NBER WORKING PAPER SERIES

BROKEN OR FIXED EFFECTS?

Charles E. Gibbons
Juan Carlos Suárez Serrato
Michael B. Urbancic

Working Paper 20342
http://www.nber.org/papers/w20342

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2014

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, Doug Miller, Juan Carlos Montoy, Enrico Moretti, Ron Oaxaca, Steve Raphael, 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. We also thank Stephen Lagos and Andrew Stanek for research assistance. Any remaining errors are the fault of the authors. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Broken or Fixed Effects?
Charles E. Gibbons, Juan Carlos Suárez Serrato, and Michael B. Urbancic
NBER Working Paper No. 20342
July 2014, Revised February 2016
JEL No. C18,C21

ABSTRACT

We replicate eight in 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 sample-weighted average treatment effect (SWE). We propose two alternative estimators that recover the SWE in the presence of group-specific heterogeneity. We document that heterogeneous treatment effects are common and the SWE 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 SWEs are either economically or statistically different from the FE estimates.

Charles E. Gibbons
The Brattle Group
201 Mission Street Suite 2800
San Francisco, CA 94105
charlie.gibbons@brattle.com

Juan Carlos Suárez Serrato
Department of Economics
Duke University
213 Social Sciences Building
Box 90097
Durham, NC 27708
and NBER
jc@jcsuarez.com

Michael B. Urbancic
Department of Economics
1285 University of Oregon
Eugene, OR 97403-1285
urbancic@uoregon.edu

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.

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:

$$y_i = x_i\beta_g + \mathbf{z}_i'\gamma + \epsilon_i, \tag{1}$$

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.

There is a long tradition in the econometrics literature considering the average partial effect (see, *e.g.*, Chamberlain, 1980, 1982, 1984, 1992; Blundell and Powell, 2003; Wooldridge, 1997, 2005; Graham and Powell, 2012). In this paper, we focus on the sample-weighted average treatment effect.[1]

Definition 1 (Sample-weighted average treatment effect (SWE))**.** *The sample-weighted average treatment effect (SWE) for Equation 1 is defined as*

$$\beta^{SWE} \equiv \sum_g \frac{N_g}{N}\beta_g,$$

where N_g is the number of observations in 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

[1]The distinguishing feature between the SWE and the average partial effect is that we are agnostic as to whether the sample is representative of a population of interest.

average that generally does not coincide with the sample-weighted effect.[2] 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.

In Section 1, we derive the fixed effects (FE) estimator (*i.e.*, an OLS model that does not account for treatment effect heterogeneity) 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 SWE under group-specific heterogeneity and derive the joint asymptotic distribution of these estimators with the FE.

Our main contribution is empirically judging the importance of the distinction between the FE and the SWE. We replicate eight influential papers from the *American Economic Review* published between 2004 and 2009.[3] Using these examples, we consider a randomized experiment in Section 2 as a case study and, in Section 3, we show generally that heterogeneous treatment effects are common and that the FE and SWE are often different in statistically and economically significant degrees. In all but one paper, there is at least one statistically significant source of treatment effect heterogeneity. In five papers, this heterogeneity induces the SWE 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 these results, we conclude that methods that consistently estimate the SWE offer more interpretable results than standard FE models.

Comparison to the literature. While all our applications are in cross-sectional contexts, there is a large literature analyzing non-separable correlated heterogeneity in panel data contexts. In particular, our approach can be viewed as a special case of the CRC model of Chamberlain (1982) (see also Chamberlain, 1984, 1992). 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

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

[3]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.

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. As a last example, Chernozhukov et al. (2013) study quantile treatment effects under non separable heterogeneity. 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. 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 Estimating the Sample-Weighted Effect

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:

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

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

where β is now a vector of coefficients. $\mathbf{Y}$, $\mathbf{X}$, and $\boldsymbol{\epsilon}$ are vectors across the N observations and $\mathbf{A}$ and $\mathbf{Z}$ are 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.

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:*

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

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

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

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

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 SWE.

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 SWE 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 2, we replicate 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 find that the SWE differs from the FE estimate by 61%.

We offer two alternative estimators for the SWE 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 SWE. 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 SWE. Thus, the IWE is given by:*

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

where[5]

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

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 SWE. 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*

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

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

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

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

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

that the RWE can be re-written as:

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

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 SWE (see Appendix A.1). Furthermore, they are consistent, which we illustrate by deriving the joint asymptotic distribution of the three estimators.[6] To do so, we first define Ω to be the variance-covariance matrix of $\boldsymbol{\epsilon}$, 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 Appendix A.1 and,* e.g.*, Wooldridge (2001)), the asymptotic distribution of the estimators is*

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

where

$$\begin{aligned} \Sigma_{FE} &= \left(\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{N}\right)^{-1} \frac{\tilde{\mathbf{X}}'\Omega\tilde{\mathbf{X}}}{N} \left(\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{N}\right)^{-1} & \Sigma_{12} &= \left(\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{N}\right)^{-1} \frac{\tilde{\mathbf{X}}'\Omega\tilde{\mathbf{A}}}{N} \left(\frac{\tilde{\mathbf{A}}'\tilde{\mathbf{A}}}{N}\right)^{-1} \mathbf{f}' \\ \Sigma_{IWE} &= \mathbf{f} \left(\frac{\tilde{\mathbf{A}}'\tilde{\mathbf{A}}}{N}\right)^{-1} \frac{\tilde{\mathbf{A}}'\Omega\tilde{\mathbf{A}}}{N} \left(\frac{\tilde{\mathbf{A}}'\tilde{\mathbf{A}}}{N}\right)^{-1} \mathbf{f}' & \Sigma_{13} &= \left(\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{N}\right)^{-1} \frac{\tilde{\mathbf{X}}'\Omega\mathbf{W}\tilde{\mathbf{X}}}{N} \left(\frac{\tilde{\mathbf{X}}'\mathbf{W}\tilde{\mathbf{X}}}{N}\right)^{-1} \\ \Sigma_{RWE} &= \left(\frac{\tilde{\mathbf{X}}'\mathbf{W}\tilde{\mathbf{X}}}{N}\right)^{-1} \frac{\tilde{\mathbf{X}}'\mathbf{W}\Omega\mathbf{W}\tilde{\mathbf{X}}}{N} \left(\frac{\tilde{\mathbf{X}}'\mathbf{W}\tilde{\mathbf{X}}}{N}\right)^{-1} & \Sigma_{23} &= \mathbf{f} \left(\frac{\tilde{\mathbf{A}}'\tilde{\mathbf{A}}}{N}\right)^{-1} \frac{\tilde{\mathbf{A}}'\Omega\mathbf{W}\tilde{\mathbf{X}}}{N} \left(\frac{\tilde{\mathbf{X}}'\mathbf{W}\tilde{\mathbf{X}}}{N}\right)^{-1}. \end{aligned}$$

Remarks.

1. 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

[6] The fixed effects that we consider denote group membership and the sizes of these groups grow with overall sample size (*i.e.*, $N_g \to \infty \, \forall \, g$). As a result, we do not have the "small T" or incidental parameters problem common in panel data models that would preclude the application of asymptotic results.

heterogeneity and the level of clustering are the same or colinear.

2. 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.

3. 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. For the details of these tests, see Appendix A.2.

4. Given the asymptotic result in Proposition 2, it is straightforward to perform a test of equality between either estimate of the SWE and the FE estimate. See Appendix A.2 for details.

5. These results can be confirmed using a Monte Carlo simulation; see Appendix B.

2 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 SWE. 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. Recall, however, that differing means in treatment do not drive the difference between the FE and SWE 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 SWE if the responsiveness to interest rates varies across risk groups.

The FE weights are given in column 2 of Table 1. These are the relative variances of treatment by group multiplied by the sample frequency of that group. Using these weights and the group effect estimated using an interacted model (given in column 4 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 SWE. 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.

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 nearly 70%.

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

Risk group	FE weight	Sample weight	Effect
Low	0.043	0.125	-32.4
Medium	0.060	0.092	-9.9
High	0.897	0.783	-2.7
Average	-4.403	-7.050	
Std. error	(1.08)	(1.92)	

Notes: The SWE 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.

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

To consider the empirical relevance of the distinction between the FE and SWE 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 FE 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 Appendix D.1.

To consider whether the difference between the FE and SWE estimators is empirically important, we test for heterogeneous treatment effects and for a difference between the FE and

SWE estimates.[7] 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 SWE 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.[8] The results for the IWE are generally very similar, as we would expect, and these results are included in the detailed tables of Appendix D.3.

We find 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 SWE and FE estimates for five papers at the 5% level and seven papers at the 10% level. Defining a difference to be "economically significant" if it exceeds 10%, we find that five papers have economically significant differences between the SWE 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 SWE. 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.[9] The SWE estimator exhibits standard errors that are less than ten percent larger than those for the FE in six of eight cases.[10] 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

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

[8] In 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.

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

[10] It is perhaps not surprising that the standard errors for Karlan and Zinman (2008) increase substantially given the large change in the estimate (nearly 70% for the RWE). But the t-statistics are similar: -4.00 using FE and -3.94 using the RWE.

(*i.e.*, the bias is high), then the SWE should be preferred for policy and interpretablity reasons.

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 irrigated Proportion other cereals Proportion rice Proportion wheat Proportion white rice Rice yield (log) Wheat yield (log)	Proportion non-landlord land	Coastal dummy, year	3	1
Bedard and Deschênes (2006)	Smoking dummy	Veteran status	Age, education, race, region	5	1
Card et al. (2008)	Saw doctor dummy Was hospitalized dummy	Age over 65 dummy	Ethnicity, gender, region, year, education level	3	6, 8
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 Palme (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

Notes: Additional details on our replications are found in Appendix D.

Table 3: *AER* replication results

Citation	Fixed effect	Joint test (p-value)	Diff. test (p-value)	Percent diff.	In paper
(1)	(2)	(3)	(4)	(5)	(6)
Banerjee and Iyer (2005)	Coastal	0.065*	0.013**	-31.7†	
(Proportion irrigated)	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 et al. (2008)	Ethnicity (*outcome: saw doctor*)	0.000***	0.291	-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***	-1.6	X
	Age (blacks only)	0.000***	0.000***	31.7†	
	Year (blacks only)	0.000***	0.000***	1.8	
	Age (whites only)	0.000***	0.000***	29.0†	
	Year (whites only)	0.000***	0.000***	-0.2	
Meghir and Palme (2005)	High father's education	0.000***	0.000***	16.0†	X
	Gender	0.326	0.517	0.3	X
	Year	0.000***	0.351	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)	0.592	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	

Notes: 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 SWE 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.

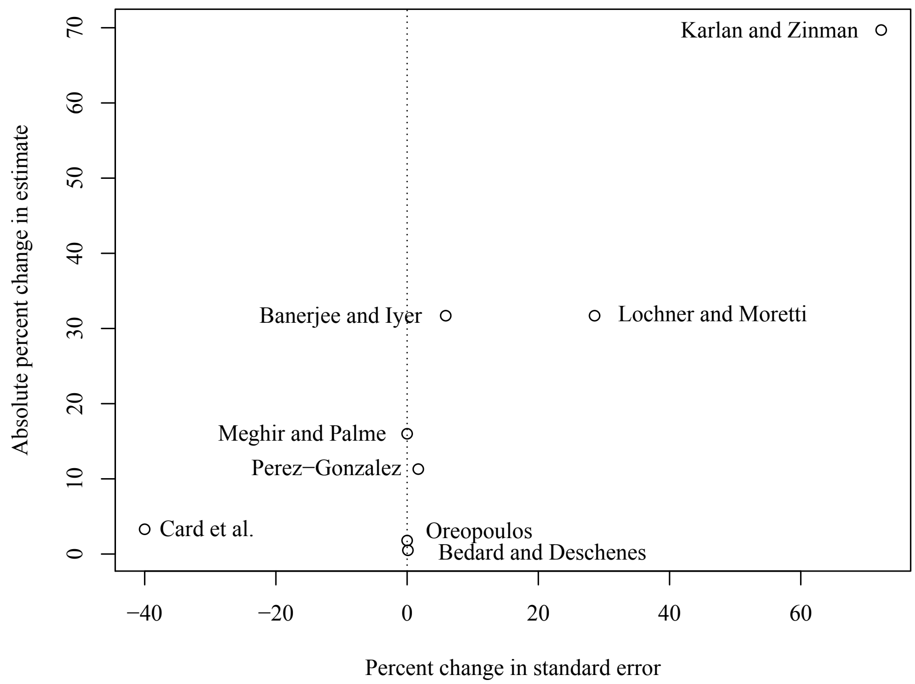

Figure 1: The relationship between the difference in the estimates and the change in variance among the *AER* replications

Notes: Figure is based on the full results presented in 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 sample-weighted 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 the interaction-weighted or regression-weighted estimators or by analyzing the group-specific weights implied by OLS with fixed effects. We believe that reporting sample-weighted 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 is to analyze heterogeneity in treatment effects across observable groups. Heterogeneity may also arise along unobservable margins (see, *e.g.*, Bitler, Gelbach and Hoynes, 2014).

References

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

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 Elsevier.

Angrist, Joshua and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. 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 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*. Vol. 2 Elsevier pp. 1247 – 1318.

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.

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.

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*. 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.

For Online Publication

A Derivation of Results

A.1 Derivation of the Asymptotic Distribution

To derive the asymptotic distribution of our estimators, we use standard assumptions for our data-generating process (see, *e.g.*, Wooldridge, 2001).

Assumptions. Consider the following assumptions for the model defined by Equation 1:

1. Exogeneity: $\mathbb{E}[\boldsymbol{\epsilon} \mid \mathbf{A}, \mathbf{Z}] = 0$
2. Identification:
   (a) $\text{Var}(\tilde{x}_i \mid g(i) = g) > 0 \quad \forall\, g \in 1, \ldots, G$
   (b) $\mathbb{E}[\mathbf{Z}'\mathbf{Z}]$ is non-singular
3. Random sampling:
   (a) (Heteroskedasticity-robust standard errors) $\{(y_i, \mathbf{a}_i, \mathbf{z}_i)\}_{i=1}^{N}$ are i.i.d. draws from a distribution that satisfies Equation 1.
   (b) (Cluster-robust standard errors) In the case of a clustered variance-covariance matrix Ω, $\{(\mathbf{Y}_c, \mathbf{A}_c, \mathbf{Z}_c)\}_{c=1}^{C}$, where $\mathbf{Y}_c$ is a vector of outcomes for observations in cluster c, with $\mathbf{A}_c$ and $\mathbf{Z}_c$ defined similarly and C fixed, are i.i.d. draws from a distribution that satisfies Equation 1.
4. Convergence of the variance-covariance matrix: The fourth moments of y_i, $\mathbf{a}_i$, and $\mathbf{z}_i$ exist.

We note that the IWE and RWE are both unbiased estimators of the SWE.

Proposition 3 (Unbiasedness of IWE)**.** *Under the assumptions above, the IWE for Equation 1 is unbiased:*

$$\mathbb{E}\left[\hat{b}^{IWE} \middle| \mathbf{X}, \mathbf{Z}, \mathbf{A}\right] = \mathbf{f}\beta = \beta^{SWE}.$$

Proposition 4 (Unbiasedness of RWE)**.** *Under the assumptions above, the RWE for Equation 1 is unbiased:*

$$\mathbb{E}\left[\hat{b}^{RWE} \middle| \mathbf{X}, \mathbf{Z}, \mathbf{A}\right] = \frac{1}{N}\sum_g N_g \frac{\widehat{\text{Var}}(\tilde{x}_i, \tilde{y}_i \mid g)}{\widehat{\text{Var}}(\tilde{x}_i \mid g)} = \sum_g \frac{N_g}{N}\beta_g = \beta^{SWE}$$

Given these assumptions, standard law of large number and central limit theorem results demonstrate that the estimators converge to their respective expected values and have the asymptotic variances given in Proposition 2 (see, *e.g.*, Wooldridge, 2001).

A.2 Testing for Heterogeneous Treatment Effects

Armed with two estimators of the SWE, 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 SWE and the FE. These tests are implemented by Stata commands and an `R` package available from the authors, as discussed in Appendix C.

A.2.1 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 5 (Wald test for modeled heterogeneity)**.** *The Wald test statistic for heterogeneous treatment effects is calculated according to*

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

where

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

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

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

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

A.2.2 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.[11] The score is calculated according to

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

Proposition 6 (Score test for unmodeled heterogeneity)**.** *A score test statistic for the presence of heterogeneous treatment effects has the form*[12]

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

where

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

and

$$\mathbf{C}=\begin{bmatrix}\mathbf{0}_{(G-1)\times(K+1)} & \mathbf{I}_{G-1}\end{bmatrix}$$

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

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

Like the Wald test above, this test statistic has an asymptotic χ^2_{G-1} distribution under the null hypothesis.[13]

A.2.3 Test for Equality Between the SWE and FE Estimates

Even if heterogeneous treatment effects are present, the SWE 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 SWE (*i.e.*, RWE or IWE) and we refer to the chosen estimator as $\hat{b}^{SWE}$.

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

[12] 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).

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

Proposition 7 (Specification test of the differences between the FE and SWE estimates)**.** *The test of the following null hypothesis*

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

can be conducted by noting that the Wald test statistic

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

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.*

B Monte Carlo Results

This appendix explores the properties of the three estimators that we consider (*i.e.*, FE, RWE, IWE) using Monte Carlo experiments. We generate 1000 simulated datasets with 1000 observations according to the following equation:

$$y_i = \alpha_g + x_i\beta_g + z_i\gamma + \epsilon_i,$$

where α_g is one of five group fixed effects with each group having an equal fraction of observations. x_i and z_i are each scalars.

We first analyze the case where the true data generating process is a model of homogenous treatment effects and show that all estimators provide consistent estimates. In particular, we set $\beta_g = 3.5$ for all g, $\gamma = 0.75$, let $\mathrm{Cov}\left(x_i, z_i\right) = 0.3$, and allow the variance of x_i to depend on g as follows:

$$\mathrm{Var}\left(x_i|g(i)=g\right)=\begin{cases} 58.33 & \text{if} \quad \mathrm{g}=1 \\ 15.03 & \text{if} \quad \mathrm{g}=2 \\ 7.39 & \text{if} \quad \mathrm{g}=3 \\ 4.57 & \text{if} \quad \mathrm{g}=4 \\ 2.18 & \text{if} \quad \mathrm{g}=5. \end{cases}$$

Panel A of Table 4 displays the means and standard deviations of the estimates and analytic standard errors for each of the estimators. The mean estimate of β is very close to the true value for all three approaches when treatment effects are heterogeneous. The mean analytic standard errors are also close to the Monte Carlo estimates of those statistics (*i.e.*, the standard deviation of the βs)

We now explore the effect of allowing β_g to vary by group as follows:

$$\beta_g=\begin{cases} \text{-0.5} & \text{if} \quad \mathrm{g}=1 \\ 1.5 & \text{if} \quad \mathrm{g}=2 \\ 3.5 & \text{if} \quad \mathrm{g}=3 \\ 5.5 & \text{if} \quad \mathrm{g}=4 \\ 7.5 & \text{if} \quad \mathrm{g}=5. \end{cases}$$

Because each group has the same number of observations, the SWE is still 3.5. Since the variance of x_i is greater for groups with below-mean β_g's, however, FE will be biased downwards. Panel B of Table 4 displays the results from this exercise and confirms this result. Note that the standard errors of the IWE and RWE estimators are very similar, thus it does not appear that either is preferred on efficiency grounds under this data-generating process.

C Implementation of the Estimators and Tests in Stata and R

As a companion to this paper, we develop Stata commands and an R package that tests for heterogeneity using both the Wald and score tests, estimates the FE, IWE, and RWE, performs the specification test for each SWE estimator, and computes the percentage difference between each SWE estimate and the OLS estimate. These packages are available from the authors; basic syntax

Table 4: Monte Carlo results

	Mean	Std. dev.
Panel A: Homogeneous Effects		
	Mean	Std. dev.
Fixed Effect: β	3.502	0.071
Fixed Effect: SE	0.069	0.002
IWE: β	3.501	0.119
IWE: SE	0.116	0.006
RWE: β	3.502	0.119
RWE: SE	0.117	0.006
Panel B: Heterogeneous Effects		
	Mean	Std. dev.
Fixed Effect: β	0.715	0.071
Fixed Effect: SE	0.096	0.002
IWE: β	3.501	0.119
IWE: SE	0.116	0.006
RWE: β	3.503	0.119
RWE: SE	0.119	0.008
Number of observations	1000	
Number of simulations	1000	

is discussed below.

C.1 Stata Commands

The `ado` file `GSSUtest.ado` contains the command `GSSUtest`, which estimates the IWE and performs the Wald test and the specification test of equality between the OLS estimate and the IWE. The command has the syntax:

```
GSSUtest y Tr FEg [varlist] [if] [in] [, vce(string) cluster(clustervar)]
```

where

- `y` is the dependent variable;
- `Tr` is the independent variable of interest (*e.g.*, treatment); and
- `FEg` is a categorical variable indexing the fixed effect group.

Other predictors can be included in `varlist`. For homoskedastic errors, ignore the `vce()` and `cluster()` options. For heteroskedastic-robust standard errors, use the option `vce(robust)` and for cluster-robust standard errors, specify `cluster(clustervar)`.

The `ado` file `GSSUwtest.ado` contains the command `GSSUwtest`, which has the same syntax as above and estimates the RWE and performs the specification test of equality between the OLS estimate and the RWE. Standard errors can be computed to be robust or cluster-robust.

The `intscoretest` command in the `ado` file `intscoretest.ado` has the same syntax and performs the score test on the interactions between the treatment variable and the fixed effects. Standard errors can be computed to be heteroskedastic robust.

The `ado` file `GSSUgetrdone.ado` offers the command `GSSUgetrdone`, which has the same syntax and runs all three commands above and displays the results. `GSSUgetrdone` automatically uses robust standard errors in its calculations.

The results from all of the commands can be accessed through matrices stored after execution. Type `ereturn list` to list them.

The Stata package can be installed using the following commands:

```
* Loads website
net from http://www.jcsuarez.com/GSSU
* Describes package
net describe GSSU
* Installs commands
net install GSSU
* Downloads example data
net get GSSU
* Installs required package for GSSUgetrdone.ado
ssc install estout, replace
```

C.2 R Package

To estimate the IWE, use the function:

```
EstimateIWE(y, treatment, group, controls, fe.other, data, subset,
  cluster.var, is.robust, is.data.returned)
```

The RWE is estimated analogously:

```
EstimateRWE(y, treatment, group, controls, fe.other, data, subset,
  cluster.var, is.robust, is.data.returned)
```

where, for both:

- `y` is the name of the outcome variable;
- `treatment` is the name of the treatment variable;
- `group` is the name of the fixed effect group of interest;
- `controls` is a character vector of the names of other control variables;
- `fe.other` is a character vector of the names of other fixed effects in the model;
- `data` is the data frame to be used for estimation;
- `subset` is an optional subset declaration;
- `cluster.var` is the name of the variable used for clustered standard errors;
- `is.robust` is a logical indicating whether robust standard errors should be used; and
- `is.data.returned` is a logical indicating whether the `data` data frame should be returned with the estimation results.

For either estimation procedure, a specification test and the score test(see Appendix A.2) are conducted by:

```
SpecTest(model, data)
ScoreTest(model, data)
```

where `model` is the result of one of the estimation procedures above and `data` is the corresponding data frame. The Wald test (see Section A.2.1) is only conducted for the IWE estimator and has the form

```
WaldTestIWE(model)
```

The `R` package can be installed using the following commands:

```
install.packages('http://cgibbons.us/research/packages/GSSU.tar.gz',
  type = 'source', repos = NULL)
```

D *AER* Replications

D.1 Paper Selection

In this paper, our goal is to determine whether the difference between an estimator of the SWE and the FE estimator is empirically important. We do this by replicating high quality papers from the *AER*. We examine a breadth of papers that covers several fields, several years, and several units of analysis and thus they serve as a decent representation of the use of fixed effects in the applied econometrics literature.

Our guidelines for paper selection are:

- The paper must have been published in the *American Economic Review*. We choose this qualification in order to limit our universe of analysis both in terms of quantity and quality of papers considered and to guarantee easy access to the necessary data.

- The paper must be published in the March 2004 issue or later (to March 2009, the issue predating our literature search). The *AER* policy during this period requires that, barring any acceptable restriction, the data for these papers be posted to the EconLit website. This leads to the condition that:

- The data necessary to replicate the main specification(s) of the paper must be readily available on the EconLit website.[14] We use these data and direct those interested to the EconLit website to obtain these files.

- The main specification(s) of the paper must have a specific effect of interest.[15]

- The main specification(s) of the paper must use some type of fixed effect. We identify papers meeting this qualification by searching the PDF files of the published papers for the terms "fixed effect" (which captures the plural "effects" as well) and for "dumm" (which captures "dummy" or "dummies," common synonyms for fixed effects).

- We limit ourselves to microeconomic analyses and do not consider papers based on financial economics issues.

[14] We determine which specifications are the "main" ones by considering the discussion of the effects in the text by the authors and ignore those specifications identified as robustness checks.

[15] In a previous version of this paper, we included a paper by Griffith, Harrison and Van Reenen (2006). Upon reflection, this paper does not satisfy this criterion and has been removed from consideration.

- We ignore papers that require special methods to handle time series issues.

We choose to replicate a total of eight papers in our analysis. To order our search, we consider papers in order of citations per year since publication. First, we use the citation counts provided by the ISI Web of Science on July 16, 2009. We limit our search to the *American Economic Review* and the years 2004–2009, as outlined above. Unfortunately, the Web of Science does not provide the volume for the papers contained therein. Instead, we create an algorithm that assigns a volume number to a paper based upon its page number; these assignments are verified as papers are considered. The total number of citations are divided by the years since publication. For example, in June 2009, a paper published in June 2004 was published 5 years before and a paper published in September 2004 was published 4.75 years before.

Citation counts are very noisy in the short time after publication that we consider here. Our citations-per-year metric might overweight later papers.[16] Nonetheless, the eight selected papers are drawn from a universe that includes all papers in this period with over 20 citations and 86% of all papers with 15 or more citations. It appears that we screen most of the highly-cited papers from this period and do not ignore the most recent papers, as would occur using the gross citation count.

Before estimating the SWE for the papers that we consider, we first ensure that we can replicate the results obtained by the authors as given in their respective papers. We can provide Stata `DO` and `log` files that generate and produce these results. We add our estimation procedures to these files as well.[17]

In choosing the fixed effects groups to consider when there are several fixed effects in the regressions, we choose such that the number of groups is not unruly (U.S. states, for example, may produce too many terms to be informative). Our interacted regressions preserve all other features of the replicated specifications (*e.g.*, clustering, robust standard errors, and inclusion of other covariates) unless otherwise noted in the text.

We do not claim that the source of heterogeneity that we consider is the most salient within the given economic situation. Additionally, we do not suggest that modeling treatment effect heterogeneity is the first-order extension of the analysis in the papers that we examine. We make

[16] In June 2009, 1 citation for a paper published in March 2009 is equal to 4 for a paper published in June 2008 and 20 for a paper published in June 2004.

[17] See Section C.

no effort to search the subsequent literature to identify other areas of concern in these papers. Lastly, many of these papers employ instrumental variables to combat endogeneity. In these cases, we use the base OLS case to illustrate our point.

D.2 Replication Details

We replicate the specifications cited in Table 2. Some of these authors include fixed effects interactions or run regressions separately for subgroups; we list these practices in Table 5. In Banerjee and Iyer (2005), the authors have eight separate outcomes of interest. In the body of the paper, we give results only for a subset of these results.

Table 5: Fixed effects interactions and regressions by subgroup conducted in the original papers

Citation	Separate regressions	Interactions
Banerjee and Iyer (2005)	Entire country, subregion	
Bedard and Deschênes (2006)		
Card, Dobkin and Maestas (2008)	Race $\times$ education	Age, age-squared
Karlan and Zinman (2008)		
Lochner and Moretti (2004)	Race (black, white)	
Meghir and Palme (2005)	Sex (male, female) Father's education (low, high) Ability (low, high) Ability $\times$ father's education $\times$ sex	Sex (male, female in full sample OLS)
Oreopoulos (2006)	Country	
Pérez-González (2006)		Less selective college attendance dummy Graduate school attendance dummy Positive R&D expenditure dummy

Notes: Separate regressions and interaction terms only listed for specifications based upon the one given in Table 2. Pérez-González (2006) does not include the dummy variables that he subsequently interacts with treatment in his base regression; hence, we do not test their interactions here.

D.3 Detailed Results

In this subsection, we presented detailed results for each paper. Because the IWE and RWE results are similar, we discuss only the RWE results in the body of the paper; here, we present both sets. If clustering was used by the paper's author, we provide both the clustered and non-clustered heteroskedasticity-robust results.[18] The estimates are given along with standard errors in parentheses. A single star indicates significance at the 10% level, two stars significance at the 5% level, and three stars indicate significance at the 1% level.

In each table, tests for heterogeneous treatment effects are given. The Wald test is used for the IWE estimator and the score test is used for the RWE estimator.[19] Specification tests for the difference between the SWE and FE estimates are conduced using the Wald statistic and an asymptotic normal approximation.

Lastly, we note that we are not able to replicate the point estimate that Oreopoulos (2006) provides for his regression of Northern Ireland and Great Britain combined; we use the specification that he provides and base our results on this model.

[18] Bedard and Deschênes (2006) and Pérez-González (2006) do not use clustered standard errors.

[19] The Wald test is natural when the interaction coefficients are actually calculated, whereas the score test is natural when they are not, hence the pairings chosen here.

Table 6: Banerjee and Iyer (2005)

(a) Fertilizer with coastal interaction

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	10.708***		10.333***	10.708***	10.867***	10.333***
	(3.345)		(3.588)	(1.020)	(0.907)	(1.008)
Het. test stat.			0.787		3.180	0.787
Het. test p-value			0.375		0.075	0.375
Spec. test stat.			0.178		1.726	2.045
Spec. test p-value			0.673		0.084	0.153
Percent change			−3.502		1.489	−3.502

(b) Fertilizer with year interactions

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	10.708***	10.740***	10.738***	10.708***	10.740***	10.738***
	(3.345)	(3.338)	(3.342)	(1.020)	(0.895)	(0.922)
Het. test stat.		124.522	139.293		263.139	139.293
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		0.563	7.230		0.172	77.485
Spec. test p-value		0.573	0.007		0.863	0.000
Percent change		0.304	0.287		0.304	0.287

(c) Log total yield with coastal interaction

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.157**		0.142*	0.157***	0.151***	0.142***
	(0.071)		(0.074)	(0.015)	(0.015)	(0.015)
Het. test stat.			5.487		26.277	5.487
Het. test p-value			0.019		0.000	0.019
Spec. test stat.			0.881		−4.386	21.152
Spec. test p-value			0.348		0.000	0.000
Percent change			−9.611		−4.239	−9.611

(d) Log total yield with year interactions

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.157**	0.157**	0.157**	0.157***	0.157***	0.157***
	(0.071)	(0.071)	(0.071)	(0.015)	(0.015)	(0.015)
Het. test stat.		274.215	126.335		82.683	126.335
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		0.275	9.096		0.002	4.412
Spec. test p-value		0.783	0.003		0.998	0.036
Percent change		0.003	0.012		0.003	0.012

(e) Log rice yield with coastal interaction

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.171^{**}		0.171^{**}	0.171^{***}	0.165^{***}	0.171^{***}
	(0.081)		(0.080)	(0.017)	(0.020)	(0.020)
Het. test stat.			1.936		18.466	1.936
Het. test p-value			0.164		0.000	0.164
Spec. test stat.			0.000		−3.765	0.000
Spec. test p-value			0.997		0.000	0.988
Percent change			0.031		−3.296	0.031

(f) Log rice yield with year interactions

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
nland	0.171^{**}	0.170^{**}	0.170^{**}	0.171^{***}	0.170^{***}	0.170^{***}
	(0.081)	(0.081)	(0.081)	(0.017)	(0.020)	(0.020)
Het. test stat.		171.874	123.681		103.150	123.681
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		−0.559	6.281		−0.074	6.626
Spec. test p-value		0.576	0.012		0.941	0.010
Percent change		−0.123	−0.123		−0.123	−0.123

(g) Percent HYV cereals with coastal interaction

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.057^{*}		0.059^{*}	0.057^{***}	0.058^{***}	0.059^{***}
	(0.031)		(0.032)	(0.010)	(0.009)	(0.010)
Het. test stat.			0.170		0.391	0.170
Het. test p-value			0.680		0.532	0.680
Spec. test stat.			0.058		0.629	0.413
Spec. test p-value			0.809		0.529	0.520
Percent change			3.281		1.131	3.281

(h) Percent HYV cereals with year interactions

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.057^{*}	0.057^{*}	0.057^{*}	0.057^{***}	0.057^{***}	0.057^{***}
	(0.031)	(0.031)	(0.031)	(0.010)	(0.009)	(0.009)
Het. test stat.		78.041	88.748		65.746	88.748
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		0.330	0.313		0.092	0.678
Spec. test p-value		0.742	0.576		0.926	0.410
Percent change		0.173	−0.191		0.173	−0.191

(i) Percent HYV rice with coastal interaction

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.079*		0.078*	0.079***	0.080***	0.078***
		(0.043)	(0.042)	(0.012)	(0.012)	(0.012)
Het. test stat.			0.041		1.231	0.041
Het. test p-value			0.840		0.267	0.840
Spec. test stat.			0.055		1.095	0.467
Spec. test p-value			0.815		0.274	0.494
Percent change			−1.725		1.099	−1.725

(j) Percent HYV rice with year interactions

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.079*	0.079*	0.079*	0.079***	0.079***	0.079***
	(0.044)	(0.044)	(0.043)	(0.012)	(0.012)	(0.012)
Het. test stat.		108.783	76.353		280.287	76.353
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		−0.205	0.005		−0.026	0.004
Spec. test p-value		0.838	0.945		0.979	0.950
Percent change		−0.079	−0.018		−0.079	−0.018

(k) Percent HYV wheat with coastal interaction

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.092**		0.072	0.092***	0.080***	0.072***
	(0.046)		(0.047)	(0.012)	(0.013)	(0.014)
Het. test stat.			0.526		82.283	0.526
Het. test p-value			0.468		0.000	0.468
Spec. test stat.			3.468		−5.412	37.519
Spec. test p-value			0.063		0.000	0.000
Percent change			−21.610		−13.337	−21.610

(l) Percent HYV wheat with year interactions

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.092**	0.091**	0.091**	0.092***	0.091***	0.091***
	(0.046)	(0.045)	(0.046)	(0.012)	(0.013)	(0.013)
Het. test stat.		179.014	69.347		126.897	69.347
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		−0.581	5.273		−0.311	2.227
Spec. test p-value		0.561	0.022		0.756	0.136
Percent change		−0.793	−0.514		−0.793	−0.514

(m) Irrigation with coastal interaction

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.065*		0.045	0.065***	0.061***	0.045***
	(0.034)		(0.036)	(0.008)	(0.007)	(0.008)
Het. test stat.			3.414		34.449	3.414
Het. test p-value			0.065		0.000	0.065
Spec. test stat.			6.219		−4.402	147.436
Spec. test p-value			0.013		0.000	0.000
Percent change			−31.655		−6.785	−31.655

(n) Irrigation with year interactions

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.065*	0.065*	0.065*	0.065***	0.065***	0.065***
	(0.034)	(0.034)	(0.034)	(0.008)	(0.007)	(0.007)
Het. test stat.		84.841	80.741		7.622	80.741
Het. test p-value		0.000	0.000		1.000	0.000
Spec. test stat.		0.053	0.017		0.010	0.017
Spec. test p-value		0.958	0.896		0.992	0.897
Percent change		0.006	0.005		0.006	0.005

Table 7: Bedard and Deschenes (2006)

(a) Age interactions

	FE	IWE	RWE
	0.078***	0.078***	0.077***
	(0.005)	(0.006)	(0.006)
Het. test stat.		11.090	11.142
Het. test p-value		0.944	0.942
Spec. test stat.		0.108	0.046
Spec. test p-value		0.914	0.830
Percent change		0.111	−0.223

(b) Education interactions

	FE	IWE	RWE
	0.078***	0.078***	0.078***
	(0.005)	(0.006)	(0.006)
Het. test stat.		14.788	14.918
Het. test p-value		0.002	0.002
Spec. test stat.		0.890	0.025
Spec. test p-value		0.374	0.875
Percent change		0.712	−0.124

(c) Race interactions

	FE	IWE	RWE
	0.078***	0.078***	0.078***
	(0.005)	(0.005)	(0.005)
Het. test stat.		3.069	3.073
Het. test p-value		0.080	0.080
Spec. test stat.		1.700	2.978
Spec. test p-value		0.089	0.084
Percent change		0.524	0.494

(d) Region interactions

	FE	IWE	RWE
	0.078***	0.078***	0.078***
	(0.005)	(0.005)	(0.005)
Het. test stat.		5.514	5.557
Het. test p-value		0.701	0.697
Spec. test stat.		1.231	0.734
Spec. test p-value		0.218	0.392
Percent change		0.245	0.075

Table 8: Card et al. (2008)

(a) Hospitalized; education interactions (whites only)

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.012**	0.012***	0.012***	0.012**	0.012**	0.012**
	(0.005)	(0.004)	(0.004)	(0.006)	(0.006)	(0.006)
Het. test stat.		14.526	6.350		11.513	6.350
Het. test p-value		0.002	0.096		0.009	0.096
Spec. test stat.		2.105	0.619		1.891	0.665
Spec. test p-value		0.035	0.431		0.059	0.415
Percent change		1.601	1.045		1.601	1.045

(b) Hospitalized; education interactions (non-whites only)

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.013	0.013**	0.013**	0.013	0.013	0.013
	(0.010)	(0.006)	(0.006)	(0.010)	(0.010)	(0.010)
Het. test stat.		0.609	1.242		0.661	1.242
Het. test p-value		0.894	0.743		0.882	0.743
Spec. test stat.		0.720	1.090		0.765	1.262
Spec. test p-value		0.472	0.296		0.444	0.261
Percent change		1.462	3.332		1.462	3.332

(c) Hospitalized; ethnicity interactions

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.012***	0.012***	0.012***	0.012**	0.012**	0.012**
	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)
Het. test stat.		16.479	16.798		15.917	16.798
Het. test p-value		0.001	0.001		0.001	0.001
Spec. test stat.		0.623	0.254		0.716	0.132
Spec. test p-value		0.533	0.614		0.474	0.717
Percent change		0.354	−0.142		0.354	−0.142

(d) Hospitalized; gender interaction

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.012**	0.012***	0.012***	0.012**	0.012**	0.012**
	(0.005)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)
Het. test stat.		22.513	22.838		22.119	22.838
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		−2.125	3.335		−1.792	3.632
Spec. test p-value		0.034	0.068		0.073	0.057
Percent change		−0.954	−0.485		−0.954	−0.485

(e) Hospitalized; region interactions

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.012**	0.012***	0.012***	0.012**	0.012**	0.012**
	(0.005)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)
Het. test stat.		10.712	13.392		10.034	13.392
Het. test p-value		0.013	0.004		0.018	0.004
Spec. test stat.		0.455	1.068		0.427	1.319
Spec. test p-value		0.649	0.301		0.670	0.251
Percent change		0.145	0.179		0.145	0.179

(f) Hospitalized; year interactions

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.012**	0.012***	0.012***	0.012**	0.012**	0.012**
	(0.005)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)
Het. test stat.		8.886	11.751		12.256	11.751
Het. test p-value		0.632	0.383		0.345	0.383
Spec. test stat.		0.320	0.606		0.327	0.734
Spec. test p-value		0.749	0.436		0.743	0.392
Percent change		0.259	−1.250		0.259	−1.250

(g) Saw doctor; education interactions (whites only)

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.008	0.008	0.008	0.008	0.008	0.008
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Het. test stat.		16.643	9.133		19.725	9.133
Het. test p-value		0.001	0.028		0.000	0.028
Spec. test stat.		−2.783	1.179		−2.414	1.752
Spec. test p-value		0.005	0.278		0.016	0.186
Percent change		−4.283	−2.008		−4.283	−2.008

(h) Saw doctor; education interactions (non-whites only)

	Clustering			No clustering		
	FE	*IWE*	*RWE*	*FE*	*IWE*	*RWE*
	0.038***	0.037***	0.038***	0.038***	0.037***	0.038***
	(0.014)	(0.011)	(0.011)	(0.014)	(0.014)	(0.014)
Het. test stat.		4.999	0.262		4.094	0.262
Het. test p-value		0.172	0.967		0.252	0.967
Spec. test stat.		−1.757	0.066		−1.722	0.062
Spec. test p-value		0.079	0.798		0.085	0.804
Percent change		−1.652	−0.370		−1.652	−0.370

(i) Saw doctor; ethnicity interactions

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.016**	0.016**	0.016**	0.016**	0.016**	0.016**
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Het. test stat.		27.968	27.804		31.706	27.804
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		−1.103	1.567		−1.144	1.416
Spec. test p-value		0.270	0.211		0.253	0.234
Percent change		−0.868	−0.501		−0.868	−0.501

(j) Saw doctor; gender interaction

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
d65	0.016*	0.015**	0.016**	0.016**	0.015**	0.016**
	(0.009)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Het. test stat.		103.383	53.782		140.021	53.782
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		−2.251	0.302		−2.190	0.812
Spec. test p-value		0.024	0.582		0.029	0.368
Percent change		−3.221	−0.371		−3.221	−0.371

(k) Saw doctor; region interactions

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.016**	0.016**	0.016**	0.016**	0.016**	0.016**
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Het. test stat.		6.137	9.083		6.637	9.083
Het. test p-value		0.105	0.028		0.084	0.028
Spec. test stat.		0.231	1.279		0.165	1.196
Spec. test p-value		0.817	0.258		0.869	0.274
Percent change		0.053	0.261		0.053	0.261

(l) Saw doctor; year interactions

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.016**	0.016**	0.016**	0.016**	0.016**	0.016**
	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Het. test stat.		10.219	14.077		8.602	14.077
Het. test p-value		0.511	0.229		0.659	0.229
Spec. test stat.		−0.937	0.271		−0.927	0.424
Spec. test p-value		0.349	0.603		0.354	0.515
Percent change		−0.667	0.805		−0.667	0.805

Table 9: Karlan and Zinman (2008)

(a) Risk interactions

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	-4.368^{***}	-7.047^{***}	-7.410^{***}	-4.368^{***}	-7.047^{***}	-7.410^{***}
	(1.093)	(1.917)	(1.883)	(1.229)	(1.880)	(1.866)
Het. test stat.		8.259	10.518		6.177	10.518
Het. test p-value		0.016	0.005		0.046	0.005
Spec. test stat.		-2.569	8.995		-2.407	7.758
Spec. test p-value		0.010	0.003		0.016	0.005
Percent change		61.323	69.652		61.323	69.652

(b) Wave interactions

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	-4.368^{***}	-4.319^{***}	-4.377^{***}	-4.368^{***}	-4.319^{***}	-4.377^{***}
	(1.093)	(1.084)	(1.091)	(1.229)	(1.026)	(1.025)
Het. test stat.		2.215	2.905		1.156	2.905
Het. test p-value		0.330	0.234		0.561	0.234
Spec. test stat.		0.206	0.077		0.917	0.070
Spec. test p-value		0.837	0.782		0.359	0.791
Percent change		-1.123	0.211		-1.123	0.211

Table 10: Lochner and Moretti (2004)

(a) Age (whites only)

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	-0.095^{***}	-0.127^{***}	-0.123^{***}	-0.095^{***}	-0.127^{***}	-0.123^{***}
	(0.003)	(0.002)	(0.004)	(0.001)	(0.002)	(0.002)
Het. test stat.		3161.624	988.593		5630.805	7928.575
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		15.163	20.085		43.613	45.505
Spec. test p-value		0.000	0.000		0.000	0.000
Percent change		33.60	28.99		33.60	28.99

(b) Year (whites only)

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	-0.095^{***}	-0.095^{***}	-0.095^{***}	-0.095^{***}	-0.095^{***}	-0.095^{***}
	(0.003)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)
Het. test stat.		1642.756	16.790		5386.106	20.525
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.			−1.668		−2.614	−2.637
Spec. test p-value			0.095		0.009	0.008
Percent change		−0.17	−0.17		−0.17	−0.17

(c) Age (blacks only)

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	-0.363^{***}	-0.488^{***}	-0.478^{***}	-0.363^{***}	-0.488^{***}	-0.478^{***}
	(0.014)	(0.013)	(0.018)	(0.008)	(0.010)	(0.010)
Het. test stat.		1469.355	2368.284		579.577	2919.031
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		13.292	30.734		22.176	36.600
Spec. test p-value		0.000	0.000		0.000	0.000
Percent change		34.51	31.7		34.51	31.7

(d) Year (blacks only)

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	-0.363^{***}	-0.379^{***}	-0.369^{***}	-0.363^{***}	-0.379^{***}	-0.369^{***}
	(0.014)	(0.015)	(0.015)	(0.008)	(0.008)	(0.008)
Het. test stat.		744.419	50.447		2263.016	70.371
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		5.113	5.838		8.010	8.211
Spec. test p-value		0.000	0.000		0.000	0.000
Percent change		1.82	1.76		1.82	1.76

(e) Race (all observations)

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	-0.116^{***}	-0.115^{***}	-0.114^{***}	-0.116^{***}	-0.115^{***}	-0.114^{***}
	(0.003)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)
Het. test stat.		1430.965	26.776		7303.985	41.131
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		−4.593	−16.541		−9.098	−26.887
Spec. test p-value		0.000	0.000		0.000	0.000
Percent change		−0.70	−1.63		−0.70	−1.63

Table 11: Meghir and Palme (2005)

(a) Female interaction

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.014	0.014	0.014	0.014***	0.014***	0.014***
	(0.009)	(0.009)	(0.009)	(0.004)	(0.004)	(0.004)
Het. test stat		0.439	0.963		2.755	2.884
Het. test p-value		0.508	0.326		0.097	0.089
Spec. test stat.		−0.332	−0.648		−1.152	−1.620
Spec. test p-value		0.74	0.517		0.249	0.105
Percent change		0.26	0.28		0.26	0.28

(b) Year interactions

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.014	0.014	0.014	0.014***	0.014***	0.014***
	(0.009)	(0.009)	(0.009)	(0.004)	(0.004)	(0.004)
Het. test stat		41.952	60.048		29.844	30.726
Het. test p-value		0.000	0.000		0.002	0.001
Spec. test stat.		−2.470	−0.933		−1.083	−0.933
Spec. test p-value		0.014	0.351		0.279	0.351
Percent change		0.52	0.10		0.52	0.10

(c) High father's education interaction

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.014	0.017**	0.016*	0.014***	0.017***	0.016***
	(0.009)	(0.008)	(0.009)	(0.004)	(0.004)	(0.004)
Het. test stat		46.318	61.023		148.436	156.900
Het. test p-value		0.000	0.000		0.000	0.000
Spec. test stat.		−1.163	−4.455		−9.489	−9.256
Spec. test p-value		0.245	0.000		0.000	0.000
Percent change		18.51	15.95		18.51	15.95

Table 12: Oreopoulos (2006)

(a) Age interaction (Great Britain)

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.075***		0.077***	0.075***	0.076***	0.077***
	(0.002)		(0.002)	(0.001)	(0.001)	(0.001)
Het. test stat			32.831		42.601	33.740
Het. test p-value			0.242		0.038	0.210
Spec. test stat.			7.480		2.851	21.853
Spec. test p-value			0.006		0.004	0.000
Percent change			1.794		1.206	1.794

(b) Age interaction (Northern Ireland)

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.106***		0.107***	0.106***	0.107***	0.107***
	(0.004)		(0.004)	(0.002)	(0.003)	(0.003)
Het. test stat			25.661		61.217	25.661
Het. test p-value			0.592		0.000	0.592
Spec. test stat.			1.192		0.574	1.518
Spec. test p-value			0.275		0.566	0.218
Percent change			0.760		0.500	0.760

(c) Age interaction (G.B. and N.I.)

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.078***		0.079***	0.078***	0.079***	0.079***
	(0.002)		(0.002)	(0.001)	(0.001)	(0.001)
Het. test stat			51.023		43.709	51.023
Het. test p-value			0.005		0.030	0.005
Spec. test stat.			3.753		1.887	14.200
Spec. test p-value			0.053		0.059	0.000
Percent change			1.222		0.668	1.222

(d) N. Ireland dummy interaction (G.B. and N.I.)

	Clustering			No clustering		
	FE	IWE	RWE	FE	IWE	RWE
	0.078***		0.079***	0.078***	0.079***	0.079***
	(0.002)		(0.002)	(0.001)	(0.001)	(0.001)
Het. test stat			43.723		91.327	43.723
Het. test p-value			0.000		0.000	0.000
Spec. test stat.			11.004		4.831	109.906
Spec. test p-value			0.001		0.000	0.000
Percent change			0.753		0.712	0.753

Table 13: Pérez-González (2006)

(a) Operating returns on assets (OROA), year interactions

	FE	IWE	RWE
	-0.027^{***}	-0.027^{***}	-0.025^{**}
	(0.010)	(0.009)	(0.010)
Het. test stat.		34.878	25.540
Het. test p-value		0.010	0.111
Spec. test stat.		0.217	0.474
Spec. test p-value		0.829	0.491
Percent change		-2.372	-7.464

(b) Market-to-book ratio (M-B), year interactions

	FE	IWE	RWE
	-0.256^{***}	-0.226^{***}	-0.227^{***}
	(0.089)	(0.083)	(0.087)
Het. test stat.		39.777	24.390
Het. test p-value		0.002	0.143
Spec. test stat.		0.978	0.963
Spec. test p-value		0.329	0.327
Percent change		-11.448	-11.278

(c) Operating returns on assets (OROA), high family ownership interaction

	FE	IWE	RWE
	-0.027^{***}	-0.030^{***}	-0.030^{***}
	(0.010)	(0.009)	(0.008)
Het. test stat.		0.492	0.642
Het. test p-value		0.483	0.423
Spec. test stat.		-0.693	0.449
Spec. test p-value		0.489	0.503
Percent change		10.368	9.390

(d) Market-to-book ratio (M-B), High family ownership interaction

	FE	IWE	RWE
	-0.256^{***}	-0.302^{***}	-0.279^{***}
	(0.089)	(0.079)	(0.077)
Het. test stat.		1.482	2.238
Het. test p-value		0.223	0.135
Spec. test stat.		-1.171	0.435
Spec. test p-value		0.243	0.510
Percent change		18.040	9.160