier.

Chamberlain, Gary. 1992. "Efficiency Bounds for Semiparametric Regression." *Econometrica* 60 (3): 567–596.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn and Whitney Newey. 2013. "Average and Quantile Effects in Nonseparable Panel Models." *Econometrica* 81 (2): 535–580.

Friedberg, Leora. 1998. "Did Unilateral Divorce Raise Divorce Rates? Evidence from Panel Data." *American Economic Review* 88 (3): 608–627.

Gentzkow, Matthew and Jesse Shapiro. 2013. "Measuring the Sensitivity of Parameter Estimates to Sample Statistics." Working paper University of Chicago.

Ghanem, Dalia. 2017. "Testing Identifying Assumptions in Nonseparable Panel Data Models." *Journal of Econometrics* 197 (2): 202–217.

Graham, Bryan S.and James L Powell. 2012. "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models." *Econometrica* 80 (5): 2105–2152.

Griffith, Rachel, Rupert Harrison and John Van Reenen. 2006. "How Special Is the Special Relationship? Using the Impact of U.S. R&D Spillovers on U.K. Firms as a Test of Technology Sourcing." *American Economic Review* 96 (5): 1859–1875.

Heckman, James J., and V. Joseph Hotz. 1989. “Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training.” *Journal of the American Statistical Association* 84 (408): 862–874.

Imai, Kosuke and In Song Kim. 2016. “When Should We Use Linear Fixed Effects Regression Moedles For Causal Inference With Longitudinal Data?” Working Paper.

Karlan, Dean S. and Jonathan Zinman. 2008. “Credit Elasticities in Less-Developed Economies: Implications for Microfinance.” *American Economic Review* 98 (3): 1040–1068.

Kline, Patrick and Andres Santos. 2012. “A Score Based Approach to Wild Bootstrap Inference.” *Journal of Econometric Methods* 1 (1): 23–41.

Kling, Jeffrey R. 2001. “Interpreting Instrumental Variables Estimates of the Returns to Schooling.” *Journal of Business & Economic Statistics* 19 (3): 358–364.

Lochner, Lance and Enrico Moretti. 2004. “The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports.” *American Economic Review* 94 (1): 155–189.

Meghir, Costas and Marten Palme. 2005. “Educational Reform, Ability, and Family Background.” *American Economic Review* 95 (1): 414–424.

Murphy, Kevin M. and Robert H. Topel. 1985. “Estimation and Inference in Two-Step Econometric Models.” *Journal of Business & Economic Statistics* 3 (4): 370–379.

Oreopoulos, Philip. 2006. “Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter.” *American Economic Review* 96 (1): 152–175.

Oster, Emily. 2014. Unobservable Selection and Coefficient Stability: Theory and Validation. Working Paper, University of Chicago.

Papke, Leslie E. 1994. “Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program.” *Journal of Public Economics* 54: 37–49.

Pérez-González, Francisco. 2006. “Inherited Control and Firm Performance.” *American Economic Review* 96 (5): 1559–1588.

Solon, Gary, Steven J. Haider and Jeffrey M. Wooldridge. 2015. “What are We Weighting For?” *Journal of Human Resources* 50 (2): 301–316.

Wooldridge, Jeffrey M. 1997. “On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model.” *Economic Letters* 56: 129–133.

Wooldridge, Jeffrey M. 2001. *Econometric Analysis of Cross-Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, Jeffrey M. 2005. “Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models.” *Review of Economics and Statistics* 87 (2): 385–390.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/jem-2017-0002).