Broken or Fixed Effects?*

Charles E. Gibbons† Juan Carlos Suárez Serrato‡ Michael B. Urbancic§

July 15, 2014

Abstract

This paper provides empirical evidence of an established theoretical result: in the presence of heterogeneous treatment effects, OLS is generally not a consistent estimator of the sample-weighted average treatment effect (SWE). We propose two alternative estimators that do recover the SWE in the presence of group-specific heterogeneity. We derive tests to detect the presence of heterogeneous treatment effects and to distinguish between the OLS and SWE. We document that heterogeneous treatment effects are common and the SWE is often statistically and economically different from the OLS estimate by extending eight influential papers. In all but one paper, there is statistically significant treatment effect heterogeneity; in five, the SWE is statistically different from the OLS estimator; and in five, the SWE and OLS estimators are economically different.

*We are grateful for comments from Michael Anderson, Alan Auerbach, Rodney Andrews, Joshua Angrist, Marianne Bitler, Henning Bohn, Moshe Buchinsky, Colin Cameron, Carlos Dobkin, Maximilian Kasy, Patrick Kline, Yolanda Kodrzycki, Trevon Logan, Fernando Lozano, 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 Brattle Group. Corresponding author; charlie.gibbons@brattle.com

‡Stanford Institute for Economic Policy Research, Stanford University and Department of Economics, Duke University

§Department of Economics, University of Oregon

1 Introduction

Fixed effects are a common means to "control for" unobservable differences among observations based upon observable features; 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 between groups conditional upon covariates, 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.

An established result is that fixed effects regressions average the group-specific slopes proportional to both the conditional variance of treatment and the proportion of the sample in each group, an average that generally does not coincide with the sample-weighted average.[1] 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 this difference can be large.

In this paper, we begin by deriving the OLS estimator under heterogeneous treatment effects and provide an interpretation as a weighted average of group-specific effects. This average is generally different from the sample-weighted average, a common parameter of interest. We propose two alternative estimators that are able to consistently estimate the sample-weighted average treatment effect under group-specific heterogeneity. We derive Wald and score tests to indicate the presence of heterogeneous treatment effects as well as specification tests that consider whether the estimates arising from OLS and our proposed alternatives are statistically distinguishable. These theoretical results are based upon well-known OLS and testing results.

Our main contribution is empirically judging the importance of the distinction between OLS and sample-weighted estimates. To do this, we replicate eight influential papers from the *American Economic Review* published between 2004 and 2009.[2] We choose these papers to represent high quality, respected empirical work and our replication results should be seen as an extension, not a critique, of these findings. The goal is to use this sample as an indication of the pervasiveness of

[1] See, *e.g.*, Angrist and Krueger (1999); Wooldridge (2005*a*); Angrist and Pischke (2009). The sample-weighted average is the average of each group's partial effect weighted by its frequency in the sample.

[2] Thanks to a policy decision by the editorial board of the *AER*, it is possible to access the data and programs used in recently published articles and to replicate the results of these studies. We only analyze the data that the authors provide openly 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 to focus on the weighting scheme applied by this standard procedure.

our findings.[3]

Using our theoretical derivations, we show empirically that heterogeneous treatment effects are common and that the OLS and sample-weighted estimates 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 OLS 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%. Even for a randomized experiment that we consider, we find the difference between the OLS estimate and the SWE exceeds 60%. Further, though a trade-off may be expected between bias and variance, we find that the variances of our estimators are not larger and are often smaller than those of OLS estimates.

Our paper begins in Section 2 by outlining the interplay of fixed effects and group-specific effects in the literature. In Section 3, we precisely define the parameter of interest in the presence of group-specific heterogeneity and show that OLS models are inconsistent estimators for the sample-weighted average except in special cases. We propose two alternative estimators that do recover this parameter. We derive relevant tests in Section 4. To illustrate these results through an empirical example, in Section 5, we use a simplified model from Karlan and Zinman (2008) to compare the weighting scheme from the OLS model to a sample-weighted approach and study the implications for the final estimate. We demonstrate the generality of these points in Section 6 in which we replicate a total of eight influential papers. We conclude in Section 7 by offering guidance to the applied researcher.

2 Empirical studies of heterogeneous treatment effects

In the presence of heterogeneous treatment effects across groups in the sample, the OLS estimator 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, demonstrating that the effects of a dichotomous

[3]See Murphy and Topel (1985), Gentskow and Shapiro (2013), and Oster (2014) for other examples of papers that replicate published studies to elucidate a methodological point.

treatment are averaged using different weights in each procedure.[4] Closest to our derivation below, Wooldridge (2005*a*) finds sufficient conditions for OLS models to produce sample-weighted averages in correlated random coefficient models. Our analysis builds upon this derivation for the case of fixed coefficients and offers a different interpretation of the necessary conditions for this result. Additionally, while these papers provide a strong theoretical reason to believe that OLS estimators do not provide sample-weighted estimates, we illustrate the empirical importance of this distinction using a broad array of microeconometric questions.

There has long been an interest in coefficient heterogeneity across cross-sectional groups. A notable early piece is Chow (1960). Here, he runs regressions separately by group, which is the most flexible way of permitting heterogeneity across groups for a given model, and compares the predictive power of the separate regressions to that of the pooled regression, forming a test for differences in slopes and intercepts. In our paper, we concentrate on whether the treatment effect alone varies across groups and our tests and estimators are derived in that spirit. Our approaches are less flexible, but more parsimonious and focused.

Many 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 interacted model or weighted approaches that we propose. Notable exceptions include Heckman and Hotz (1989), who consider the specific case of individual-specific time trends, which they call the random growth rate model. Papke (1994) and Friedberg (1998) also use the random growth model and find that the results of their studies are greatly influenced by trends that vary across geographic districts. These examples, however, use interactions on predictors to avert omitted variables bias or to improve the fit of their models.[5]

In contrast to these works, the message of our analysis is that models that do not account for heterogeneous effects may provide inconsistent estimates of average effects. This point has recently been made in Solon, Haider and Wooldridge (2013), who note that weighted least squares can be used to recover the average partial effect in the presence of unmodeled treatment effect heterogeneity. We build upon their discussion by deriving the necessary weights and provide applications to illustrate empirically the importance of the difference between weighted and OLS estimates.

[4] See also Angrist and Pischke (2009).

[5] In a different approach, Lochner and Moretti (2011) consider non-linearities in treatment effects, but do not estimate heterogeneous treatment effects across groups as we do here.

We extend this literature in three ways. First, while Wooldridge (2005*a*) gives sufficient conditions for a fixed effects model to deliver the sample-weighted treatment effect, we offer an alternative exposition and describe the estimate that is given by an OLS model when these conditions fail. We focus on treatment effect heterogeneity and illustrate how it can be characterized and incorporated into a model in two parsimonious ways. Next, we derive tests that can be used to consider the presence of treatment effect heterogeneity and a test that distinguishes between sample-weighted estimates and OLS estimates. Our most important contribution is to show that these estimators can be substantially different in empirical applications.

A limitation of this approach, however, is that it only considers heterogeneity that can be quantified according to observable group membership. In a different approach, a recent strand of literature has focused on characterizing and interpreting heterogeneous effects using quantile regression techniques (*e.g.*, Bitler, Gelbach and Hoynes, 2006). Examples in this literature suggest that allowing for group-specific treatment effects (such as in our approach) might not sufficiently characterize within-group heterogeneity that arises from unobservable margins (Bitler, Gelbach and Hoynes, 2014).

3 Estimation in the presence of heterogeneous treatment effects

In this section, we consider a specific model of heterogeneous treatment effects. We show that OLS provides a weighted average of these effects, where the weights depend upon the within-group conditional variance of treatment. We contrast this average to an average weighted by group size, the sample-weighted average. We offer two estimators that are able to recover the sample-weighted average, one that uses a weighting approach and another that explicitly models the heterogeneity. We provide Stata and `R` packages to perform estimation and, as we shall see in Section 4, testing. We conclude this section by contrasting our proposed approaches.

3.1 The sample-weighted effect (SWE)

Underlying our discussion is a basic linear model with treatment effects β_g that vary across G groups:

$$\begin{aligned} y_i &= \alpha_g + \mathbf{w}_i\boldsymbol{\gamma} + x_i\beta_g + \nu_i \\ &= \alpha + (\alpha_g - \alpha)\mathbb{I}_g + \mathbf{w}_i\boldsymbol{\gamma} + x_i\beta_g + \nu_i. \end{aligned} \tag{1}$$

In this model, y_i is an outcome for observation i, which is a member of group g, x_i is a treatment variable, $\mathbb{I}_g$ is a vector of group fixed effects, and $\mathbf{w}_i$ is a vector of additional covariates.[6] The variation in the level of y_i across groups is captured by fixed effects; to foreshadow, the interacted approach that we propose below will interact these variables with treatment to model treatment effect heterogeneity as well.

Though it may be instructive to model and examine the heterogeneity in treatment effects across groups, researchers often want a single summary of the effect. A natural candidate would be the sample-weighted effect, as explored in Wooldridge (2005*b*).

Definition 1 (Sample-weighted treatment effect (SWE))**.** *The sample-weighted treatment effect for the model in Equation 1 is*

$$\beta_{SWE} = \sum_g \widehat{\Pr}(g)\beta_g,$$

where $\widehat{\Pr}(g) = \frac{N_g}{N}$*, N is the total number of observations in the sample and N_g is the number of observations belonging to fixed effect group $g \in 1, \ldots, G$.*

3.2 The OLS estimator

Suppose that OLS is used to estimate the model

$$\begin{aligned} \mathbb{E}[y_i \mid w_i, x_i] &= \alpha + (\alpha_g - \alpha)\mathbb{I}_g + \mathbf{w}_i\boldsymbol{\gamma} + x_i\beta_{OLS} \\ &= \mathbf{A}\boldsymbol{\theta}_{OLS} + \mathbf{x}\beta_{OLS}, \end{aligned} \tag{2}$$

[6] Though there are G groups, there are $G-1$ fixed effects included in the model for identification purposes. Assume that group G is the excluded group.

where $\mathbf{A}$ contains the fixed effects and covariates other than treatment. Clearly $\hat{\beta}_{OLS}$ will be an average of the β_g. We are left to determine the weights used in this averaging.

Following the Frisch-Waugh-Lovell theorem, we can find the coefficient estimate $\hat{\beta}_{OLS}$ by multiplying both sides of this expression by the annihilator matrix $\mathbf{M}_A = \mathbf{I} - (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$, giving

$$\mathbf{M}_A\mathbf{y} = \mathbf{M}_A\mathbf{x}\beta_{OLS}, \tag{3}$$

which leads to the result that

$$\hat{\beta}_{OLS} = \left(\mathbf{x}'\mathbf{M}_A\mathbf{x}\right)^{-1}\mathbf{x}'\mathbf{M}_A\mathbf{y} = \frac{\widehat{\mathrm{Cov}}\left(\tilde{x}_i, y\right)}{\widehat{\mathrm{Var}}\left(\tilde{x}_i\right)},$$

where $\tilde{x}_i$ is the projected value of treatment for observation i. It can be shown (see Appendix A.1) that

$$\hat{\beta}_{OLS} = \sum_{g \in G} \widehat{\Pr}(g)\hat{\beta}_g \left(\frac{\widehat{\mathrm{Var}}\left(\tilde{x}_i \mid g\right)}{\sum_{g'=1}^{G} \widehat{\Pr}(g')\widehat{\mathrm{Var}}\left(\tilde{x}_i \mid g'\right)} \right). \tag{4}$$

From Equation 4, we see that, while OLS models do provide a weighted average of group effects, these effects are generally not weighted by sample frequencies. Instead, these weights depend upon sample variances, thereby producing estimates that are less readily interpretable.

The OLS and SWE estimators are the same when the treatment effects are homogeneous or the variance of the projected treatment is the same across all groups. Otherwise, the OLS estimator overweights groups that have larger variances of treatment conditional upon other covariates and underweights groups with smaller conditional variances.

Proposition 1 (Sufficient condition for consistent estimation of sample-weighted treatment effects by OLS)**.** *OLS consistently estimates the sample-weighted average in the presence of heterogeneous treatment effects if the variance of treatment conditional on all other covariates is the same across all groups;* i.e., $\widehat{\mathrm{Var}}\left(\tilde{x}_i \mid g\right) = \widehat{\mathrm{Var}}\left(\tilde{x}_i\right) \; \forall \, g$. *(see Appendix A.1).*

For example, a regression on data from a perfectly randomized experiment where treatment has the same variance across groups yields the sample-weighted treatment effect. Such perfection is likely unattainable in observational or experimental settings, however. Indeed, in Section 5, we replicate a randomized experiment from Karlan and Zinman (2008) as a case study. In that

experiment, treatment (an interest rate on a microloan in South Africa) is randomized within different fixed effects groups (the risk category of the borrower), but the variances of the (multi-valued) treatment are not the same across groups. In this case, we find that the sample-weighted treatment effect differs from the OLS estimate by 61%.

3.3 The regression-weighted estimator (RWE)

The result in Equation 4 hints that observations could be weighted in such a way to undo the OLS weighting and achieve sample frequency weighting. Begin from Equation 3 and multiply both sides by a diagonal weighting matrix with elements

$$d_i = \left[\widehat{\text{Var}}\left(\tilde{x}_i \mid g\right)\right]^{-1/2}. \tag{5}$$

This weighted least squares estimator becomes

$$\begin{aligned}
\hat{\beta}_{RWE} &= \frac{\widehat{\text{Cov}}\left(d_i\tilde{x}_i, d_i y\right)}{\widehat{\text{Var}}\left(d_i\tilde{x}_i\right)} \\
&= \sum_{g=1}^{G} \widehat{\Pr}(g)\hat{\beta}_g \left(\frac{\widehat{\text{Var}}\left(d_i\tilde{x}_i \mid g\right)}{\sum_{g'=1}^{G} \widehat{\Pr}(g')\widehat{\text{Var}}\left(d_i\tilde{x}_i \mid g'\right)} \right) \\
&= \sum_{g=1}^{G} \widehat{\Pr}(g)\hat{\beta}_g \left(\frac{d_i^2\widehat{\text{Var}}\left(\tilde{x}_i \mid g\right)}{\sum_{g'=1}^{G} \widehat{\Pr}(g')d_i^2\widehat{\text{Var}}\left(\tilde{x}_i \mid g'\right)} \right) \\
&= \sum_{g=1}^{G} \widehat{\Pr}(g)\hat{\beta}_g,
\end{aligned}$$

an estimator for the SWE. We call this WLS-based estimator the *regression-weighted estimator (RWE)*.

Definition 2 (Regression-weighted estimator (RWE))**.** *The regression-weighted estimator (RWE) is found by the following procedure:*

1. *Calculate the annihilator matrix* $\mathbf{M}_A$.
2. *Calculate* $\tilde{\mathbf{x}} = \mathbf{M}_A\mathbf{x}$ *and* $\tilde{\mathbf{y}} = \mathbf{M}_A\mathbf{y}$.
3. *Find the weights* d_i *for each observation according to Equation 5.*

4. Perform weighted least squares for a regression of $\tilde{y}_i$ on $\tilde{x}_i$ using weights d_i.

This gives the estimator

$$\hat{\beta}_{RWE} = \frac{\sum d_i^2 \tilde{x}_i \tilde{y}_i}{\sum d_i^2 \tilde{x}_i^2}.$$

The variance of the RWE is

$$\text{Var}\left(\hat{\beta}_{RWE} \middle| X\right) = \frac{1}{\left(\sum d_i^2 \tilde{x}_i^2\right)^2} \text{Var}\left(\sum d_i^2 \tilde{x}_i \tilde{y}_i\right).$$

We have

$$\text{Var}\left(\sum d_i^2 \tilde{x}_i \tilde{y}_i \middle| X\right) = \begin{cases} \hat{\sigma}^2 \sum_i d_i^4 \tilde{x}_i^2 & \text{under homoskedasticity} \\ \sum_i d_i^4 \tilde{x}_i^2 e_i^2 & \text{for a heteroskedasticity-robust estimator} \\ \sum_{c \in C} \sum_{j \in N_c} \sum_{i \in N_c} \left[d_j^2 d_i^2 \tilde{x}_j \tilde{x}_i e_j e_i \right] & \text{for a cluster-robust estimator} \end{cases}$$

Note that we calculate

$$\hat{\sigma}^2 = \frac{1}{N-K} \sum_i d_i^2 e_i^2,$$

where K is the number of variables in the model, including those that were annihilated.

3.4 The interaction-weighted estimator (IWE)

Fixed effects are used to model the heterogeneity in the level of y directly. Analogously, interactions between fixed effects and the treatment x can be used to model heterogeneous treatment effects. Adding these interactions to the model gives

$$\begin{aligned} y_i &= \alpha + (\alpha_g - \alpha)\mathbb{I}_g + \mathbf{w}_i \boldsymbol{\gamma} + x_i \beta + x_i \mathbb{I}_g (\beta_g - \beta) + \nu_i \\ \mathbf{y} &= \mathbf{Z}\boldsymbol{\theta}_{INT} + \boldsymbol{\nu}. \end{aligned} \tag{6}$$

We estimate the SWE by averaging the interaction terms, weighted by the sample frequency of the group. This gives the *interaction-weighted effect (IWE).*

Definition 3 (Interaction-weighted effect (IWE) estimator)**.** *The interaction-weighted estimator*

(IWE) from an interacted model following Equation 6 is

$$\hat{\beta}_{IWE} = \hat{\beta} + \sum_{g=1}^{G-1} \widehat{\Pr}(g)\hat{\beta}_g.$$

Suppose that there are K covariates that do not involve treatment.[7] Then, the variance of the IWE is

$$\text{Var}\left(\beta_{IWE}\right) = \mathbf{f}'\text{Var}\left(\hat{\boldsymbol{\theta}}_{INT}\right)\mathbf{f},$$

where the frequency vector

$$\mathbf{f} = \frac{1}{N}\begin{bmatrix} \mathbf{0}_K & N & N_1 & \dots & N_{G-1} \end{bmatrix}'$$

and $\text{Var}\left(\hat{\boldsymbol{\theta}}_{INT}\right)$ can be estimated robustly using standard techniques.

3.5 A comparison of the approaches

We have shown that, in general, OLS does not estimate the SWE. We have proposed two alternatives: the regression-weighted and the interaction-weighted estimators. Both of these approaches do consistently estimate the SWE under the DGP of Equation 1. Each has advantages.

The IWE models 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.

In the presence of heterogeneous treatment effects, the IWE can reduce standard errors by modeling these effects directly. Although, if adding these effects induces correlation between the covariates and the error term, bias can arise.

The weighting scheme employed by OLS yields a more efficient estimator in the absence of heterogeneous treatment effects. This suggests that OLS may be more efficient if the heterogeneity is relatively unimportant. As we have shown, however, OLS is a generally inconsistent estimator of

[7] That is, there is an intercept term, $G-1$ fixed effects, and $K-G$ covariates in the $\mathbf{w}_i$ vector.

the SWE. This presents a classic bias-variance trade-off. We examine the empirical extent of this trade-off in Section 6.2 and find little empirical evidence to suggest that the SWE estimators are generally less efficient than OLS.

4 Testing for heterogeneous treatment effects

Armed with two estimators of the SWE, we next consider testing.[8] First, we derive tests for the presence of heterogeneous treatment effects using both a Wald test and a score test. Then, we offer a specification test for equality between the SWE estimator and the OLS estimator.

4.1 Wald test for modeled heterogeneity

If the IWE is estimated following Equation 6, 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 2 (Wald test for modeled heterogeneity)**.** *The Wald test statistic for heterogeneous treatment effects is calculated according to*

$$T_W = \mathbf{p}' \text{Var}\left(\hat{\theta}_{INT}\right) \mathbf{p},$$

where the $(K+G) \times (G-1)$ *matrix*

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

and $\text{Var}\left(\hat{\boldsymbol{\theta}}_{INT}\right)$ *can be estimated robustly using standard approaches. Asymptotically, this test statistic has a* χ^2_{G-1} *distribution under the null hypothesis.*

4.2 Score test for unestimated 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

[8] The fixed effects that we consider denote group membership and the sizes of these groups grow with overall sample size. 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.

IWE can be tested.

This procedure begins by estimating the standard OLS model of Equation 2 and obtaining the residuals for each observation e_i. Next, the score is calculated according to

$$\mathbf{s}\left(\mathbf{z}_i', \hat{\boldsymbol{\theta}}_{OLS}\right) = e_i \mathbf{z}_i',$$

where $\mathbf{z}_i$ is the row of covariates for observation i from the $\mathbf{Z}_{INT}$ interacted matrix defined implicitly in Equation 6.

Proposition 3 (Score test for unestimated heterogeneity)**.** *A score test statistic for the presence of heterogeneous treatment effects has the form*[9]

$$T_S = N \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{z}_i', \hat{\boldsymbol{\theta}}_{OLS}\right) \right)' \mathbf{S}_0^{-1} \mathbf{C}' \left(\mathbf{C} \mathbf{S}_0^{-1} \mathbf{C}' \right)^{-1} \mathbf{C} \mathbf{S}_0^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{z}_i', \hat{\boldsymbol{\theta}}_{OLS}\right) \right),$$

where[10]

$$\mathbf{S}_0 = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{z}_i', \hat{\boldsymbol{\theta}}_{OLS}\right) \mathbf{s}\left(\mathbf{z}_i', \hat{\boldsymbol{\theta}}_{OLS}\right)'$$

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*

$$\mathbf{S}_0 = \sum_{c=1}^{C} \sum_{j=1}^{N_c} \sum_{i=1}^{N_c} \mathbf{s}\left(\mathbf{z}_i', \hat{\boldsymbol{\theta}}_{OLS}\right) \mathbf{s}\left(\mathbf{z}_j', \hat{\boldsymbol{\theta}}_{OLS}\right)'.$$

Like the Wald test above, this test statistic has an asymptotic χ^2_{G-1} distribution under the null hypothesis. This test may outperform the Wald test when a clustered variance-covariance matrix is used (Kline and Santos, 2012).

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

[10] $\mathbf{S}_0$ is analogous to $\mathbf{S}$ in Appendix A.2, but the 0 subscript here indicates that it is calculated under the null hypothesis.

4.3 Specification test of equality between the SWE and OLS estimates

Even if heterogeneous treatment effects are present, the SWE and OLS estimators may be equal or at least statistically indistinguishable. In this subsection, we derive a test that is able to distinguish between the two estimators. 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 $\hat{\beta}_{SWE}$.

Proposition 4 (Specification test of the differences between the OLS and SWE estimates)**.** *The test of the following null hypothesis*

$$H_0 : \text{plim}\left(\hat{\beta}_{SWE} - \hat{\beta}_{OLS}\right) = 0$$
$$H_a : \text{plim}\left(\hat{\beta}_{SWE} - \hat{\beta}_{OLS}\right) \neq 0$$

can be conducted by noting that the Wald test statistic

$$T_E = \left(\hat{\beta}_{SWE} - \hat{\beta}_{OLS}\right)' \left(N^{-1}\widehat{\text{Var}}\left[\hat{\beta}_{SWE} - \hat{\beta}_{OLS}\right]\right)^{-1} \left(\hat{\beta}_{SWE} - \hat{\beta}_{OLS}\right)$$

has an asymptotic $\chi^2(1)$ *distribution under* H_0*. Robust estimation of the variance-covariance matrix of the difference is addressed in Appendix A.2. This test is implemented by Stata commands and an* `R` *package available from the authors, as discussed in Appendix B.*

5 A case study: Karlan and Zinman (2008)

In this section, we provide a detailed case study of one of our selected *AER* papers. This example illuminates the exposition of Section 3.2 and further clarifies the relationship between the OLS and SWE estimates.

In Section 3.2, we show that, even if an experiment randomizes such that treatment is independent of any other covariates, the OLS estimator might not be a consistent estimator of the sample-weighted average. More specifically, a sufficient condition is that all covariates are precisely uncorrelated with treatment within each group *and* the variance of treatment is the same across all groups (see Equation 4). Among our *AER* replications, we have one experiment that we might consider more closely. Karlan and Zinman (2008) randomize the interest rate offered for a microloan

across a population of South Africans and estimate the credit elasticity.

In the case of Karlan and Zinman (2008), the authors include two sets of covariates other than the treatment: the "pre-approved risk category" of the borrower (low, medium, or high) and the borrower's mailer wave. The distributions of treatment and risk level are nearly uncorrelated with the mailer wave, hence, we ignore these fixed effects in this section for expository purposes (though we do include them in the replications of Section 6). But, to offer interest rates commensurate with prevailing market rates, the authors need to charge higher rates to higher risk individuals. Recall that differing means in treatment do not drive the difference between the OLS and SWE estimates, but rather differences in variances.

The authors offer not only higher rates to higher risk borrowers, but also offer a greater range of rates to this group and, as a result, the variance of treatment differs across the groups. As a result, the OLS estimate will not be equal to the SWE if the responsiveness to interest rates varies across risk groups.

The OLS 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 from an interacted model, given in column 4 of Table 1, we can calculate the OLS estimate; this estimate is given in the bottom row of the table in the "OLS weight" column.

Now compare the weights from the OLS model to the sample frequencies used to calculate a SWE (these weights appear in column 3 of Table 1). Note that high risk individuals are overweighted in the OLS model and the low and medium risk individuals are underweighted. This accords with the design of the study—high risk borrowers had a wider range of interest rate offers and this relatively high variance in treatment leads to overweighing in the OLS estimate.

Differences in weighting scheme are only important if the treatment effect is heterogeneous. We find that high-risk borrowers are much less responsive to the interest rate than low-risk borrowers. Because high-risk individuals are overweighted and have a smaller (in absolute value) treatment effect, the OLS estimate underestimates the sample-weighted responsiveness of individuals to the interest rate by nearly 70%.[11]

[11]The estimate that we calculate is not precisely equal to the OLS estimate given in the paper. This is because we do not include mailer wave fixed effects in this exposition, explaining the difference between the 70% difference here and the 61% difference reported in later summary tables.

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

Risk group	OLS 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 corresponds to the IWE estimator. Note that the OLS estimate here, -4.40, does not precisely equal the OLS 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 (*i.e.*, treatment). Subsequent replication results in our paper do recover the actual values reported in the replicated papers, including this one, unless otherwise noted.

6 Comparing OLS and SWE estimates: an *AER* investigation

We have seen that OLS models generally do not provide the sample-weighted estimate in the presence of heterogeneous treatment effects, even in randomized experiments. To produce the SWE, we offer two approaches: the RWE and the IWE. To consider the empirical relevance of the distinction between the OLS and SWE estimators, we turn to highly-cited papers published in the *American Economic Review* between 2004 and 2009. We choose this publication due to its influence and the quality of its papers and limit our analysis to recent years in order to capitalize upon the *AER* editorial board's decision to require posting of data and other replication details to the EconLit online repository. The papers that we choose are well known in their respective fields and 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, which 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 here, 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 C.1.[12]

[12]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.

6.1 Replication results

To consider whether the difference between the OLS and SWE estimators is empirically important, we first test for heterogeneous treatment effects, then test for a difference between the OLS and SWE estimates. When testing for the significance of heterogeneity, we provide results using both the Wald and score tests. We also provide estimates and tests of equality with the OLS estimator for both the RWE and IWE estimators. We develop a Stata command and `R` package to perform these analyses.[13]

Our results are summarized in Table 3. For each paper, we list the fixed effects groups that we consider as potential sources of treatment effect heterogeneity along with a test indicating the presence of heterogeneity, a specification test comparing the SWE and OLS estimates, and the percent difference in these two estimates. In the final column, we indicate whether the author considers treatment effect heterogeneity for these fixed effects groups. These statistics are all for the RWE estimator. We compute standard errors following the level of clustering used by the original authors.[14] The results for the IWE estimator are generally very similar, as we would expect, and these results are included in the detailed tables of Appendix C.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 OLS 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 OLS estimates.

[13] See Appendix B. The authors have posted these resources online for researchers interested in implementing these tests.

[14] In Appendix C.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.

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 C.

Table 3: *AER* replication results

(1) Citation	(2) Fixed effect	(3) Joint test (p-value)	(4) Diff. test (p-value)	(5) Percent diff.	(6) In paper
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 sample-weighted estimate and the OLS estimate. Column 5 gives the percent difference between these two estimates; we report the percent difference because it is easier for readers not familiar with a particular paper to understand the implications of our results. Raw differences in the estimates are reported in Appendix C.2. 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.

6.2 The interacted and OLS models and the bias-variance tradeoff

Our implementation of the IWE incorporates group-specific treatment effects into a standard fixed effects regression. The choice between the standard OLS model and the interacted version, then, can be viewed as the choice between short and long versions of a regression. To this point, we have focused on the bias of OLS estimators relative to the SWE in a world of treatment effect heterogeneity. Of course, we are concerned with the variance of our estimators as well.

Suppose that the variance of the OLS estimator is lower than that of the IWE. Goldberger (1991) provides rationales choosing for a short, potentially biased, regression over a long regression that has higher variance following a bias-variance tradeoff framework. We consider these rationales in the context of the OLS and interacted models using the empirical evidence found in our meta-study. The rationales are:

- The researcher believes that the treatment effects are homogeneous and thus the coefficients on the interactions are expected to be zero. Fortunately, this is assumption can be tested using a joint significance test of the coefficients on the interaction variables. The interactions are significant in the vast majority of the cases that we consider, rendering this an inappropriate general justification for choosing the OLS estimator.

- The researcher believes that treatment effects may be heterogeneous, but would accept an imperfect approximation, $\hat{\beta}_{OLS}$, with smaller standard errors. This choice depends upon the magnitude of the difference between the estimators in terms of both bias and variance.

To evaluate the bias-variance tradeoff in our replications, we can examine the relationship between the largest absolute difference between the OLS and RWE estimates for each paper and compare that to the percent difference in standard errors of the treatment effect between the two estimators; Figure 1 shows this relationship.[15] See that the SWE estimator exhibits standard errors that are less than ten percent larger than those for the OLS estimator in six of eight cases. It is perhaps not surprising that the standard errors for the Karlan and Zinman (2008) paper increase substantially given the large change in the estimate (nearly 70% for the RWE). But, the t-statistics are similar, -4.00 using OLS and -3.94 using the RWE. Overall, the results indicate that there is

[15] If the difference in the standard errors is positive, the SWE from the interacted model has a larger standard error.

not generally a strong bias-variance trade-off unless the differences between the estimates are great.

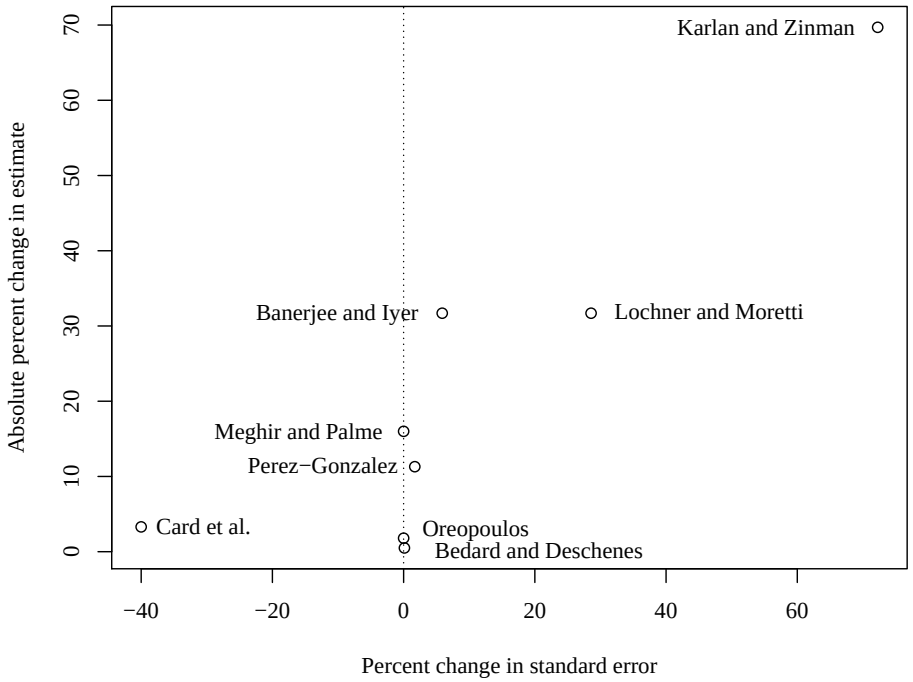

Figure 1: The relationship between the difference in the estimates and the change in variance among the *AER* replications

Notes: Figure is based on the full results presented in Appendix C.3. Figure plots estimates from the RWE and corresponding standard errors at the level of clustering used by the original authors, where applicable.

7 Conclusion

This paper assesses the empirical relevance of a well-known result: that, under treatment effect heterogeneity, OLS with group fixed effects may provide a biased estimate of the sample-weighted treatment effect. We illustrate this point and develop tests for the presence of heterogeneity and for equality between the OLS estimate and the sample-weighted effect. We also provide statistical packages to implement our proposed estimation and testing procedures.

Our replication exercise reveals that, in most cases, treatment effects are indeed heterogeneous. This fact in-and-of-itself is not surprising, as many of the original authors explore treatment effect heterogeneity across observable groups. Our main result, however, is that the underlying heterogeneity in treatment effects is often combined with unequal conditional variances in the treatment across groups, resulting in OLS estimates that are both statistically and economically

different from sample-weighted effects.

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 randomized 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).

Finally, our focus in this paper is to analyze heterogeneity in treatment effects across observable groups. Heterogeneity may instead arise along unobservable margins or within observable groups (Bitler, Gelbach and Hoynes, 2014).

We show that OLS with group fixed effects is often 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 main estimates by considering the IWE or RWE as proposed above or by analyzing the underlying 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.

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.

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. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96(4):988–1012.

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.

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.

Chow, Gregory C. 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica* 28(2):591–605.

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

Gentskow, Matthew and Jesse Shapiro. 2013. Measuring the sensitivity of parameter estimates to sample statistics. Working paper University of Chicago.

Goldberger, Arthur S. 1991. *A Course in Econometrics.* Harvard University Press.

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.

Lochner, Lance and Enrico Moretti. 2011. Estimating and Testing Non-Linear Models Using Instrumental Variables. Working Paper 17039 National Bureau of Economic Research.

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 Wooldridge. 2013. What are we Weighting For? Working Paper 18859 National Bureau of Economic Research.

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

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

Wooldridge, Jeffrey M. 2005*b*. Unobserved Heterogeneity and Estimation of Average Partial Effects. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. Donald Andrews and James Stock. Cambridge University Press.

A Derivations of results

A.1 Sufficient conditions for estimation of sample-weighted treatment effects by OLS

In this appendix, we utilize well-established results to derive the properties of the OLS estimator. Suppose that a researcher estimates a fixed-effects model

$$\begin{aligned}\mathbb{E}[y_i \mid w_i, x_i] &= \alpha + (\alpha_g - \alpha)\mathbb{I}_g + \mathbf{w}_i\boldsymbol{\gamma} + x_i\beta_{OLS} \\ &= \mathbf{A}\boldsymbol{\theta}_{OLS} + \mathbf{x}\beta_{OLS}.\end{aligned}$$

Following the Frisch-Waugh-Lovell theorem, we find the estimator $\hat{\beta}_{OLS}$ by multiplying both sides of this expression by the annihilator matrix $\mathbf{M}_A = \mathbf{I} - \mathbf{A}\left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{A}'$, giving

$$\begin{aligned}\mathbf{M}_A\mathbf{y} &= \mathbf{M}_A\mathbf{x}b + \mathbf{M}_A\mathbf{u} \\ \Rightarrow \hat{\beta}_{OLS} &= \left(\mathbf{x}'\mathbf{M}_A\mathbf{x}\right)^{-1}\mathbf{x}'\mathbf{M}_A\mathbf{y} = \frac{\widehat{\mathrm{Cov}}\left(\tilde{x}_i, y\right)}{\widehat{\mathrm{Var}}\left(\tilde{x}_i\right)},\end{aligned}$$

where $\tilde{x}_i$ is the projected value of treatment for observation i. Define the group-specific effect as

$$\hat{\beta}_g = \frac{\widehat{\mathrm{Cov}}\left(\tilde{x}_i, y \mid g\right)}{\widehat{\mathrm{Var}}\left(\tilde{x}_i \mid g\right)}.$$

We can decompose the estimate of $\hat{b}$ following

$$\begin{aligned}\hat{\beta}_{OLS} &= \frac{\widehat{\mathrm{Cov}}\left(\tilde{x}_i, y_i\right)}{\widehat{\mathrm{Var}}\left(\tilde{x}_i\right)} \\ &= \frac{\sum_{g=1}^{G}\widehat{\Pr}(g)\widehat{\mathrm{Cov}}\left(\tilde{x}_i, y_i \mid g\right)}{\widehat{\mathrm{Var}}\left(\tilde{x}_i\right)} \\ &= \frac{\sum_{g=1}^{G}\widehat{\Pr}(g)\hat{\beta}_g\widehat{\mathrm{Var}}\left(\tilde{x}_i \mid g\right)}{\widehat{\mathrm{Var}}\left(\tilde{x}_i\right)} \\ &= \sum_{g=1}^{G}\widehat{\Pr}(g)\hat{\beta}_g\left(\frac{\widehat{\mathrm{Var}}\left(\tilde{x}_i \mid g\right)}{\widehat{\mathrm{Var}}\left(\tilde{x}_i\right)}\right).\end{aligned}$$

The second equality follows because we are considering a specific type of covariate—discrete fixed effects. Note further that the within-between variance decomposition demonstrates that

$$\begin{aligned}(N-1)\widehat{\text{Var}}\left(\tilde{x}_i\right) &= \sum_{i=1}^{N}\left(\tilde{x}_i-\tilde{x}_g\right)^2+\sum_{i=1}^{N}\left(\tilde{x}_g-\bar{\tilde{x}}\right)^2 \\ &= \sum_{i=1}^{N}\left(\tilde{x}_i-\tilde{x}_g\right)^2 \\ &= \sum_{g=1}^{G}\left(N_g-1\right)\frac{1}{\left(N_g-1\right)}\sum_{i=1}^{N_g}\left(\tilde{x}_i-\tilde{x}_g\right)^2 \\ &= \sum_{g=1}^{G}\left(N_g-1\right)\widehat{\text{Var}}(\tilde{x}_i \mid g),\end{aligned}$$

where $\tilde{x}_g$ is the mean of $\tilde{x}_i$ in group g, $\bar{\tilde{x}}$ is the mean of $\tilde{x}_i$ across all groups (*i.e.*, , 0), and N_g is the number of observations in group g. The second (*i.e.*, between) term in the sum drops out because the $\tilde{x}_i$ are demeaned within each fixed effect group. Thus, we find

$$\hat{\beta}_{OLS}=\sum_{g=1}^{G}\widehat{\Pr}(g)\hat{\beta}_g\left(\frac{\widehat{\text{Var}}\left(\tilde{x}_i \mid g\right)}{\sum_{g'=1}^{G}\widehat{\Pr}(g')\frac{N_g-1}{N-1}\widehat{\text{Var}}(\tilde{x}_i \mid g')}\right).$$

It is clear that the estimate of the treatment effect arising from the fixed effects model is not simply a frequency-weighted average of the group-specific effects. This is only the case if the conditional variances of the treatment within each group are the same.

The bias of the OLS model in estimating the sample-weighted average, $\hat{\beta}_{SWE}$, has the following limit:

$$\begin{aligned}\text{plim}\left(\hat{\beta}_{SWE}-\hat{\beta}_{OLS}\right) &= \sum_g\left(\Pr(\mathbb{I}_g=1)-\frac{\Pr(\mathbb{I}_g=1)\text{Var}(\tilde{x}|\mathbb{I}_g=1)}{\mathbb{E}_G[\text{Var}(\tilde{x} \mid g)]}\right)\beta_g \\ &= \sum_g \Pr(\mathbb{I}_g=1)\left(1-\frac{\text{Var}(x|\mathbb{I}_g=1)}{\mathbb{E}_G[\text{Var}(\tilde{x} \mid g)]}\right)\beta_g.\end{aligned}$$

Again, this difference is 0 if $\text{Var}\left(\tilde{x}_i \mid g\right)=\text{Var}\left(\tilde{x}_i\right)_i \; \forall \, g$.

A.2 Specification test for the difference between the OLS and SWE estimators

Next, we turn to testing whether the SWE and OLS estimates can be statistically distinguished. We derive the distribution of the test statistic through joint estimation of the models using a Method of Moments (MM) approach. We first derive the joint distribution of the estimators, then we develop a specification test for our particular hypothesis.

Recall the models of Equations 2, 3, and 6. We form the moment conditions for the OLS estimator as

$$\mathbf{0} = \sum_{i=1}^{N} \mathbf{h}_{OLS,i}\left(\hat{\boldsymbol{\theta}}_{OLS}\right) = \sum_{i=1}^{N} \mathbf{a}_i'\left(y_i - \mathbf{a}_i'\hat{\boldsymbol{\theta}}_{OLS}\right).$$

For the SWE, they are

$$\mathbf{0} = \sum_{i=1}^{N} \mathbf{h}_{SWE,i}\left(\hat{\boldsymbol{\theta}}_{SWE}\right) = \begin{cases} \sum_{i=1}^{N} d_i^2 \tilde{x}_i\left(\tilde{y}_i - \tilde{x}_i\hat{\beta}_{RWE}\right) & \text{for the RWE} \\ \sum_{i=1}^{N} \mathbf{z}_i'\left(y_i - \mathbf{z}_i'\hat{\boldsymbol{\theta}}_{INT}\right) & \text{for the IWE.} \end{cases}$$

We stack the relevant conditions into

$$\mathbf{0} = \sum_{i=1}^{N} \mathbf{h}_i\left(\hat{\boldsymbol{\theta}}\right),$$

where $\hat{\boldsymbol{\theta}}' = \left[\hat{\boldsymbol{\theta}}_{OLS}'\ \hat{\boldsymbol{\theta}}_{SWE}'\right]$ and let $\boldsymbol{\theta}_0' = \left[\boldsymbol{\theta}_{OLS}'\ \boldsymbol{\theta}_{SWE}'\right]$. Applying standard MM arguments (see, *e.g.*, Cameron and Trivedi, 2005), it follows that $\widehat{\boldsymbol{\theta}}$ has the property that

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{G}^{-1}\mathbf{S}\mathbf{G}^{-1}\right),$$

where

$$\mathbf{G} = \text{plim}\frac{1}{N}\sum_{i=1}^{N}\left[\left.\frac{\partial \mathbf{h}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] \text{ and } \mathbf{S} = \text{plim}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\left.\mathbf{h}_i(\boldsymbol{\theta})\mathbf{h}_j'(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right].$$

Note that, by partitioning the matrix $\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix}$ and using the fact that

$$\frac{\partial \mathbf{h}_{OLS,i}(\boldsymbol{\theta}_{OLS})}{\partial \boldsymbol{\theta}_{SWE}'} = \mathbf{0} \quad \text{and} \quad \frac{\partial \mathbf{h}_{SWE,i}(\boldsymbol{\theta}_{SWE})}{\partial \boldsymbol{\theta}_{OLS}'} = \mathbf{0},$$

it follows that $\mathbf{G}_{21} = \mathbf{G}_{12} = \mathbf{0}$.

As is standard (once again, see Cameron and Trivedi, 2005), we estimate $\mathbf{G}$ via

$$\widehat{\mathbf{G}} = \frac{1}{N}\sum_{i=1}^{N}\left[\left.\frac{\partial \mathbf{h}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right].$$

To estimate $\mathbf{S}$, we consider two cases. First, assuming independence over i, an estimator robust to heteroskedasticity is

$$\widehat{\mathbf{S}}_R = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{h}_i\left(\widehat{\boldsymbol{\theta}}\right)'.$$

A second estimator that incorporates clustered errors is

$$\widehat{\mathbf{S}}_C = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j=1}^{N_c}\mathbf{h}_{ic}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{h}_{jc}\left(\widehat{\boldsymbol{\theta}}\right)'.$$

Thus, robust and clustered estimators of the variance of $\widehat{\boldsymbol{\theta}}$ are $\widehat{\text{Var}}\left[\widehat{\boldsymbol{\theta}}\right] = N^{-1}\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{S}}_e\widehat{\mathbf{G}}^{-1}$ for $e = R, C$ respectively.

Now we turn to the specific hypothesis that we would like to consider—namely that the SWE and OLS estimates are indistinguishable:

$$H_0 : \text{plim}\left(\hat{\beta}_{SWE} - \hat{\beta}_{OLS}\right) = 0$$
$$H_0 : \text{plim}\left(\hat{\beta}_{SWE} - \hat{\beta}_{OLS}\right) \neq 0.$$

The test statistic is

$$T_E = \mathbf{r}\hat{\boldsymbol{\theta}}\left(\mathbf{r}\widehat{\text{Var}}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{r}'\right)^{-1}\hat{\boldsymbol{\theta}}'\mathbf{r}',$$

where

$$\mathbf{r} = \begin{cases} \begin{bmatrix} \mathbf{0}_K & -1 & 1 \end{bmatrix} & \text{for RWE} \\ \frac{1}{N}\begin{bmatrix} \mathbf{0}_K & -N & \mathbf{0}_K & N & N_1 & \dots & N_{G-1} \end{bmatrix} & \text{for IWE.} \end{cases}$$

For Online Publication

B 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 RWE and IWE and provides the OLS result, 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 is discussed below.

B.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
```

B.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:

```
EstimateIWE(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 (see Appendix A.2) and the score test (see Section 4.2) 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 4.1) 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)
```

C *AER* replications

C.1 Paper selection

In this paper, our goal is to determine whether the difference between an estimator of the SWE and the OLS 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.[16] 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.[17]

- 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.

- 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

[16] 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.

[17] 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.

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.[18] 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.[19]

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 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.

[18] 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.

[19] See Section B.

C.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 4. 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 4: 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 × 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 × father's education × 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.

C.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 IWE 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.[20] 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.[21] Specification tests for the difference between the SWE and OLS 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.

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

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

Table 5: 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 6: Bedard and Deschenes (2006)

(a) Age interactions

	OLS	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

	OLS	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

	OLS	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

	OLS	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 7: 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 8: Karlan and Zinman (2008)

(a) Risk interactions

	Clustering			No clustering		
	OLS	IWE	RWE	OLS	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		
	OLS	IWE	RWE	OLS	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 9: Lochner and Moretti (2004)

(a) Age (whites only)

	Clustering			No clustering		
	OLS	IWE	RWE	OLS	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		
	OLS	IWE	RWE	OLS	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		
	OLS	IWE	RWE	OLS	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		
	OLS	IWE	RWE	OLS	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		
	OLS	IWE	RWE	OLS	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 10: Meghir and Palme (2005)

(a) Female interaction

	Clustering			No clustering		
	OLS	IWE	RWE	OLS	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		
	OLS	IWE	RWE	OLS	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		
	OLS	IWE	RWE	OLS	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 11: Oreopoulos (2006)

(a) Age interaction (Great Britain)

	Clustering			No clustering		
	OLS	*IWE*	*RWE*	*OLS*	*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		
	OLS	*IWE*	*RWE*	*OLS*	*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		
	OLS	*IWE*	*RWE*	*OLS*	*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		
	OLS	*IWE*	*RWE*	*OLS*	*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 12: Pérez-González (2006)

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

	OLS	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

	OLS	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

	OLS	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

	OLS	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