NBER WORKING PAPER SERIES

# AT WHAT LEVEL SHOULD ONE CLUSTER STANDARD ERRORS IN PAIRED EXPERIMENTS, AND IN STRATIFIED EXPERIMENTS WITH SMALL STRATA?

Clément de Chaisemartin
Jaime Ramirez-Cuellar

Working Paper 27609
http://www.nber.org/papers/w27609

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2020

We are very grateful to Antoine Deeb, Jake Kohlhepp, Heather Royer, Dick Startz, Doug Steigerwald, Gonzalo Vasquez-Bare, members of the econometrics and labor groups at UCSB, participants of the Advances in Field Experiments Conference 2019, California Econometrics Conference 2019, LAMES 2019, and LACAE 2019 for their helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At What Level Should One Cluster Standard Errors in Paired Experiments, and in Stratified Experiments with Small Strata?
Clément de Chaisemartin and Jaime Ramirez-Cuellar
NBER Working Paper No. 27609
July 2020
JEL No. C01,C21

**ABSTRACT**

In paired experiments, units are matched into pairs, and one unit of each pair is randomly assigned to treatment. To estimate the treatment effect, researchers often regress their outcome on a treatment indicator and pair fixed effects, clustering standard errors at the unit-of-randomization level. We show that the variance estimator in this regression may be severely downward biased: under constant treatment effect, its expectation equals 1/2 of the true variance. Instead, we show that researchers should cluster their standard errors at the pair level. Using simulations, we show that those results extend to stratified experiments with few units per strata.

Clément de Chaisemartin
Department of Economics
University of California at Santa Barbara
Santa Barbara, CA 93106
and NBER
clementdechaisemartin@ucsb.edu

Jaime Ramirez-Cuellar
University of California at Santa Barbara
Santa Barbara, CA 93106
jramirezcuellar@ucsb.edu

# 1 Introduction

In randomized experiments, the units included in the randomization, e.g. villages, are often matched into pairs, and then one unit of each pair is randomly assigned to treatment. Alternatively, units may be grouped into small strata of, say, less than ten units, and then a fixed number of units gets treated in each stratum. Paired experiments or stratified experiments with a small number of units per strata are commonly used in economics. The American Economic Journal: Applied Economics published 50 randomized field experiments from 2014 to 2018. Of those, four used a paired design, and seven used a stratified design with 10 units or less per strata. Indeed, grouping units into pairs or small strata reduces the variance of the treatment effect estimator, if the variables on which the units are grouped predict the outcome (Athey and Imbens, 2017). In our survey, we also found that to estimate the treatment effect, researchers usually regress their outcome on a treatment indicator and pair or strata fixed effects, and cluster their variance estimator at the unit-of-randomization level, namely, at the village level in our example. In this paper, we assess whether this inference method is appropriate. We do so assuming that the units participating in the experiment are a convenience sample (Abadie et al., 2020) rather than an i.i.d. sample drawn from a super population, so the only source of randomness is the assignment to the treatment. Our survey shows that this set-up is applicable to a majority of paired- and small-strata experiments conducted in economics, where experimental units are rarely drawn from a larger population.

We start by considering paired designs. The treatments of the two villages in the same pair are perfectly negatively correlated: if village A is treated, then village B must be untreated, and conversely. The pair-clustered variance estimator accounts for that correlation. Accordingly, it is unbiased for the variance of the treatment effect estimator when the effect does not vary across pairs, and conservative otherwise. On the other hand, the village-clustered variance estimator does not account for that correlation and may be biased. The direction of the bias crucially depends on whether pair fixed effects are included in the regression. When fixed effects are included, as is often the case in practice, we show that if all villages have the same number of villagers, the village-clustered variance estimator is exactly equal to a half of the pair-clustered one. Then, if the treatment effect does not vary across pairs, when the number of pairs goes to infinity the $t$-statistic using village-clustered standard errors converges to a $\mathcal{N}(0, 2)$ distribution. Accordingly, comparing that $t$-statistic to, e.g., 1.96, the critical value one would use in a 5% level test, actually yields a 16.5% type 1 error rate. In Section E of the Web Appendix, we show that this result is not very sensitive to the assumption that all villages have the same number of villagers.

When pair fixed effects are not included in the regression, which is less often the case in practice, there is no longer a fixed relationship between the village- and pair-clustered variance estimators. However, we show that the expectation of the difference between the former and the latter is proportional to the difference between the between-pair and within-pair covariance of the two potential outcomes. Both covariances should typically be positive, and one may also expect their difference to be positive. For instance, in the extreme case where the two units in the same pair have equal potential outcomes, the second covariance is equal to 0, and the village-clustered variance estimator without fixed effects is too conservative.

We apply our results to revisit the paired randomized experiments we found in our survey. 372 regressions in those papers have pair fixed effects. Using standard errors clustered at the unit-of-randomization level, the authors found a 5%-level significant treatment effect in 162 regressions. Using standard errors clustered at the pair level, we find a significant effect in 109 regressions. 54 regressions do not have pair fixed effects. Using standard errors clustered at the unit-of-randomization level, the authors found a 5%-level significant treatment effect in 31 regressions. Using standard errors clustered at the pair level, we find a significant effect in 36 regressions.

With heterogeneous treatment effects across pairs, the pair-clustered variance estimator is conservative. To increase power, one may want to use a less conservative estimator. We study two alternatives: the pair-of-pairs variance estimator proposed by Abadie and Imbens (2008), and a variance estimator proposed by Bai et al. (2019). The properties of these two estimators have not been studied yet in the finite-population set-up we consider. We show that as the pair-clustered variance estimator, those two estimators are conservative. They are less conservative than the pair-clustered estimator when the treatment effect is less heterogeneous within than across pairs of pairs, and more conservative otherwise. We estimate the three estimators in the regressions we consider in our empirical application, and find that they are on average equivalent, so it does not seem one can expect large power gains from using those estimators. Moreover, simulations based on the data from the paired experiment conducted by Crépon et al. (2015) show that $t$-tests using those two estimators have a drawback relative to the $t$-test using the pair-clustered estimator. They are approximately normally distributed only if the sample has more than a couple hundred pairs, a condition met in only one of the paired experiments in our survey. On the other hand, the $t$-test based on the pair-clustered estimator is approximately normally distributed with as few as 20 pairs.

Finally, we consider stratified experiments with a small number of units per strata. Using simulations, we show that our results for paired designs extend to that case. There as well, the treatments of units in the same stratum are negatively correlated, so this correlation should be accounted for. Our simulations show that $t$-tests based on strata-clustered standard errors have correct size, irrespective of whether strata fixed effects are included in the regression. On the other hand, $t$-tests based on randomization-unit-clustered standard errors do not have correct size. They tend to be liberal when strata fixed effects are included, conservative otherwise. When strata fixed effects are included, the liberality of those $t$-tests decreases with the number of randomization units per strata: the larger a stratum, the less correlated its units' treatments. With 5 units per strata, a 5% level $t$-test is rejected 8.0% of the time, and with 10 units per strata, it is rejected 6.6% of the time. With more than 10 units per strata, size distortions become smaller. This is why we use this 10 units per strata threshold in our survey, though we acknowledge it is somewhat arbitrary.

Our paper is related to several other papers that have considered paired experiments. Our results may seem to contradict those in Bruhn and McKenzie (2009). Using simulations, they show that in paired experiments with only one observation per randomization unit (e.g. one villager per village), $t$-tests based on fixed effects regressions with no clustering have correct size. This is due to the fact that in the one-observation-per-randomization-unit case, the fixed effects regression has half as many regressors as observations. Consequently, the degrees-of-freedom correction embedded in most statistical software amounts to multiplying the unclustered variance estimator by two, which then makes it equivalent to the pair-clustered variance estimator. Therefore, the issue we highlight does not apply to the one-observation-per-randomization-unit case. However, in our survey of paired and small-strata experiments, only one paper has one observation per randomization unit. Random assignment is clustered in all the other papers,[1] and the median number of observations per unit of randomization is large (99 for paired experiments, 26 for small-strata experiments). With clustered assignment, the degree-of-freedom adjusted standard error clustered at the unit of randomization is no longer equivalent to the pair-clustered one.

Athey and Imbens (2017) and Bai et al. (2019) have also shown that when pair fixed effects are not included in the regression, standard errors clustered at the unit-of-randomization level tend to be conservative. We show that when pair fixed effects are included, these standard errors actually become very liberal.

Imai et al. (2009) have also proposed an estimator of the treatment-effect estimator's variance

[1] This is in line with Muralidharan and Niehaus (2017), who find that assignment is clustered in 62% of the RCTs published in top 5 journals in 2001-2016.

in paired experiments, and have shown it is unbiased if the treatment effect is constant across pairs, and conservative in general. With respect to this paper, our contribution is to show that when the number of observations is the same in all villages, the pair-clustered variance estimator is equal to their estimator, up to a degrees-of-freedom correction. Thus, we justify the use of the pair-clustered variance estimator. Moreover, we present large sample results for $t$-tests based on pair-clustered variance estimators, while their paper focuses on finite-sample results.

Finally, two of the variance estimators we study were proposed by Abadie and Imbens (2008) and Bai et al. (2019). Assuming that units participating in the experiment are an i.i.d. sample drawn from a super population, both papers have shown that once properly normalized, their proposed estimator is consistent for the asymptotic variance of the treatment effect estimator. Our paper is the first to show that those estimators are actually conservative in a finite-population set-up, and it is also the first to compare them to the pair-clustered variance estimator in a wide range of empirical applications.

The paper is organized as follows. Section 2 presents our survey of paired and small-strata experiments in economics. Section 3 introduces the setup and the notation. Section 4 presents our main results, for paired designs. Section 5 presents our simulation study. Section 6 presents our empirical applications. In Section B of the Web Appendix, we use simulations to show that our results for paired designs extend to stratified experiments with small strata.

## 2 Paired and small-strata experiments in economics

We searched the 2014-2018 issues of the American Economic Journal: Applied Economics (AEJ Applied) for paired randomized experiments or stratified experiments with ten or less randomization units per strata. 50 field experiments papers were published over that period. Three relied on a paired randomization for all of their analysis, while one relied on a paired randomization for part of its analysis.[2] 21 papers mentioned that the randomization was stratified (though not paired). For 18 of those papers, we could compute the average number of randomization units per strata, either by reading the paper or by opening the paper's data set. For the remaining three papers, either the data set was not available online, or it did not include the stratification variable. Among those 18 papers, seven have 10 or less randomization units per strata. Overall, at least 22% of the 50 field

[2] Beuermann et al. (2015) use a paired design to estimate the spillover effects of the intervention they consider. Their estimation of the direct effects of that intervention relies on another type of randomization. We only include their spillover analysis in our survey and in our replication.

experiments published by the AEJ Applied over that period are paired or stratified experiments with ten or less randomization units per strata.

To increase our sample of paired experiments, we searched the AEA's registry website for randomized controlled trials (`https://www.socialscienceregistry.org`). We looked at all completed projects, whose randomization method included the prefix "pair" and that had either a working or a published paper. Thus, we found five more paired experiments.

In total we found 16 papers. The list is in Table 5 in the Web Appendix. 14 are published, two are not. In one of them, the regression is at the level of the randomization unit. For example, researchers randomly assigned some firms to the treatment or to the control, and then their regressions are at the firm level. In 15 of them, the regression is at a more disaggregated level than the randomization unit. For instance, researchers randomly assigned some schools to the treatment or to the control, and then their regressions are at the student level.

Across the nine paired experiments, the median number of pairs is 28, and the median number of observations per unit of randomization is 99. To estimate the treatment effect, five articles include pair fixed effects in all of their regressions, three articles include pair fixed effects in some but not all of their regressions, and one article does not include pair fixed effects in any regression. To conduct inference, eight articles out of nine cluster standard errors at the randomization-unit level, and one article does not cluster standard errors. None clusters standard errors at the pair level.

Across the seven small-strata experiments, the median number of units per strata is 7, the median number of strata is 48, and the median number of observations per unit of randomization is 26. To estimate the treatment effect, six articles include strata fixed effects in all their regressions, and one article does not include strata fixed effects in any regression. To conduct inference, all articles cluster standard errors at the randomization-unit level.

In the following sections, we focus on paired experiments. However, in Section B of the Web Appendix, we use simulations to show that the main results we derive for paired experiments extend to small-strata experiments.

# 3 Setup and variance estimators

We consider a population of $2P$ randomization units. Unlike Abadie and Imbens (2008) or Bai et al. (2019), we do not assume that the randomization units are an i.i.d. sample drawn from a super population. Instead, that population is fixed, and its characteristics are not random. This modelling framework is similar to that in Neyman (1923) or Abadie et al. (2020). Our survey suggests it is applicable to a majority of paired- and small-strata experiments conducted in economics. The randomization units are drawn from a larger population in only one of the nine paired experiments we found.[3] In all the other paired experiments, and in all the stratified experiments, the sample is a convenience sample, consisting of volunteers to receive the treatment, or of units located in areas where the RCT implementing partner operates, or of units located in areas where conducting the research was easier.[4] When the randomization units are an i.i.d. sample drawn from a super population, our results still hold, conditional on the sample drawn.

The $2P$ units are matched into $P$ pairs. Pairs are created by grouping together units with the closest value of some baseline variables predicting the outcome. In our fixed-population framework, pairing is not random, as it depends on fixed units' characteristics. The pairs are indexed by $p \in \{1, \ldots, P\}$, and the two randomization units in pair $p$ are indexed by $g \in \{1, 2\}$. Unit $g$ in pair $p$ has $n_{gp}$ observations, so that pair $p$ has $n_p = n_{1p} + n_{2p}$ observations, and the population has $n = \sum_{p=1}^{P} n_p$ observations.

Treatment is assigned as follows. For all $p \in \{1, \ldots, P\}$ and $g \in \{1, 2\}$, let $W_{gp}$ be an indicator variable equal to 1 if unit $g$ in pair $p$ is treated, and to 0 otherwise. We assume that the treatments satisfy the following conditions.

**Assumption 1** (Paired assignment)**.**

1. For all $p$, $W_{1p} + W_{2p} = 1$.
2. $\mathbb{P}(W_{gp} = 1) = \frac{1}{2}$ for all $g$ and $p$.
3. For all $p \neq p'$, $(W_{1p}, W_{2p}) \perp\!\!\!\perp (W_{1p'}, W_{2p'})$.

Point 1 requires that in each pair, one of the two randomization units gets treated. Point 2 requires that the two units have the same probability of being treated. Finally, Point 3 requires that the treatments be independent across pairs.

Let $y_{igp}(1)$ and $y_{igp}(0)$ represent the potential outcomes of observation $i$ in randomization unit $g$ and pair $p$ with and without the treatment, respectively. We follow the randomization

[3] This is in line with Muralidharan and Niehaus (2017), who show that in only 31% of the RCTs published in top 5 journals between 2001 and 2016, the randomization units are drawn from a larger population.

[4] For instance, Glewwe et al. (2016) conducted their study in rural counties of the Gansu Chinese province that were located close to the provincial capital, which eased monitoring by Gansu's Center for Disease Control.

inference literature (see Abadie et al., 2020) and assume that potential outcomes are fixed.[5] Let $\tau_{igp} = y_{igp}(1) - y_{igp}(0)$ be the treatment effect of that observation. The observed outcome is $Y_{igp} = y_{igp}(1)W_{gp} + y_{igp}(0)(1 - W_{gp})$. We focus on the average treatment effect

$$\tau = \frac{1}{n}\sum_{p=1}^{P}\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}(y_{igp}(1) - y_{igp}(0)).$$

We consider two estimators of $\tau$. Let $T_p = n_{1p}W_{1p}+n_{2p}W_{2p}$ and $C_p = n_{1p}(1-W_{1p})+n_{2p}(1-W_{2p})$ be the number of treated and untreated observations in pair $p$. Let $T = \sum_{p=1}^{P} T_p$ and $C = \sum_{p=1}^{P} C_p$ be the total number of treated and untreated observations. The first estimator $\widehat{\tau}$ is the OLS estimator from the regression of the observed outcome $Y_{igp}$ on a constant and $W_{gp}$:

$$Y_{igp} = \widehat{\alpha} + \widehat{\tau}W_{gp} + \varepsilon_{igp} \qquad i = 1, 2, \ldots, n_{gp};\ g = 1, 2;\ p = 1, \ldots, P. \tag{1}$$

$\widehat{\tau}$ is the well-known difference-in-means estimator:

$$\widehat{\tau} = \sum_{p=1}^{P}\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}\frac{Y_{igp}W_{gp}}{T} - \sum_{p=1}^{P}\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}\frac{Y_{igp}(1 - W_{gp})}{C}.$$

The second estimator is the pair-fixed-effects estimator, $\widehat{\tau}_{fe}$, obtained from the regression of the observed outcome $Y_{igp}$ on $W_{gp}$ and a set of pair fixed effects $(\delta_{ig1}, \ldots, \delta_{igP})$:

$$Y_{igp} = \widehat{\tau}_{fe}W_{gp} + \sum_{p=1}^{P}\widehat{\gamma}_p\delta_{igp} + u_{igp}, \qquad i \in \{1, \ldots, n_{gp}\};\ g \in \{1, 2\};\ p \in \{1, \ldots, P\}. \tag{2}$$

It follows from, e.g., Equation (3.3.7) in Angrist and Pischke (2008) and a few lines of algebra that

$$\widehat{\tau}_{fe} = \sum_{p=1}^{P}\omega_p\sum_{g=1}^{2}\left[W_{gp}\sum_{i=1}^{n_{gp}}\frac{Y_{igp}}{n_{gp}} - (1 - W_{gp})\sum_{i=1}^{n_{gp}}\frac{Y_{igp}}{n_{gp}}\right], \qquad \text{where} \quad \omega_p = \frac{\left(n_{1p}^{-1} + n_{2p}^{-1}\right)^{-1}}{\sum_{p'=1}^{P}\left(n_{1p'}^{-1} + n_{2p'}^{-1}\right)^{-1}}.$$

We study the variance estimators of $\widehat{\tau}$ and $\widehat{\tau}_{fe}$ arising from Regressions (1) and (2) above, when the regression is clustered at the pair or at the randomization-unit level.[6] Lemma 3.1 below gives simple expressions of those four variance estimators. Let $SET_p = \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} W_{gp}\varepsilon_{igp}$ and

[5] In a previous version of the paper, we allowed potential outcomes to be stochastic. For instance, in Crépon et al. (2015), a villager's income may be affected by stochastic events like weather shocks. Having stochastic potential outcomes does not change our main results, see de Chaisemartin and Ramirez-Cuellar (2020).

[6] The clustered-variance estimators we study are those proposed in Liang and Zeger (1986).

$SEU_p = \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}(1-W_{gp})\varepsilon_{igp}$ respectively be the sum of the residuals $\varepsilon_{igp}$ for the treated and untreated observations in pair $p$. Similarly, let $SET_{p,fe} = \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} W_{gp}u_{igp}$ and $SEU_{p,fe} = \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}(1-W_{gp})u_{igp}$ respectively be the sum of the residuals $u_{igp}$ for the treated and untreated observations in pair $p$.

**Lemma 3.1** (Clustered variance estimators for $\widehat{\tau}$ and $\widehat{\tau}_{fe}$)**.**

1. *The pair-clustered variance estimator (PCVE) of* $\widehat{\tau}$ *is* $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = \sum_{p=1}^{P}\left(\frac{SET_p}{T} - \frac{SEU_p}{C}\right)^2$.
2. *The randomization-unit-clustered variance estimator (UCVE) of* $\widehat{\tau}$ *is* $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) = \sum_{p=1}^{P}\left(\frac{SET_p^2}{T^2} + \frac{SEU_p^2}{C^2}\right)$.
3. *The PCVE of* $\widehat{\tau}_{fe}$ *is* $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P}\omega_p^2 SET_{p,fe}^2\left(\frac{1}{n_{1p}} + \frac{1}{n_{2p}}\right)^2$.
4. *The UCVE of* $\widehat{\tau}_{fe}$ *is* $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P}\omega_p^2 SET_{p,fe}^2\left(\frac{1}{n_{1p}^2} + \frac{1}{n_{2p}^2}\right)$.

*Proof.* See Appendix D. □

We also study two other estimators of $\mathbb{V}(\widehat{\tau})$. Those estimators have been proposed in the one-observation-per-randomization unit special case, but it is straightforward to extend them to the case where all randomization units have the same number of observations, as stated below:[7]

**Assumption 2.** For all $p$, $n_{1p} = n_{2p} = \frac{n}{2P}$.

Let $\widehat{\tau}_p = \sum_g [W_{gp}\frac{1}{n_{gp}}\sum_i Y_{igp} - (1-W_{gp})\frac{1}{n_{gp}}\sum_i Y_{igp}]$ denote the treatment-effect estimator in pair $p$. The first alternative estimator we consider is a slightly modified version of the pairs-of-pairs variance estimator (POPVE) proposed by Abadie and Imbens (2008). We only define it when the number of pairs $P$ is even, but in our application in Section 6 we propose a simple method to extend it to cases where the number of pairs is odd. Let $x_{g,p}$ denote the value of a predictor of the outcome in pair $p$'s unit $g$. Pairs are ordered according to their value of $\frac{x_{1,p}+x_{2,p}}{2}$, the two pairs with the lowest value are matched together, the next two pairs are matched together, and so on and so forth. Let $R = \frac{P}{2}$. For any $r \in \{1, ..., R\}$ and for any $p \in \{1, 2\}$, let $\widehat{\tau}_{pr}$ denote the treatment effect estimator in pair $p$ of pair of pairs $r$. Then, the POPVE is defined as

$$\widehat{\mathbb{V}}_{pop}(\widehat{\tau}) = \frac{1}{P^2}\sum_{r=1}^{R}(\widehat{\tau}_{1r} - \widehat{\tau}_{2r})^2.$$

$x_{g,p}$, the variable used to match pairs into pairs of pairs, could be the average value of the outcome at baseline in pair $p$'s unit $g$. Or it could be the covariate used to form the pairs, when only one

[7] Extending those variance estimators when Assumption 2 fails is left for future work.

covariate is used. In our application in Section 6, we use the baseline outcome to match pairs into pairs of pair, because the covariates used to match units into pairs are unavailable in most of the data sets of the papers we revisit. Based on Lemma 4.1, we will argue below that the baseline outcome should often be a good choice to match pairs into pairs of pairs. The variable one uses to form pairs of pairs should be pre-specified and not a function of the treatment assignment. Otherwise, researchers could try to find the variable minimizing the POPVE, which would obviously lead to incorrect inference.

There are two differences between $\widehat{\mathbb{V}}_{pop}(\widehat{\tau})$ and the variance estimator proposed in Equation (3) in Abadie and Imbens (2008). First, we match pairs with respect to a single covariate, while Abadie and Imbens (2008) consider matching with respect to a potentially multidimensional vector of covariates. This difference is not of essence: we could easily allow pairs to be matched on several covariates. We focus on the unidimensional case as that is the one we use in our application, where the matching is done based on the baseline outcome. Second, the estimator in Abadie and Imbens (2008) matches pairs with replacement, while $\widehat{\mathbb{V}}_{pop}(\widehat{\tau})$ matches pairs without replacement. If after ordering pairs according to their value of $\frac{x_{1,p}+x_{2,p}}{2}$, pair 2 is closer to pair 3 than pair 4, pair 2 will be matched both to pairs 1 and 3 in Abadie and Imbens (2008), while $\widehat{\mathbb{V}}_{pop}(\widehat{\tau})$ will match pair 1 to pair 2, and pair 3 to pair 4. Matching without replacement makes the finite-sample properties of $\widehat{\mathbb{V}}_{pop}(\widehat{\tau})$ easier to analyze, but should not change its large-sample properties.

The second alternative variance estimator we consider is that proposed by Bai et al. (2019) in their Equation (20) (BRSVE). Again, we define this estimator when the number of pairs $P$ is even. With our notation, their estimator is

$$\widehat{V}_{brs}(\widehat{\tau}) = \frac{1}{P^2}\sum_{p=1}^{P}\widehat{\tau}_p^2 - \frac{1}{2}\left(\frac{2}{P^2}\sum_{r=1}^{R}\widehat{\tau}_{1r}\widehat{\tau}_{2r} + \frac{\widehat{\tau}^2}{P}\right).$$

Bai et al. (2019) propose another variance estimator in their Equation (27). That estimator is less amenable to simple comparisons with the UCVE, PCVE, and POPVE, so we do not analyze its properties. However, we compute it in our applications, and find that it is typically similar to the POPVE and BRSVE.

# 4 Main results

In this section, we present our main findings, that are derived under Assumption 2. Then, the number of treated and untreated observations are equal: $T = C = n/2$, so

$$\widehat{\tau} = \sum_{p=1}^{P} \frac{2}{n} \sum_{g=1}^{2} \left[ W_{gp} \sum_{i=1}^{n_{gp}} Y_{igp} - (1 - W_{gp}) \sum_{i=1}^{n_{gp}} Y_{igp} \right] = \sum_{p=1}^{P} \frac{\widehat{\tau}_p}{P}.$$

$\widehat{\tau}$ is the average of the treatment-effect estimators in each pair. Then, by Point 3 of Assumption 1,

$$\mathbb{V}(\widehat{\tau}) = \sum_{p=1}^{P} \mathbb{V}(\widehat{\tau}_p)/P^2. \tag{3}$$

Finally, under Assumption 2, one can show that the difference-in-means and the fixed-effects estimators are equal: $\widehat{\tau} = \widehat{\tau}_{fe}$,[8] and that both are unbiased estimators of the ATE.

## 4.1 Finite-sample results

In this section, we consider the finite-sample properties of the six variance estimators we study. Let $\tau_{\cdot r} = \frac{1}{2}(\tau_{1r} + \tau_{2r})$ denote the average treatment effect in pair of pairs $r$.

**Lemma 4.1.** *If Assumptions 1 and 2 hold and $P$ is even,*[9]

1. $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = \widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$, *and* $\mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right] = \mathbb{V}(\widehat{\tau}) + \frac{1}{P(P-1)}\sum_{p=1}^{P}(\tau_p - \tau)^2$.
2. $\mathbb{E}\left[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})\right] = \mathbb{V}(\widehat{\tau}) + \frac{1}{P^2}\sum_{r=1}^{R}(\tau_{1r} - \tau_{2r})^2$.
3. $\widehat{\mathbb{V}}_{brs}(\widehat{\tau}) = \frac{1}{2}\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) + \frac{1}{2}\widehat{\mathbb{V}}_{pop}(\widehat{\tau})$
4. *If* $\frac{1}{R}\sum_{r=1}^{R}\sum_{p=1,2}\frac{1}{2}(\tau_{pr} - \tau_{\cdot r})^2 \leq \frac{1}{R-1}\sum_{r=1}^{R}(\tau_{\cdot r} - \tau)^2$,
   (a) $\mathbb{E}\left[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})\right] \leq \mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right]$,
   (b) $\mathbb{E}\left[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})\right] \leq \mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{brs}(\widehat{\tau})\right]$,
   (c) $\mathbb{E}\left[\widehat{\mathbb{V}}_{brs}(\widehat{\tau})\right] \leq \mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right]$.

*Proof.* See Appendix A. ☐

[8] For this result to hold, we can relax Assumption 2 as long as the two randomization units in a pair have the same number of observations.

[9] Point 1 of the lemma holds even if $P$ is uneven.

Point 1 shows that the PCVEs without and with pair fixed effects are equal, and that after a degrees-of-freedom correction, they are conservative estimators of the variance of $\widehat{\tau}$. This second statement follows from Proposition 1 in Imai et al. (2009), once noted that the PCVE without pair fixed effects is equal to the sample variance of the pair-level-treatment-effect estimators $\widehat{\tau}_p$. The PCVEs are unbiased if the treatment effect is constant across pairs.

Point 2 shows that the POPVE is conservative in general, and unbiased if the treatment effect is constant within pairs-of-pairs. The less treatment effect heterogeneity within pairs of pairs, the less conservative the POPVE. An important practical consequence of Point 2 is that the variable used to form pairs of pairs should be a good predictor of pairs' treatment effect. The baseline value of the outcome may often be a good predictor of pairs' treatment effect. For instance, treatments sometimes produce a stronger effect on units with the lowest baseline outcome, thus leading to a catch-up mechanism (see for instance Glewwe et al., 2016). Point 2 of Lemma 4.1 is related to Theorem 1 in Abadie and Imbens (2008), though there are a few differences. Abadie and Imbens (2008) assume that the experimental units are drawn from a super population, and show that once properly normalized, their estimator is consistent for the normalized variance of $\widehat{\tau}$ conditional on the covariates used for pairing. The fact that the POPVE is conservative in Lemma 4.1 and consistent in their Theorem 1 comes from the fact we do not assume that the experimental units are an i.i.d. sample from a super population. In our setting, $P\widehat{\mathbb{V}}_{pop}(\widehat{\tau})$ remains conservative even when the number of units and pairs goes to infinity, as shown in Theorem 4.4 below.

Point 3 shows that the BRSVE is equal to the average of the PCVE and POPCVE. Then, it follows from Points 1 and 2 that $\frac{P}{P-1}\widehat{\mathbb{V}}_{brs}(\widehat{\tau})$ is conservative. Point 3 is related to Lemma 6.4 and Theorem 3.3 in Bai et al. (2019), where the authors show that $P\widehat{\mathbb{V}}_{brs}(\widehat{\tau})$ is consistent for the normalized variance of $\widehat{\tau}$. Here as well, the fact that $P\widehat{\mathbb{V}}_{brs}(\widehat{\tau})$ is conservative in Lemma 4.1 and consistent in Bai et al. (2019) comes from the fact we do not assume that the experimental units are an i.i.d. sample drawn from a super population.

Finally, Point 4 shows that if the treatment effect varies less within than across pairs of pairs, the POPVE is less conservative than the degrees-of-freedom-adjusted PCVE and BRSVE, and the BRSVE is less conservative than the degrees-of-freedom-adjusted PCVE. A sufficient condition to have that the treatment effect varies less within than across pairs of pairs is $\frac{1}{R}\sum_{r=1}^{R}(\tau_{1r}-\tau)(\tau_{2r}-\tau)\geq 0$, meaning that the covariance between the treatment effects of the two pairs in the same pair of pairs is positive. In Section 6, we compare the three variance estimators in a number of empirical applications.

Overall, Lemma 4.1 shows that four of the variance estimators we consider are conservative. In the fixed population framework of Neyman (1923) and Abadie et al. (2020) we adopt here, it is common to have conservative variance estimators when the treatment effect is heterogeneous. For instance, Abadie et al. (2020) show that in randomized experiments where $n_1$ units out of $n$ are assigned to the treatment, the standard heteroskedasticity-robust variance estimator is also conservative if the treatment effect is heterogeneous, so having a conservative variance estimator is not specific to our paper.

Contrary to the four estimators in Lemma 4.1, the UCVE with fixed effects may be very liberal.

**Lemma 4.2.** *If Assumption 2 holds, then* $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = 2\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$.

*Proof.* See Appendix D. □

Lemmas 4.1 and 4.2 imply that the UCVE with pair fixed effects may be downward biased: under constant treatment effects, its expectation is equal to a half of the true variance of $\widehat{\tau}$. Then, using that estimator may severely distort inference on $\tau$. In Section E of the Web Appendix, we show that Lemma 4.2 still approximately holds when Assumption 2 fails, unless randomization units in the same pair have very heterogeneous numbers of observations. Specifically, Lemma E.1 shows that $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ is included between 1/2 and 5/9 as long as $n_{1p}/n_{2p}$ is included between 0.5 and 2 for all $p$, meaning that in each pair the first randomization unit has between half and twice as many observations as the second one. This condition should hold in most applications.

Intuitively, the UCVE is biased because it does not account for the perfect negative correlation of the treatments of the two units in the same pair, a correlation accounted for by the PCVE.[10] However, the direction of the bias depends on whether fixed effects are included in the regression. Indeed, the next lemma shows that without fixed effects, the UCVE will often be upward biased, and more upward biased than the PCVE without fixed effects. For all $d \in \{0, 1\}$, let $\overline{y}_p(d) \equiv \frac{1}{2}\sum_g \overline{y}_{gp}(d)$, and $\overline{y}(d) \equiv \sum_p \overline{y}_p(d)/P$.

[10] In matching studies, Abadie and Spiess (2016) highlight a different but related phenomenon: matching on covariates leads to a correlation between the matched units' outcomes, which has to be accounted for, for instance by clustering.

**Lemma 4.3.** *If Assumptions 1 and 2 hold, then*

$$\mathbb{E}\left[\frac{P}{P-1}\left(\widehat{\mathbb{V}}_{unit}(\widehat{\tau})-\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right)\right]=\frac{2}{P}\left(\frac{1}{P-1}\sum_p\left(\overline{y}_p(0)-\overline{y}(0)\right)\left(\overline{y}_p(1)-\overline{y}(1)\right)\right.$$
$$\left.-\frac{1}{P}\sum_p\sum_g\frac{1}{2}\left(\overline{y}_{gp}(0)-\overline{y}_p(0)\right)\left(\overline{y}_{gp}(1)-\overline{y}_p(1)\right)\right).$$

*Proof.* See Appendix D. □

Lemma 4.3 shows that the expectation of the difference between the UCVE and PCVE in regressions without fixed effects is proportional to the difference between the between-pair and within-pair covariance of the two potential outcomes. In most applications, both terms should be positive, as the two potential outcomes should be positively correlated. One may also expect their difference to be positive, as units in the same pair should have more similar potential outcomes than units in different pairs. For instance, in the extreme case where units in the same pair have equal potential outcomes, the second term is equal to 0. Then, the expectation of the difference between the UCVE and PCVE is positive, and it follows from Point 1 of Lemma 4.1 that the UCVE is upward biased. Contrary to the PCVEs, POPVE, and BRSVE, the UCVE in regressions without fixed effects is biased even when the treatment effect is homogeneous across pairs.

The clustered-variance estimators we study are those proposed in Liang and Zeger (1986). Typically, statistical softwares report degrees-of-freedom (DOF) adjusted versions of those estimators: the Liang and Zeger estimator is multiplied by $n/(n-K)$, where $n$ is the sample size and $K$ the number of regressors (see (StataCorp, 2017)).[11] In general, this DOF adjustment does not change the estimator very much. An exception is when the regression has pair fixed effects, and when each randomization unit has only one observation. Then, $n/(n-K)=2P/(P-1)$: the regression has $2P$ observations and $P+1$ regressors. This fact and Lemma 4.2 imply that in this special case, the unclustered DOF adjusted variance estimator is almost equal to the PCVE. Then, Theorem 4.4 below implies that if the treatment effect is constant across pairs, $t$-tests using the unclustered DOF adjusted variance estimator have nominal size, as found by Bruhn and McKenzie (2009). On the other hand, the DOF-adjusted PCVE is too large, and becomes conservative. Outside of the one-observation-per-randomization-unit special case, the DOF adjustment does not matter much, especially when randomization units have a large number of observations. In our survey of paired

[11] In Stata, this degrees-of-freedom adjustment is implemented when one uses the regress command with pair indicators, not when one uses the xtregress command (see Cameron and Miller (2015)).

experiments above, the median number of observations per randomization unit is equal to 99.

### 4.2 Large sample results

Let

$$\begin{aligned}
\sigma^2_{pair} &= \lim_{P\to+\infty} \frac{P\mathbb{V}(\widehat{\tau})}{P\mathbb{V}(\widehat{\tau}) + \frac{1}{P}\sum_p(\tau_p-\tau)^2} \\
\sigma^2_{pop} &= \lim_{P\to+\infty} \frac{P\mathbb{V}(\widehat{\tau})}{P\mathbb{V}(\widehat{\tau}) + \frac{1}{P}\sum_r(\tau_{1r}-\tau_{2r})^2} \\
\sigma^2_{brs} &= \lim_{P\to+\infty} \frac{P\mathbb{V}(\widehat{\tau})}{P\mathbb{V}(\widehat{\tau}) + \frac{1}{2P}\sum_r(\tau_{1r}-\tau_{2r})^2 + \frac{1}{2P}\sum_p(\tau_p-\tau)^2} \\
\Delta_{cov,P} &= \frac{1}{P}\sum_p (\overline{y}_p(0)-\overline{y}(0))(\overline{y}_p(1)-\overline{y}(1)) - \frac{1}{P}\sum_p \frac{1}{2}\sum_g \left(\overline{y}_{gp}(0)-\overline{y}_p(0)\right)\left(\overline{y}_{gp}(1)-\overline{y}_p(1)\right) \\
\sigma^2_{unit} &= \lim_{P\to+\infty} \frac{P\mathbb{V}(\widehat{\tau})}{P\mathbb{V}(\widehat{\tau}) + \frac{1}{P}\sum_p(\tau_p-\tau)^2 + 2\Delta_{cov,P}},
\end{aligned}$$

where Assumption 3 below ensures the limits in the previous display exist.

**Assumption 3.**

1. For every $d$, $g$ and $p$, $\left|\overline{y}_{gp}(d)\right|^{2+\epsilon} \leq M < +\infty$, for some $M$, $\epsilon > 0$.

2. When $P \to +\infty$, $\frac{1}{P}\sum_p \tau_p$, $\frac{1}{P}\sum_p(\tau_p-\tau)^2$, $\frac{1}{P}\sum_r(\tau_{1r}-\tau_{2r})^2$, and $\Delta_{cov,P}$ converge towards finite limits, and $P\mathbb{V}(\widehat{\tau})$ and $P\mathbb{V}(\widehat{\tau}) + \frac{1}{P}\sum_p(\tau_p-\tau)^2 + 2\Delta_{cov,P}$ converge towards strictly positive finite limits.

3. As $P \to \infty$, $\sum_{p=1}^P \mathbb{E}[|\widehat{\tau}_p-\tau_p|^{2+\epsilon}]/S_P^{2+\epsilon} \to 0$ for some $\epsilon > 0$, where $S_P^2 \equiv P^2\mathbb{V}(\widehat{\tau})$.

Point 1 of Assumption 3 guarantees that we can apply the strong law of large numbers (SLLN) in Lemma 1 in Liu (1988) to the sequence $(\widehat{\tau}_p^2)_{p=1}^{+\infty}$. Point 2 ensures that $P\mathbb{V}(\widehat{\tau})$ and $P\widehat{\mathbb{V}}_{unit}(\widehat{\tau})$ do not converge towards 0. Point 3 guarantees that we can apply the Lyapunov central limit theorem to $(\widehat{\tau}_p)_{p=1}^{+\infty}$. Then,

**Theorem 4.4.** *(t-stats' asymptotic behavior) Under Assumptions 1, 2 and 3,*

1. $(\widehat{\tau}-\tau)/\sqrt{\widehat{\mathbb{V}}_{pair}(\widehat{\tau})} = (\widehat{\tau}_{fe}-\tau)/\sqrt{\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})} \xrightarrow{d} \mathcal{N}(0,\sigma^2_{pair})$. $\sigma^2_{pair} \leq 1$, *and if* $\tau_p = \tau$ *for every* $p$, $\sigma^2_{pair} = 1$.

2. $(\widehat{\tau}-\tau)/\sqrt{\widehat{\mathbb{V}}_{pop}(\widehat{\tau})} \xrightarrow{d} \mathcal{N}(0,\sigma^2_{pop})$. $\sigma^2_{pop} \leq 1$, *and if* $\tau_{1r} = \tau_{2r}$ *for every* $r$, $\sigma^2_{pop} = 1$.

3. $(\widehat{\tau}-\tau)/\sqrt{\widehat{\mathbb{V}}_{brs}(\widehat{\tau})} \xrightarrow{d} \mathcal{N}(0,\sigma^2_{brs})$. $\sigma^2_{brs} \leq 1$, *and if* $\tau_p = \tau$ *for every* $p$, $\sigma^2_{brs} = 1$.

4. $(\widehat{\tau}_{fe}-\tau)/\sqrt{\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})} \xrightarrow{d} \mathcal{N}(0,2\sigma^2_{pair})$.

5. $(\widehat{\tau}-\tau)/\sqrt{\widehat{\mathbb{V}}_{unit}(\widehat{\tau})} \xrightarrow{d} \mathcal{N}(0,\sigma^2_{unit})$.

6. (a) $\sigma^2_{pair} \leq \sigma^2_{brs} \leq \sigma^2_{pop}$ *if and only if* $0 \leq \lim_{P\to+\infty} \frac{1}{R}\sum_{r=1}^{R}(\tau_{1r}-\tau)(\tau_{2r}-\tau)$.

   (b) $\sigma^2_{unit} \leq \sigma^2_{pair}$ *if and only if* $\Delta_{cov,P}$ *converges towards a positive limit.*

*Proof.* See Appendix D.

Point 1 (resp. 2, 3) shows that when the number of pairs grows, the $t$-statistic of the difference-in-means or fixed-effects estimator using the PCVEs (resp. POPVE, BRSVE) converges to a normal distribution with a mean equal to 0 and a variance lower than 1 in general, but equal to 1 when the treatment effect is homogenous across pairs. Therefore, those t-tests are conservative. Point 6a shows that whenever there is a positive correlation between the treatment effects of the two pairs in the same pair of pair, the t-test using the POPVE is less conservative than that using the BRSVE, which is itself less conservative than that using the PCVE.

Point 4 shows that the $t$-statistic of the fixed-effects estimator using the UCVE converges to a normal distribution with a mean equal to 0 and a variance twice as large as that of the t-statistic using the PCVE. Therefore, comparing that $t$-statistic to critical values of a standard normal could lead to a test with a size larger than its nominal size. For instance, if $\tau = 0$ and the treatment effect is homogenous across pairs, comparing $\left|\widehat{\tau}_{fe}/\sqrt{\widehat{V}_{unit}(\widehat{\tau}_{fe})}\right|$ to 1.96 would lead the analyst to reject the null hypothesis that $\tau = 0$ 16.5% of the times. With heterogeneous treatment effects across pairs, the t-tests using the PCVEs, POPVE, and BRSVE may be conservative, while that using the UCVE with fixed effects may be exact. However, in practice, we do not know if the treatment effect is constant or heterogeneous, and it is common to require that a test control size uniformly across all possible data generating processes. The $t$-tests making use of the PCVEs, BRSVE, and POPVE satisfy that property, unlike the $t$-test making use of the UCVE with fixed effects.

Finally, Point 5 shows that in general the $t$-statistic of the difference-in-means estimator using the UCVE converges towards a normal distribution with a mean equal to 0 and a variance that differs from 1. As shown in Point 6b, the asymptotic variance of this $t$-statistic is lower than that of the $t$-statistic using the PCVE whenever the difference between the between- and within-pairs covariance between the two potential outcomes is positive. Then, that $t$-statistic is conservative, and it is more conservative than that using the PCVE. Point 5 is related to Theorem 3.1 in Bai et al. (2019), who show that when $n_{gp} = 1$, the $t$-test in Point 5 is conservative. The asymptotic variance we obtain is different from theirs, because our two results are derived under different assumptions. For instance, we assume a fixed population, while Bai et al. (2019) assume that the experimental units are an i.i.d. sample drawn from an infinite superpopulation, and that asymptotically the expectation of the potential outcomes of two units in the same pair become equal.

# 5 Simulations using real data

To assess if in practice, the size of the $t$-tests we consider is close to that predicted by Theorem 4.4, we performed Monte-Carlo simulations using a real data set. We use the data from the microfinance experiment in Crépon et al. (2015). The authors matched 162 Moroccan villages into 81 pairs, and in each pair, they randomly assigned one village to a microfinance treatment. They sampled households from each village and measured their outcomes such as their credit access and income. In the paper, the authors report the effect of the microfinance intervention on 82 outcome variables.

For each outcome variable, we construct potential outcomes under the assumption of no effects, i.e., $y_{igpk}(0) = y_{igpk}(1) = Y_{igpk}$, where $Y_{igpk}$ is the value of outcome $k$ for household $i$ in village $g$ and pair $p$. We then simulate 1000 vectors of treatment assignments $W_k^j = ((W_{11,k}^j, W_{21,k}^j), \ldots, (W_{1P,k}^j, W_{2P,k}^j))$, assigning one of the two villages to treatment in each pair. Then, we regress $Y_{igpk}$ on the simulated treatment $W_{gp,k}^j$. We estimate regressions with and without pair-fixed effects, clustering standard errors at the pair level and at the village (unit-of-randomization) level. Thus, we obtain four $t$-statistics, and four 5% level $t$-tests. The estimated size of each $t$-test is just the percentage of times the test is rejected across the 82,000 regressions (82 outcomes $\times$ 1000 simulations). Because the data is generated under the hypothesis of no treatment effect, these $t$-tests should be rejected 5% of the time.

Table 1 shows the estimated sizes of the four $t$-tests. The sizes of the $t$-tests using pair-clustered standard errors are close to 5%, irrespective of whether pair fixed effects are included in the regression. On the other hand, when standard errors are clustered at the village level and pair fixed effects are included, the size of the $t$-test is 17%, very close to the 16.5% level predicted by point 4 of Theorem 4.4. Finally, the size of the $t$-test with village-clustered standard errors without fixed effects is equal to 1.3%. In this application, this $t$-test is very conservative.

The fourth column of the table shows that we obtain similar results if we use a random sample of 20 of the 81 pairs. With less than 20 pairs, standard errors clustered at the pair level become liberal. One may then have to use randomization inference tests. Similarly, in a small-strata experiment with too few strata to cluster at that level, one could use randomization inference, or the variance estimator proposed in Section 9.5.1 of Imbens and Rubin (2015), provided each stratum has at least two treated and two control units.

Table 1: Fraction of times $t$-test is rejected

| Clustering level | Pair Fixed Effects | 5% level $t$-test size | |
|---|---|---|---|
| | | With 81 pairs | With 20 pairs |
| Pair | Yes | 0.0505 | 0.0581 |
| Pair | No | 0.0521 | 0.0595 |
| Village | Yes | 0.1719 | 0.1851 |
| Village | No | 0.0132 | 0.0204 |

Table 1 reports the empirical size of four 5% level $t$-tests in Crépon et al. (2015). For each of the 82 outcomes in the paper, we randomly drew 1000 simulated treatment assignments, following the paired assignment used by the authors, and regressed the outcome on the simulated treatment. The four $t$-tests are computed, respectively, without and with fixed effects in the regression, and clustering standard errors at the village or at the pair level. The size of each test is the percent of times it is rejected across the 82,000 regressions (82 outcomes × 1000 replications). Column 3 (resp. 4) shows the results using the original sample of 81 pairs (resp. a smaller sample of 20 randomly selected pairs).

For 26 of the 82 regressions in (Crépon et al., 2015), the baseline outcome is available in the authors' data set, so for those outcomes we can simulate the POPVE and BRSVE as well. Those estimators are defined under Assumption 2, which does not hold. Therefore in those simulations, we aggregate the data at the village level. We use two samples of 80 and 20 randomly selected pairs out of the original 81 pairs, so as to have an even number of pairs. For each outcome, we simulate 3,000 vectors of treatment assignments, assigning one of the two villages to treatment in each pair. Then, we compute $\widehat{\tau}$, $\widehat{\mathbb{V}}_{pair}(\tau)$, $\widehat{\mathbb{V}}_{pop}(\tau)$, and $\widehat{\mathbb{V}}_{brs}(\tau)$, and the three corresponding 5% level $t$-tests. The estimated size of each $t$-test is shown in Table 2 below. As above, the $t$-test using the PCVE has close to nominal size with as few as 20 pairs. On the other hand, the $t$-tests using the POPVE and BRSVE have greater than nominal size, even with 80 pairs. Accordingly, we run simulations again, duplicating the random sample of 80 pairs twice to have 160 pairs. The $t$-test using the BRSVE now has close to nominal size, but the $t$-test using the POPVE still has greater than nominal size. With a sample of 320 pairs obtained by duplicating the random sample of 80 pairs four times, all tests have close to nominal size. With 20 and 80 pairs, we find in our simulations that the correlation between $\widehat{\mathbb{V}}_{pop}(\tau)$ and $|\widehat{\tau}|$ is much weaker than that between $\widehat{\mathbb{V}}_{pair}(\tau)$ and $|\widehat{\tau}|$. This explains why the t-test using $\widehat{\mathbb{V}}_{pop}(\tau)$ is liberal, despite the fact $\widehat{\mathbb{V}}_{pop}(\tau)$ is unbiased: when $|\widehat{\tau}|$ is large, $\widehat{\mathbb{V}}_{pop}(\tau)$ is less likely to be large than $\widehat{\mathbb{V}}_{pair}(\tau)$, so the POPVE t-test is more liberal. With 160 and 320 pairs, this phenomenon becomes less pronounced. Overall, the asymptotic approximations in Points 2 and 3 of Theorem 4.4 seem to hold only with a large number of pairs, contrary to that in Point 1. In our survey of paired experiments, only one paper has more than 160 pairs, so it seems that $t$-tests based on the POPVE and BRSVE can only be used in a minority of paired experiments.

Table 2: Simulations with data aggregated at village-level to compute $\widehat{\mathbb{V}}_{pop}$ and $\widehat{\mathbb{V}}_{brs}$

| Variance estimator | 5% level $t$-test size | | | |
|---|---|---|---|---|
| | With 20 pairs | With 80 pairs | With 160 pairs | With 320 pairs |
| $\widehat{\mathbb{V}}_{pair}(\tau)$ | 0.0591 | 0.0515 | 0.0500 | 0.0510 |
| $\widehat{\mathbb{V}}_{pop}(\tau)$ | 0.1307 | 0.0834 | 0.0638 | 0.0568 |
| $\widehat{\mathbb{V}}_{brs}(\tau)$ | 0.0826 | 0.0623 | 0.0551 | 0.0531 |

Table 1 reports the empirical size of three 5% level $t$-tests in Crépon et al. (2015), aggregating the data at the village level. For each of the 26 outcomes in the paper for which the baseline outcome is available, we randomly drew 3,000 simulated treatment assignments, following the paired assignment used by the authors, and computed the treatment effect estimator $\widehat{\tau}$, the variance estimators $\widehat{\mathbb{V}}_{pair}(\tau)$, $\widehat{\mathbb{V}}_{pop}(\tau)$, and $\widehat{\mathbb{V}}_{brs}(\tau)$, and the three corresponding $t$-tests. The size of each test is the percent of times it is rejected across the 78,000 regressions (26 outcomes $\times$ 3,000 replications). Column 2 (resp. 3, 4, 5) shows the results using a random sample of 20 pairs (resp. a random sample of 80 pairs, the same random sample of 80 pairs duplicated twice, the same random sample of 80 pairs duplicated four times).

## 6 Application

In this section, we revisit the paired randomized experiments we found in our survey. The data used in four of those papers is publicly available. Therein, the authors estimated the effect of the treatment in 294 regressions, clustering standard errors at the unit-of-randomization level. In Panel A of Table 3, we re-estimate those regressions, clustering standard errors at the pair level, and including the same controls as the authors. In the 240 regressions with fixed effects, the average ratio of the UCVE and PCVE is equal to 0.548. Those ratios are not all exactly equal to 1/2 because Assumption 2 is not always satisfied, but they all are quite close to 1/2, as predicted by Lemma E.1 in the Web Appendix. The authors originally found that the treatment has a 5%-level significant effect in 110 regressions. Using pair-clustered standard errors, we find significant effects in 74 regressions. In the 54 regressions without fixed effects, the UCVE is on average 1.18 times larger than the PCVE. The authors originally found 31 significant effects, whereas we find 36 significant significant effects using PCVE.

Of the remaining five papers, one used standard errors assuming homoskedastic errors. This is not an inference method we consider so we do not include it in our replication. Three papers estimated 132 regressions with fixed effects, clustering standard errors at the unit-of-randomization level. For those regressions, we multiply the UCVE by the average ratio of the PCVE and UCVE found in Panel A of Table 3 to predict the value of the PCVE. Panel B of Table 3 shows that while the authors originally found a 5%-level significant effect in 52 regressions, we find significant effects in 35 regressions using our predicted PCVE. The last paper only estimated regressions without pair fixed effects. As the ratio of the PCVE and UCVE can vary a lot across applications, we do not try to predict the PCVE in that paper.

Table 3: Using unit- or pair-level clustered variance estimators in paired experiments

| | Unit-level divided by pair-level clustered variance estimators | Number of 5%-level significant effects with UCVE | Number of 5%-level significant effects with PCVE | Number of Regressions |
|---|---|---|---|---|
| *Panel A: Articles with publicly available data* | | | | |
| with pair fixed effects | 0.548 | 110 | 74 | 240 |
| without pair fixed effects | 1.184 | 31 | 36 | 54 |
| *Panel B: Articles without publicly available data* | | | | |
| with pair fixed effects | | 52 | 35 | 132 |

The table shows the effect of using pair- rather than unit-level clustered standard errors in seven of the paired randomized experiments we found in our survey. In Panel A, we consider four papers whose data is available online, and re-estimate their regressions clustering standard errors at the pair level. Column 1 shows the ratio of the unit- and pair-level clustered variance estimators, separately for regressions without and with pair fixed effects. Column 2 (resp. 3) shows the number of 5%-level significant effects using unit- (resp. pair-) clustered standard errors. In Panel B, we consider three other papers whose data is not available online, and use the average ratio of the unit- and pair- clustered variance estimators found in Panel A to predict the value of the pair-clustered estimator in the regressions with fixed effects estimated by those papers. Column 2 (resp. 3) shows the number of 5%-level significant effects using unit- (resp. predicted pair-) clustered standard errors.

For 152 of the 294 regressions in Panel A of Table 3, the baseline outcome is available in the authors' data set, so we can estimate the POPVE and BRSVE as well. Those estimators are defined under Assumption 2, which does not hold in all those regressions. Therefore, we compute the POPVE and BRSVE after aggregating the data at the unit-of-randomization level. When the number of pairs is odd, we compute the POPVE twice, first excluding the pair with the lowest value of the baseline outcome, then excluding the pair with the highest value of the baseline outcome, and we finally take the average of the two estimators. We do the same for the BRSVE when the number of pairs is odd. We also recompute the PCVE without pair fixed effects with the aggregated data, using the exact same sample as that used to compute the POPVE and BRSVE. Across those 152 regressions, the POPVE divided by the PCVE is on average equal to 1.003. The BRSVE divided by the PCVE is on average equal to 1.002.[12] Overall, it does not seem one can expect large power gains from using the POPVE and BRSVE.

## 7 Conclusion

Researchers conducting paired or small-strata experiments often use pair- or strata-fixed-effects regressions and cluster standard errors at the unit-of-randomization level to make inference about the average treatment effect. We show that the corresponding $t$-test can overreject the null of no effect. Instead, we recommend using standard errors clustered at the pair or strata level.

[12] The second variance estimator proposed by Bai et al. (2019) in their Equation (27) is also on average higher than the PCVE.

## References


Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. (2017), When should you adjust standard errors for clustering?, Technical report, National Bureau of Economic Research.

Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. M. (2020), 'Sampling-based versus design-based uncertainty in regression analysis', *Econometrica* **88**(1), 265–296.
**URL:** *https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA12675*

Abadie, A. and Imbens, G. W. (2008), 'Estimation of the conditional variance in paired experiments', *Annales d'Economie et de Statistique* pp. 175–187.

Abadie, A. and Spiess, J. (2016), Robust post-matching inference, *in* 'Unpublished Paper). MIT and Harvard University. Retrieved from https://editorialexpress. com/cgi-bin/conference/down load. cgi'.

Ambler, K., Aycinena, D. and Yang, D. (2015), 'Channeling remittances to education: a field experiment among migrants from el salvador', *American Economic Journal: Applied Economics* **7**(2), 207–32.

Angelucci, M., Karlan, D. and Zinman, J. (2015), 'Microcredit impacts: Evidence from a randomized microcredit program placement experiment by compartamos banco', *American Economic Journal: Applied Economics* **7**(1), 151–82.

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

Ashraf, N., Karlan, D. and Yin, W. (2006), 'Deposit collectors', *Advances in Economic Analysis & Policy* **5**(2).

Athey, S. and Imbens, G. W. (2017), Chapter 3 - the econometrics of randomized experiments, *in* A. V. Banerjee and E. Duflo, eds, 'Handbook of Field Experiments', Vol. 1 of *Handbook of Economic Field Experiments*, North-Holland, pp. 73 – 140.

Attanasio, O., Augsburg, B., De Haas, R., Fitzsimons, E. and Harmgart, H. (2015), 'The impacts of microfinance: Evidence from joint-liability lending in mongolia', *American Economic Journal: Applied Economics* **7**(1), 90–122.

Bai, Y., Romano, J. P. and Shaikh, A. (2019), Inference in experiments with matched pairs, Technical Report 2019-63, University of Chicago, Becker Friedman Institute for Economics Working Paper.

Banerjee, A., Duflo, E., Glennerster, R. and Kinnan, C. (2015), 'The miracle of microfinance? evidence from a randomized evaluation', *American Economic Journal: Applied Economics* **7**(1), 22–53.

Banerji, R., Berry, J. and Shotland, M. (2017), 'The impact of maternal literacy and participation programs: Evidence from a randomized evaluation in india', *American Economic Journal: Applied Economics* **9**(4), 303–37.

Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O. and Cruz-Aguayo, Y. (2015), 'One laptop per child at home: Short-term impacts from a randomized experiment in peru', *American Economic Journal: Applied Economics* **7**(2), 53–80.

Björkman Nyqvist, M., de Walque, D. and Svensson, J. (2017), 'Experimental evidence on the long-run impact of community-based monitoring', *American Economic Journal: Applied Economics* **9**(1), 33–69.

Bruhn, M., Leão, L. d. S., Legovini, A., Marchetti, R. and Zia, B. (2016), 'The impact of high school financial education: Evidence from a large-scale evaluation in brazil', *American Economic Journal: Applied Economics* **8**(4), 256–95.

Bruhn, M. and McKenzie, D. (2009), 'In pursuit of balance: Randomization in practice in development field experiments', *American Economic Journal: Applied Economics* **1**(4), 200–232.

Cameron, A. C. and Miller, D. L. (2015), 'A practitioner's guide to cluster-robust inference', *Journal of Human Resources* **50**(2), 317–372.

Crépon, B., Devoto, F., Duflo, E. and Parienté, W. (2015), 'Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in morocco', *American Economic Journal: Applied Economics* **7**(1), 123–50.

de Chaisemartin, C. and Ramirez-Cuellar, J. (2020), 'At what level should one cluster standard errors in paired experiments, and in stratified experiments with small strata?', *arXiv preprint arXiv:1906.00288v4* .

Fryer Jr, R. G. (2017), Management and student achievement: Evidence from a randomized field experiment, Technical report, National Bureau of Economic Research.

Fryer Jr, R. G., Devi, T. and Holden, R. T. (2016), Vertical versus horizontal incentives in education: Evidence from randomized trials, Technical report, National Bureau of Economic Research.

Glewwe, P., Park, A. and Zhao, M. (2016), 'A better vision for development: Eyeglasses and academic performance in rural primary schools in china', *Journal of Development Economics* **122**, 170–182.

Imai, K., King, G. and Nall, C. (2009), 'The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation', *Statistical Science* **24**(1), 29–53.

Imbens, G. W. and Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Lafortune, J., Riutort, J. and Tessada, J. (2018), 'Role models or individual consulting: The impact of personalizing micro-entrepreneurship training', *American Economic Journal: Applied Economics* **10**(4), 222–45.

Liang, K.-Y. and Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

Liu, R. Y. (1988), 'Bootstrap procedures under some non-iid models', *The Annals of Statistics* **16**(4), 1696–1708.

Muralidharan, K. and Niehaus, P. (2017), 'Experimentation at scale', *Journal of Economic Perspectives* **31**(4), 103–24.

Neyman, J. (1923), 'Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes', *Roczniki Nauk Rolniczych* **10**, 1–51.

Panagopoulos, C. and Green, D. P. (2008), 'Field experiments testing the impact of radio advertisements on electoral competition', *American Journal of Political Science* **52**(1), 156–168.

Somville, V. and Vandewalle, L. (2018), 'Saving by default: Evidence from a field experiment in rural india', *American Economic Journal: Applied Economics* **10**(3), 39–66.

StataCorp, L. (2017), *Stata User's Guide*, 15 edn, College Station, Texas.

# A Main text proofs

## Proof of Lemma 4.1

### Point 1

*Proof of* $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = \widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$

It follows from Equations (1) and (2) that

$$\widehat{\alpha} + \widehat{\tau} W_{gp} + \varepsilon_{igp} = \widehat{\tau}_{fe} W_{gp} + \sum_{p=1}^{P} \widehat{\gamma}_p \delta_{igp} + u_{igp}.$$

Rearranging and using the fact that under Assumption 2 $\widehat{\tau} = \widehat{\tau}_{fe}$, one obtains that for every $p$:

$$\varepsilon_{igp} = \widehat{\gamma}_p - \widehat{\alpha} + u_{igp}. \tag{4}$$

Then,

$$\begin{aligned}
\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) &= \frac{1}{T^2} \sum_{p=1}^{P} (SET_p - SEU_p)^2 \\
&= \frac{1}{T^2} \sum_p \left[ \sum_g \sum_i (2W_{gp} - 1) \varepsilon_{igp} \right]^2 \\
&= \frac{1}{T^2} \sum_p \left[ \sum_g \sum_i (2W_{gp} - 1)(\widehat{\gamma}_p - \widehat{\alpha} + u_{igp}) \right]^2 \\
&= \frac{1}{T^2} \sum_p \left[ \sum_g \sum_i (2W_{gp} - 1) u_{igp} + (\widehat{\gamma}_p - \widehat{\alpha}) \sum_g \sum_i (2W_{gp} - 1) \right]^2 \\
&= \frac{4}{T^2} \sum_p \left( \sum_g \sum_i W_{gp} u_{igp} \right)^2. \qquad (5)
\end{aligned}$$

The first equality follows from Point 1 of Lemma 3.1 and Assumption 2. The third equality follows from Equation (4). The fifth follows from the following two facts. First, $\sum_g \sum_i (2W_{gp} - 1) u_{igp} = 2 \sum_g \sum_i W_{gp} u_{igp} - \sum_g \sum_i u_{igp} = 2 \sum_g \sum_i W_{gp} u_{igp}$, since $\sum_g \sum_i u_{igp} = 0$ by definition of $u_{igp}$. Second, $(\widehat{\gamma}_p - \widehat{\alpha}) \sum_g \sum_i (2W_{gp} - 1) = (\widehat{\gamma}_p - \widehat{\alpha}) \left[ \sum_g \sum_i W_{gp} - \sum_g \sum_i (1 - W_{gp}) \right] = (\widehat{\gamma}_p - \widehat{\alpha})[T_p - C_p] = 0$, where the last equality comes from the fact that $n_{1p} = n_{2p}$ by Assumption 2.

Similarly,

$$\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \frac{4}{T^2}\sum_{p=1}^{P} SET_{p,fe}^2 = \frac{4}{T^2}\sum_{p=1}^{P}\left(\sum_g \sum_i W_{gp} u_{igp}\right)^2, \tag{6}$$

where the first equality follows from Point 3 of Lemma 3.1 and Assumption 2. Combining Equations (5) and (6) yields $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = \widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$.

*Proof of* $\mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right] = \mathbb{V}(\widehat{\tau}) + \frac{1}{P(P-1)}\sum_{p=1}^{P}(\tau_p - \tau)^2$

Under Assumption 2, $T = C = n/2$, so

$$\begin{aligned}
\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) &= \sum_{p=1}^{P}\left(\frac{SET_p}{T} - \frac{SEU_p}{C}\right)^2 \\
&= \frac{4}{n^2}\sum_{p=1}^{P}(SET_p - SEU_p)^2 \\
&= \frac{4}{n^2}\sum_{p=1}^{P}\left(\sum_g \sum_i (W_{gp}\varepsilon_{igp} - (1 - W_{gp})\varepsilon_{igp})\right)^2 \\
&= \frac{4}{n^2}\sum_{p=1}^{P}\left(\sum_g \sum_i (2W_{gp} - 1)\varepsilon_{igp}\right)^2 \\
&= \frac{4}{n^2}\sum_{p=1}^{P}\left(\sum_g (2W_{gp} - 1)\sum_i (Y_{igp} - \widehat{\tau}W_{gp} - \widehat{\alpha})\right)^2 \\
&= \frac{4}{n^2}\sum_{p=1}^{P}\left(\sum_g (2W_{gp} - 1)\left(\sum_i Y_{igp} - \widehat{\tau}W_{gp}\frac{n_p}{2} - \widehat{\alpha}\frac{n_p}{2}\right)\right)^2 \\
&= \frac{4}{n^2}\sum_{p=1}^{P}\left(\sum_g (2W_{gp} - 1)\sum_i Y_{igp} - \widehat{\tau}\frac{n_p}{2}\sum_g (2W_{gp} - W_{gp}) - \widehat{\alpha}\frac{n_p}{2}\sum_g (2W_{gp} - 1)\right)^2 \\
&= \frac{4}{n^2}\sum_{p=1}^{P}\left(\sum_g (2W_{gp} - 1)\sum_i Y_{igp} - \widehat{\tau}\frac{n_p}{2}\sum_g W_{gp}\right)^2 \\
&= \frac{4}{n^2}\sum_{p=1}^{P}\left(\sum_g (2W_{gp} - 1)\sum_i Y_{igp} - \widehat{\tau}\frac{n_p}{2}\right)^2 \\
&= \frac{4}{n^2}\sum_{p=1}^{P}\left(\widehat{\tau}_p\frac{n_p}{2} - \widehat{\tau}\frac{n_p}{2}\right)^2 \\
&= \frac{1}{P^2}\sum_{p=1}^{P}(\widehat{\tau}_p - \widehat{\tau})^2.
\end{aligned} \tag{7}$$

The third equality comes from the definition of $SET_p$ and $SEU_p$. The fifth equality follows from the Equation (1). The sixth equality follows from $n_{1p} = n_{2p} = n_p/2$, which is a consequence of Assumption 2. The eighth equality comes from the fact that $\sum_g (2W_{gp} - 1) = 0$, which follows from Point 1 of Assumption 1. The ninth equality follows from Point 1 of Assumption 1. The tenth equality follows from $\sum_g (2W_{gp} - 1) \sum_i Y_{igp} = \sum_g W_{gp} \sum_i Y_{igp} - \sum_g (1 - W_{gp}) \sum_i Y_{igp} = n_p \widehat{\tau}_p/2$. The eleventh equality follows from Assumption 2.

Now, consider Equation (7). Adding and subtracting $\tau$ and $\tau_p \equiv \mathbb{E}[\widehat{\tau}_p]$,

$$\begin{aligned}
\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) &= \frac{1}{P^2} \sum_{p=1}^{P} \left( (\widehat{\tau}_p - \tau_p) - (\widehat{\tau} - \tau) + (\tau_p - \tau) \right)^2 \\
&= \frac{1}{P^2} \sum_{p=1}^{P} \left[ (\widehat{\tau}_p - \tau_p)^2 + (\widehat{\tau} - \tau)^2 + (\tau_p - \tau)^2 - 2(\widehat{\tau}_p - \tau_p)(\widehat{\tau} - \tau) \right. \\
&\qquad \left. + 2(\widehat{\tau}_p - \tau_p)(\tau_p - \tau) - 2(\widehat{\tau} - \tau)(\tau_p - \tau) \right].
\end{aligned}$$

Taking the expected value, and given that $\mathbb{E}[\widehat{\tau}] = \tau$ and $\mathbb{E}[\widehat{\tau}_p] = \tau_p$,

$$\begin{aligned}
\mathbb{E}[\widehat{\mathbb{V}}_{pair}(\widehat{\tau})] &= \frac{1}{P^2} \sum_{p=1}^{P} \left[ \mathbb{V}(\widehat{\tau}_p) + \mathbb{V}(\widehat{\tau}) + (\tau_p - \tau)^2 - 2\mathrm{Cov}(\widehat{\tau}, \widehat{\tau}_p) \right] \\
&= \frac{1}{P^2} \sum_{p=1}^{P} \left[ \left(1 - \frac{2}{P}\right) \mathbb{V}(\widehat{\tau}_p) + \mathbb{V}(\widehat{\tau}) + (\tau_p - \tau)^2 \right] \\
&= \left(1 - \frac{2}{P}\right) \mathbb{V}(\widehat{\tau}) + \frac{1}{P^2} \sum_{p=1}^{P} \mathbb{V}(\widehat{\tau}) + \frac{1}{P^2} \sum_{p=1}^{P} (\tau_p - \tau)^2 \\
&= \left(1 - \frac{1}{P}\right) \mathbb{V}(\widehat{\tau}) + \frac{1}{P^2} \sum_{p=1}^{P} (\tau_p - \tau)^2.
\end{aligned}$$

The second equality follows from the fact that by Point 3 of Assumption 1 and Assumption 2, $\mathrm{Cov}(\widehat{\tau}_p, \widehat{\tau}) = \mathrm{Cov}\left(\widehat{\tau}_p, \sum_{p'} \frac{1}{P} \widehat{\tau}_{p'}\right) = \frac{1}{P} \mathbb{V}(\widehat{\tau}_p)$. The third equality comes from Equation (3). This proves the result.

**Point 2**

$$
\begin{aligned}
\widehat{\mathbb{V}}_{pop}(\widehat{\tau}) &= \frac{1}{P^2}\sum_{r=1}^{R}(\widehat{\tau}_{1r} - \widehat{\tau}_{2r})^2, \\
&= \frac{1}{P^2}\sum_{r=1}^{R}(\widehat{\tau}_{1r}^2 + \widehat{\tau}_{2r}^2 - 2\widehat{\tau}_{1r}\widehat{\tau}_{2r}).
\end{aligned}
$$

Taking expected value,

$$
\begin{aligned}
\mathbb{E}[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})] &= \frac{1}{P^2}\sum_{r=1}^{R}\mathbb{E}(\widehat{\tau}_{1r}^2 + \widehat{\tau}_{2r}^2 - 2\widehat{\tau}_{1r}\widehat{\tau}_{2r}), \\
&= \frac{1}{P^2}\sum_{r=1}^{R}(\mathbb{V}(\widehat{\tau}_{1r}) + \mathbb{V}(\widehat{\tau}_{2r}) + \tau_{1r}^2 + \tau_{2r}^2 - 2\tau_{1r}\tau_{2r}), \\
&= \frac{1}{P^2}\sum_{p=1}^{P}\mathbb{V}(\widehat{\tau}_p) + \frac{1}{P^2}\sum_{r=1}^{R}(\tau_{1r} - \tau_{2r})^2, \\
&= \mathbb{V}(\widehat{\tau}) + \frac{1}{P^2}\sum_{r=1}^{R}(\tau_{1r} - \tau_{2r})^2. \qquad (8)
\end{aligned}
$$

The second equality follows from properties of the variance and that $\mathbb{E}[\widehat{\tau}_{1r}] = \tau_{1r}$ and $\mathbb{E}[\widehat{\tau}_{2r}] = \tau_{2r}$. The third equality follows from $P = 2R$. The fourth equality follows from Equation (3).

**Point 3**

$$
\begin{aligned}
\widehat{\mathbb{V}}_{brs}(\widehat{\tau}) &= \frac{1}{P^2}\sum_{p}\widehat{\tau}_p^2 - \frac{1}{2}\left(\frac{2}{P^2}\sum_{r}\widehat{\tau}_{1r}\widehat{\tau}_{2r} + \frac{\widehat{\tau}^2}{P}\right). \\
&= \frac{1}{2P^2}\sum_{p}(\widehat{\tau}_p - \widehat{\tau})^2 + \frac{1}{2P^2}\sum_{r}(\widehat{\tau}_{1r}^2 + \widehat{\tau}_{2r}^2 - 2\widehat{\tau}_{1r}\widehat{\tau}_{2r}). \\
&= \frac{1}{2}\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) + \frac{1}{2}\widehat{\mathbb{V}}_{pop}(\widehat{\tau}).
\end{aligned}
$$

**Point 4**

$$
\begin{aligned}
\mathbb{E}[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})] &\leq \mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right],\\
\Leftrightarrow (2R-1)\sum_{r=1}^{R}(\tau_{1r}-\tau_{2r})^2 &\leq 2R\sum_{p=1}^{P}(\tau_p-\tau)^2,\\
\Leftrightarrow (2R-1)\sum_{r=1}^{R}(\tau_{1r}^2+\tau_{2r}^2-2\tau_{1r}\tau_{2r}) &\leq 2R\sum_{r=1}^{R}[\tau_{1r}^2-2\tau_{1r}\tau+\tau^2+\tau_{2r}^2-2\tau_{2r}\tau+\tau^2],\\
\Leftrightarrow 0 &\leq \sum_{r=1}^{R}(\tau_{1r}-\tau_{2r})^2+2R\sum_{r=1}^{R}[2\tau_{1r}\tau_{2r}-2(\tau_{1r}+\tau_{2r})\tau+2\tau^2],\\
\Leftrightarrow 0 &\leq \sum_{r=1}^{R}(\tau_{1r}-\tau_{2r})^2+4R\sum_{r=1}^{R}(\tau_{1r}-\tau)(\tau_{2r}-\tau).
\end{aligned}
$$

The second inequality follows from Points 1 and 2 of this lemma. Let $\tau_{\cdot r}=\frac{1}{2}(\tau_{1r}+\tau_{2r})$. Then,

$$
\begin{aligned}
\mathbb{E}[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})] &\leq \mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right],\\
\Leftrightarrow 0 &\leq \sum_{r=1}^{R}\sum_{p=1,2}2(\tau_{pr}-\tau_{\cdot r})^2+4R\sum_{r=1}^{R}(\tau_{1r}-\tau_{\cdot r}+\tau_{\cdot r}-\tau)(\tau_{2r}-\tau_{\cdot r}+\tau_{\cdot r}-\tau),\\
\Leftrightarrow 0 &\leq \sum_{r=1}^{R}\sum_{p=1,2}\frac{1}{2}(\tau_{pr}-\tau_{\cdot r})^2+R\sum_{r=1}^{R}[(\tau_{1r}-\tau_{\cdot r})(\tau_{2r}-\tau_{\cdot r})+(\tau_{\cdot r}-\tau)^2],\\
\Leftrightarrow 0 &\leq \sum_{r=1}^{R}\sum_{p=1,2}\frac{1}{2}(\tau_{pr}-\tau_{\cdot r})^2+R\sum_{r=1}^{R}\left[-\sum_{p=1,2}\frac{1}{2}(\tau_{pr}-\tau_{\cdot r})^2+(\tau_{\cdot r}-\tau)^2\right],\\
\Leftrightarrow \frac{1}{R}\sum_{r=1}^{R}\sum_{p=1,2}\frac{1}{2}(\tau_{pr}-\tau_{\cdot r})^2 &\leq \frac{1}{R-1}\sum_{r=1}^{R}(\tau_{\cdot r}-\tau)^2.
\end{aligned}
$$

This proves inequality a).

Then, if $\frac{1}{R}\sum_{r=1}^{R}\sum_{p=1,2}\frac{1}{2}(\tau_{pr}+\tau_{\cdot r})^2 \leq \frac{1}{R-1}\sum_{r=1}^{R}(\tau_{\cdot r}-\tau)^2$, it follows from Point 3 of the lemma and the previous display that

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})\right] \leq& \frac{1}{2}\,\mathbb{E}\left[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})\right]+\frac{1}{2}\,\mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right]\\
\leq& \frac{1}{2}\,\mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pop}(\widehat{\tau})\right]+\frac{1}{2}\,\mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right]\\
=& \mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{brs}(\widehat{\tau})\right],
\end{aligned}
$$

which proves inequality b).

Similarly, if $\frac{1}{R}\sum_{r=1}^{R}\sum_{p=1,2}\frac{1}{2}(\tau_{pr}+\tau_{\cdot r})^2 \leq \frac{1}{R-1}\sum_{r=1}^{R}(\tau_{\cdot r}-\tau)^2$, it follows from Point 3 of the lemma and the previous display that

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathbb{V}}_{brs}(\widehat{\tau})\right] \leq & \frac{1}{2}\,\mathbb{E}\left[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})\right] + \frac{1}{2}\,\mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right] \\
\leq & \frac{1}{2}\,\mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right] + \frac{1}{2}\,\mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right] \\
= & \,\mathbb{E}\left[\frac{P}{P-1}\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right],
\end{aligned}$$

which proves inequality c).

QED.

## Proof of Lemma 4.2

It follows from Lemma 3.1 that

$$\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P}\omega_p^2 SET_{p,fe}^2\left(\frac{1}{n_{1p}}+\frac{1}{n_{2p}}\right)^2$$

and

$$\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P}\omega_p^2 SET_{p,fe}^2\left(\frac{1}{n_{1p}^2}+\frac{1}{n_{2p}^2}\right).$$

Under Assumption 2, $n_{1p}=n_{2p}=n_p/2$, for all $p$. Then,

$$\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P}\omega_p^2 SET_{p,fe}^2\left(\frac{2}{n_p}+\frac{2}{n_p}\right)^2 = 16\sum_{p=1}^{P}\frac{\omega_p^2 SET_{p,fe}^2}{n_p^2}.$$

Similarly,

$$\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P}\omega_p^2 SET_{p,fe}^2\left(\frac{4}{n_p^2}+\frac{4}{n_p^2}\right) = 8\sum_{p=1}^{P}\frac{\omega_p^2 SET_{p,fe}^2}{n_p^2}.$$

QED.

**Proof of Lemma 4.3**

Let $\overline{Y}_{gp} \equiv \sum_i Y_{igp}/n_{gp}$, $\widehat{Y}_p(1) \equiv \sum_g W_{gp}\overline{Y}_{gp}$, $\widehat{Y}_p(0) \equiv \sum_g (1-W_{gp})\overline{Y}_{gp}$, and $\widehat{Y}(d) \equiv \sum_p \widehat{Y}_p(d)/P$, $d = 0, 1$.

$$\mathbb{E}[\widehat{Y}_p(1)] = \mathbb{E}\left[\sum_g W_{gp}\overline{y}_{gp}(1)\right] = \frac{1}{2}\sum_g \overline{y}_{gp}(1) = \overline{y}_p(1). \tag{9}$$

The second equality follows from Points 2 of Assumption 1. Similarly,

$$\mathbb{E}[\widehat{Y}_p(0)] = \mathbb{E}[\overline{y}_p(0)] \tag{10}$$

$$\mathbb{E}[\widehat{Y}(d)] = \overline{y}(d), \quad \text{ for } d \in \{0, 1\}. \tag{11}$$

Then, one has

$$\begin{aligned}
\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) - \widehat{\mathbb{V}}_{pair}(\widehat{\tau}) &= \frac{8}{n^2}\sum_p SET_p SEU_p \\
&= \frac{8}{n^2}\sum_p \left(\sum_g W_{gp}\sum_i (y_{igp}(1) - \widehat{Y}(1))\right)\left(\sum_g (1-W_{gp})\sum_i (y_{igp}(0) - \widehat{Y}(0))\right) \\
&= \frac{8}{n^2}\sum_p \frac{n_p^2}{4}\left(\sum_g W_{gp}\sum_i \frac{y_{igp}(1)}{n_{gp}} - \widehat{Y}(1)\right)\left(\sum_g (1-W_{gp})\sum_i \frac{y_{igp}(0)}{n_{gp}} - \widehat{Y}(0)\right) \\
&= \frac{2}{P^2}\sum_p \widehat{Y}_p(1)\widehat{Y}_p(0) - \frac{2}{P}\widehat{Y}(1)\widehat{Y}(0)
\end{aligned} \tag{12}$$

The first equality follows from Points 1 and 2 of Lemma 3.1 and Assumption 2. The second equality follows from the definitions of $SET_p$, $SEU_p$, and $\varepsilon_{igp}$. The third equality follows from Point 1 of Assumption 1, and Assumption 2. The fourth equality follows from Assumption 2 and some algebra. Taking the expectation of (12),

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) - \widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right] \\
&= \frac{2}{P^2}\sum_p \left(\mathrm{Cov}(\widehat{Y}_p(1), \widehat{Y}_p(0))\right) + \frac{2}{P^2}\sum_p (\overline{y}_p(1) - \overline{y}(1))(\overline{y}_p(0) - \overline{y}(0)) - \frac{2}{P}\mathrm{Cov}(\widehat{Y}(1), \widehat{Y}(0)) \\
&= \frac{2}{P^2}\sum_p \left(\mathrm{Cov}(\widehat{Y}_p(1), \widehat{Y}_p(0))\right) + \frac{2}{P^2}\sum_p (\overline{y}_p(1) - \overline{y}(1))(\overline{y}_p(0) - \overline{y}(0)) - \frac{2}{P}\mathrm{Cov}\left(\frac{1}{P}\sum_p \widehat{Y}_p(1), \frac{1}{P}\sum_p \widehat{Y}_p(0)\right) \\
&= \frac{2(P-1)}{P^3}\sum_p \left(\mathrm{Cov}(\widehat{Y}_p(1), \widehat{Y}_p(0))\right) + \frac{2}{P^2}\sum_p (\overline{y}_p(1) - \overline{y}(1))(\overline{y}_p(0) - \overline{y}(0)).
\end{aligned}$$

The first equality follows from adding and substracting $\frac{2}{P}\,\mathbb{E}[\widehat{Y}(1)]\,\mathbb{E}[\widehat{Y}(0)]$ and $\frac{2}{P^2}\sum_p \mathbb{E}[\widehat{Y}_p(1)]\,\mathbb{E}[\widehat{Y}_p(0)]$, and from Equations (9), (10) and (11). The third equality follows from Point 3 of Assumption 1. Therefore,

$$\frac{P}{P-1}\,\mathbb{E}\left[\widehat{\mathbb{V}}_{unit}(\widehat{\tau})-\widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right]=\frac{2}{P^2}\sum_p\left(\mathrm{Cov}(\widehat{Y}_p(1),\widehat{Y}_p(0))\right)+\frac{2}{P(P-1)}\sum_p(\overline{y}_p(0)-\overline{y}(0))(\overline{y}_p(1)-\overline{y}(1)). \tag{13}$$

Finally,

$$\begin{aligned}
\mathrm{Cov}\left(\widehat{Y}_p(1),\widehat{Y}_p(0)\right) &= \mathbb{E}[\widehat{Y}_p(1)\widehat{Y}_p(0)]-\mathbb{E}[\widehat{Y}_p(1)]\,\mathbb{E}[\widehat{Y}_p(0)]\\
&=\left(\frac{1}{2}\overline{y}_{1p}(1)\overline{y}_{2p}(0)+\frac{1}{2}\overline{y}_{2p}(1)\overline{y}_{1p}(0)\right)-\left(\frac{1}{2}\sum_g\overline{y}_{gp}(1)\right)\left(\frac{1}{2}\sum_g\overline{y}_{gp}(0)\right)\\
&=\frac{1}{4}\overline{y}_{1p}(1)\overline{y}_{2p}(0)+\frac{1}{4}\overline{y}_{2p}(1)\overline{y}_{1p}(0)-\frac{1}{4}\overline{y}_{1p}(1)\overline{y}_{1p}(0)-\frac{1}{4}\overline{y}_{2p}(1)\overline{y}_{2p}(0)\\
&=\frac{1}{4}\left(\overline{y}_{1p}(1)-\overline{y}_{2p}(1)\right)\left(\overline{y}_{2p}(0)-\overline{y}_{1p}(0)\right)\\
&=-\frac{1}{2}\sum_g\left(\overline{y}_{gp}(0)-\overline{y}_p(0)\right)\left(\overline{y}_{gp}(1)-\overline{y}_p(1)\right)
\end{aligned} \tag{14}$$

The second equality follows from Points 1 and 2 of Assumption 1, and Equations (9) and (10). The third, fourth, and fifth equalities follow after some algebra. The result follows plugging Equation (14) into (13).

QED.

# For Online Publication

## B Extension: stratified experiments with few units per strata

In this section, we perform Monte-Carlo simulations to assess how our results in section 4 extend to stratified experiments where the number of units per strata is larger than two, but still fairly small. Three main findings emerge. First, $t$-tests using stratum-clustered standard errors have nominal size. Second, $t$-tests using standard errors clustered at the unit of randomization level are liberal in regressions with strata fixed effects, but become less liberal as the number of randomization units per strata increases. With 5 units per strata, the empirical size of a 5% level test with UCVE and fixed effects is around 7.8%, while with 10 units per strata it is around 6.6%. Finally, $t$-tests using standard errors clustered at the unit of randomization level are typically conservative in regressions without strata fixed effects.

We draw the potential and observed outcomes from the following data generating process (DGP),

$$Y_{igp} = W_{gp} y_{igp}(1) + (1 - W_{gp}) y_{igp}(0) + \gamma_p, \qquad i = 1, \dots, n_{gp};\ g = 1, \dots, G;\ p = 1, \dots, P, \quad (15)$$

where $y_{igp}(1)$ and $y_{igp}(0)$ are independent and both follow a $\mathcal{N}(0, 1)$ distribution, $\{\gamma_p\}_p \sim$ iid $\mathcal{N}(0, \sigma_\gamma^2)$, and $(y_{igp}(1), y_{igp}(0)) \perp\!\!\!\perp \gamma_p$. We either let $\sigma_\eta = 0$ or $\sigma_\eta = \sqrt{0.1}$. $\sigma_\eta = 0$ corresponds to a model with no stratum common shock, while $\sigma_\eta = \sqrt{0.1}$ corresponds to a model with a shock. We draw potential outcomes once and keep them fixed, so $y_{igp}(1)$, $y_{igp}(0)$ and $\gamma_p$ do not vary across simulations.

Each stratum has $G$ randomization units. We vary $G$ from two to ten. If $G$ is even, then half of the units are randomly assigned to the control and the remaining to the treatment. If $G$ is odd, then $(G + 1)/2$ units are randomly assigned to the control. We also set $n_{gp} = 5$ or $n_{gp} = 100$, and we let $P = 100$.

For each simulation, treatment is randomly assigned to $G/2$ or $(G-1)/2$ units per stratum. We compute $t$-tests based on unit- and stratum-clustered standard errors in regressions of the outcome on the treatment with and without strata fixed effects. We perform 10,000 simulations for each DGP. Table 4 presents the size of the $t$-tests in each DGP.

$t$-tests using stratum-clustered standard errors achieve 5% size for all values of the number of units per strata, $G$. In contrast, $t$-tests based on unit-clustered standard errors in regressions with fixed effects overreject the true null of no treatment effect. These results are in line with Points 1 and 2 of Theorem 4.4, which covered the special case where $G = 2$. $t$-tests based on unit-clustered standard errors in regressions with fixed effects overreject less as the number of units per strata increases from two (column 2) to ten (column 10). Interestingly, it seems that unit-clustered standard errors are approximately equal to $\sqrt{\frac{G-1}{G}}$ times the stratum-clustered standard errors. If $G = 2$, the ratio of those two standard errors is exactly equal to $\sqrt{(2-1)/2} = \sqrt{1/2}$ as shown in Lemma 4.1, but this relationship seems to still hold in expectation for larger values of $G$.

In Panel A, $t$-tests based on unit-clustered standard errors in regressions without fixed effects have the right size. When $\sigma_\eta = 0$, there is no between and within strata heterogeneity in $\overline{y}_{gp}(0)$, so it follows from Point 5 of Theorem 4.4 that in the special case where $G = 2$, $t$-tests based on unit-clustered standard errors in regressions without fixed effects have correct size. Our simulations suggest that this result still holds when $G > 2$. However, in Panel B, $t$-tests using unit-clustered standard errors in regressions without fixed effects are conservative, because there is now between strata heterogeneity in $\overline{y}_{gp}(0)$.

We obtain similar results with five observations per unit of randomization (Panels C and D). The only change is that the DOF correction in regressions with fixed effects makes the stratum-clustered variance estimator slightly conservative and the unit-cluster variance estimator slightly less liberal than in Panels A and B.

Table 4: Size of $t$-test in simulated stratified experiments with small strata

| | Number of units per strata | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| *Panel A. iid standard normal potential outcomes and* $n_{gp} = 100$ | | | | | | | | | |
| UCVE without FE | 0.0342 | 0.0470 | 0.0491 | 0.0511 | 0.0492 | 0.0472 | 0.0532 | 0.0476 | 0.0525 |
| UCVE with FE | 0.1633 | 0.1121 | 0.0907 | 0.0797 | 0.0724 | 0.0667 | 0.0703 | 0.0619 | 0.0663 |
| SCVE without FE | 0.0516 | 0.0504 | 0.0534 | 0.0517 | 0.0514 | 0.0506 | 0.0548 | 0.0513 | 0.0567 |
| SCVE with FE | 0.0510 | 0.0502 | 0.0533 | 0.0516 | 0.0511 | 0.0502 | 0.0547 | 0.0512 | 0.0566 |
| $\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})/\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})$ | 0.7053 | 0.8164 | 0.8703 | 0.8996 | 0.9181 | 0.9320 | 0.9405 | 0.9487 | 0.9559 |
| *Panel B. Stratum-level shock affecting potential outcomes and* $n_{gp} = 100$ | | | | | | | | | |
| UCVE without FE | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| UCVE with FE | 0.1683 | 0.1096 | 0.0901 | 0.0841 | 0.0730 | 0.0683 | 0.0702 | 0.0633 | 0.0629 |
| SCVE without FE | 0.0538 | 0.0520 | 0.0540 | 0.0560 | 0.0522 | 0.0513 | 0.0546 | 0.0533 | 0.0532 |
| SCVE with FE | 0.0529 | 0.0516 | 0.0535 | 0.0558 | 0.0520 | 0.0512 | 0.0545 | 0.0533 | 0.0532 |
| $\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})/\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})$ | 0.7053 | 0.8166 | 0.8722 | 0.8980 | 0.9178 | 0.9309 | 0.9409 | 0.9487 | 0.9549 |
| *Panel C. iid standard normal potential outcomes and* $n_{gp} = 5$ | | | | | | | | | |
| UCVE without FE | 0.0696 | 0.0505 | 0.0518 | 0.0486 | 0.0487 | 0.0514 | 0.0523 | 0.0550 | 0.0534 |
| UCVE with FE | 0.1496 | 0.1019 | 0.0807 | 0.0722 | 0.0697 | 0.0648 | 0.0635 | 0.0631 | 0.0653 |
| SCVE without FE | 0.0560 | 0.0522 | 0.0514 | 0.0509 | 0.0531 | 0.0511 | 0.0530 | 0.0545 | 0.0553 |
| SCVE with FE | 0.0425 | 0.0448 | 0.0462 | 0.0475 | 0.0488 | 0.0483 | 0.0503 | 0.0517 | 0.0520 |
| $\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})/\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})$ | 0.7053 | 0.8166 | 0.8697 | 0.8977 | 0.9183 | 0.9308 | 0.9403 | 0.9494 | 0.9546 |
| *Panel D. Stratum-level shock affecting potential outcomes and* $n_{gp} = 5$ | | | | | | | | | |
| UCVE without FE | 0.0296 | 0.0152 | 0.0173 | 0.0140 | 0.0191 | 0.0167 | 0.0199 | 0.0248 | 0.0219 |
| UCVE with FE | 0.1422 | 0.1032 | 0.0830 | 0.0752 | 0.0680 | 0.0669 | 0.0603 | 0.0632 | 0.0611 |
| SCVE without FE | 0.0507 | 0.0544 | 0.0523 | 0.0542 | 0.0517 | 0.0542 | 0.0504 | 0.0530 | 0.0517 |
| SCVE with FE | 0.0408 | 0.0473 | 0.0479 | 0.0499 | 0.0478 | 0.0512 | 0.0480 | 0.0505 | 0.0489 |
| $\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})/\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})$ | 0.7053 | 0.8168 | 0.8692 | 0.9008 | 0.9183 | 0.9324 | 0.9415 | 0.9481 | 0.9555 |

The table shows the size of $t$-tests based on unit- and stratum-clustered standard errors in regressions with and without stratum fixed effects. Across simulations, we vary the number of randomization units per strata from two to ten ($G = 2, \ldots, 10$); we vary the number of observations per randomization unit to either $n_{gp} = 5$ or $n_{gp=}100$; and we set the number of strata to $P = 100$. For each value of $G$, we simulated 10,000 samples from the following data generating processes: independent and identically distributed (iid) standard normal potential outcomes in Panel A, and a model with an additive stratum-level shock affecting both potential outcomes in Panel B. UCVE and SCVE stand for unit- and stratum-clustered variance estimators, respectively. FE stands for strata fixed effects. $\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})/\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})$ is the average across simulations of the ratio of standard errors clustering at the unit and stratum levels in regressions with stratum fixed effects.

# C Articles in our survey of paired or small strata experiments

Table 5: Paired experiments and stratified experiments with small strata

| Reference | Search source |
|---|---|
| Paired Experiments | |
| Ashraf et al. (2006) | AEA registry |
| Panagopoulos and Green (2008) | AEA registry |
| Banerjee et al. (2015) | AEJ: Applied |
| Crépon et al. (2015) | AEJ: Applied |
| Beuermann et al. (2015)[1] | AEJ: Applied |
| Fryer Jr et al. (2016) | AEA registry |
| Glewwe et al. (2016) | AEA registry |
| Bruhn et al. (2016) | AEJ: Applied |
| Fryer Jr (2017) | AEA registry |
| Small-strata experiments | |
| Attanasio et al. (2015) | AEJ: Applied |
| Angelucci et al. (2015) | AEJ: Applied |
| Ambler et al. (2015) | AEJ: Applied |
| Björkman Nyqvist et al. (2017) | AEJ: Applied |
| Banerji et al. (2017) | AEJ: Applied |
| Lafortune et al. (2018) | AEJ: Applied |
| Somville and Vandewalle (2018) | AEJ: Applied |

The table presents economics papers that have conducted paired experiments or stratified experiments with ten or less units per strata. We searched the AEJ: Applied Economics for papers published in 2014-2018 and using the words "random" and "experiment" in the abstract, title, keywords, or main text. Four of those papers had conducted a paired randomized experiment and seven had conducted a stratified experiment with ten units or less per stratum. We also searched the AEA's registry website for randomized controlled trials (`https://www.socialscienceregistry.org`). We looked at all completed projects, whose randomization method included the prefix "pair" and that had either a working or a published paper. Thus, we found five more papers that had conducted a paired randomized experiment.

# D Online appendix proofs

## Proof of Lemma 3.1

### Point 1

First, we introduce the formulas for the PCVE and UCVE in a general linear regression. Let $\epsilon_{igp}$ be the residual from the regression of $Y_{igp}$ on a $K$-vector of covariates $\boldsymbol{X}_{igp}$, and $\boldsymbol{X}$ the $(n \times K)$ matrix whose rows are $\boldsymbol{X}'_{igp}$. The PCVE of the OLS estimator, $\widehat{\boldsymbol{\beta}}$, is defined as follows (Liang and Zeger (1986), Abadie et al. (2017))

$$\widehat{\mathbb{V}}_{pair}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left( \sum_{p=1}^{P} \left( \sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} \epsilon_{igp} \boldsymbol{X}_{igp} \right) \left( \sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} \epsilon_{igp} \boldsymbol{X}_{igp} \right)' \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{16}$$

The UCVE of the OLS estimator, $\widehat{\boldsymbol{\beta}}$, is defined as follows

$$\widehat{\mathbb{V}}_{unit}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left( \sum_{p=1}^{P} \sum_{g=1}^{2} \left( \sum_{i=1}^{n_{gp}} \epsilon_{igp} \boldsymbol{X}_{igp} \right) \left( \sum_{i=1}^{n_{gp}} \epsilon_{igp} \boldsymbol{X}_{igp} \right)' \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{17}$$

Subtract from Equation (1) the average outcome in the population $\overline{Y} \equiv \frac{1}{n} \sum_p \sum_g \sum_i Y_{igp} = \widehat{\alpha} + \widehat{\tau}\overline{W}_{gp} + \overline{\varepsilon}$, where $\overline{W} \equiv \frac{1}{n} \sum_p \sum_g \sum_i W_{gp}$, and $\overline{\varepsilon} \equiv \frac{1}{n} \sum_p \sum_g \sum_i \varepsilon_{igp} = 0$ by construction. Then,

$$Y_{igp} - \overline{Y} = \widehat{\tau}(W_{gp} - \overline{W}) + \varepsilon_{igp}. \tag{18}$$

Apply Equation (16) to the residuals and covariates of the regression defined by Equation (18).[13] Then,

$$\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = \frac{\sum_p \left[ \sum_g (W_{gp} - \overline{W}) \sum_i \varepsilon_{igp} \right]^2}{\left[ \sum_p \sum_g \sum_i (W_{gp} - \overline{W})^2 \right]^2}. \tag{19}$$

[13] The clustered variance estimator using the residuals from the demeaned formula is equivalent to the clustered variance estimator including an intercept as in Equation (1) (Cameron and Miller, 2015).

The numerator of $\widehat{\mathbb{V}}_{pair}(\widehat{\tau})$ equals

$$\begin{aligned}
\sum_p \left[\sum_g (W_{gp} - \overline{W}) \sum_i \varepsilon_{igp}\right]^2 &= \sum_p \left[(1-\overline{W})SET_p - \overline{W}SEU_p\right]^2 \\
&= \sum_p \left[\frac{C}{n}SET_p - \frac{T}{n}SEU_p\right]^2. \qquad (20)
\end{aligned}$$

The first equality follows from the definition of $SET_p$ and $SEU_p$. The second equality follows from the definition of $T$ and $C$.

The denominator of $\widehat{\mathbb{V}}_{pair}(\widehat{\tau})$ equals

$$\begin{aligned}
\left[\sum_p \sum_g \sum_i (W_{gp} - \overline{W})^2\right]^2 &= \left[\sum_p \sum_g (W_{gp} - \overline{W})^2 n_{gp}\right]^2 \\
&= \left[(1-\overline{W})^2 \sum_p T_p + \overline{W}^2 \sum_p C_p\right]^2 \\
&= \left[\frac{C^2}{n^2}T + \frac{T^2}{n^2}C\right]^2 \\
&= \left[\frac{CT}{n}\right]^2. \qquad (21)
\end{aligned}$$

The first equality follows from $(W_{gp} - \overline{W})$ being constant across units. The second equality follows from the definition of $T_p$ and $C_p$. The third equality follows from the definition of $T$ and $C$.

Then, combining Equations (19), (20) and (21),

$$\begin{aligned}
\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) &= \frac{\sum_p \left[\frac{C}{n}SET_p - \frac{T}{n}SEU_p\right]^2}{\left[\frac{CT}{n}\right]^2} \\
&= \sum_p \left[\frac{SET_p}{T} - \frac{SEU_p}{C}\right]^2.
\end{aligned}$$

**Point 2**

Apply Equation (17) to the residuals and covariates of the regression defined by Equation (18). Then,

$$\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) = \frac{\sum_p \sum_g \left[(W_{gp} - \overline{W}) \sum_i \varepsilon_{igp}\right]^2}{\left[\sum_p \sum_g \sum_i (W_{gp} - \overline{W})^2\right]^2}. \qquad (22)$$

The numerator of $\widehat{\mathbb{V}}_{unit}(\widehat{\tau})$ equals

$$\begin{aligned}\sum_p \sum_g \left[(W_{gp} - \overline{W}) \sum_i \varepsilon_{igp}\right]^2 &= \sum_p \sum_g (W_{gp} - \overline{W})^2 \left(\sum_i \varepsilon_{igp}\right)^2 \\ &= \sum_p \left[(1-\overline{W})^2 SET_p^2 + \overline{W}^2 SEU_p^2\right] \\ &= \sum_p \left[\frac{C^2}{n^2} SET_p^2 + \frac{T^2}{n^2} SEU_p^2\right]. \end{aligned} \tag{23}$$

The second equality follows from the defintion of $SET_p$ and $SEU_p$. The third equality follows from the definition of $T$ and $C$. Then, combining Equations (21), (22) and (23),

$$\begin{aligned}\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) &= \frac{\sum_p \left[\frac{C^2}{n^2} SET_p^2 + \frac{T^2}{n^2} SEU_p^2\right]}{\left[\frac{CT}{n}\right]^2} \\ &= \sum_p \left[\frac{SET_p^2}{T^2} + \frac{SEU_p^2}{C^2}\right].\end{aligned}$$

**Point 3**

First, consider Equation (2) and, for each pair $p$, take averages across units to obtain the following

$$\overline{Y}_p = \widehat{\tau}_{fe} \overline{W}_p + \widehat{\gamma}_p + \overline{u}_p, \tag{24}$$

where $\overline{Y}_p = \frac{1}{n_p} \sum_{g=1}^2 \sum_{i=1}^{n_{gp}} Y_{igp}$, $\overline{W}_p = \frac{1}{n_p} \sum_{g=1}^2 \sum_{i=1}^{n_{gp}} W_{gp} = \frac{1}{n_p} \sum_{g=1}^2 W_{gp} n_{gp} = \frac{T_p}{n_p}$. Substract Equation (24) from Equation (2) to remove the fixed effect $\widehat{\gamma}_p$ (note that $\delta_{igp'} = 0$ for all $p' \neq p$)

$$Y_{igp} - \overline{Y}_p = \widehat{\tau}_{fe}(W_{gp} - \overline{W}_p) + u_{igp} - \overline{u}_p. \tag{25}$$

Given that $\{u_{ijp'}\}$ is an OLS residual, then it is orthogonal to any regressor by construction. In particular, $\{u_{ijp'}\}$ is orthogonal to the pair-$p$ fixed effect indicator $\{\delta_{igp}\}$,

$$\sum_{p'=1}^{P} \sum_{g=1}^{2} \sum_{i=1}^{n_{jp'}} u_{ijp'} \delta_{igp} = 0.$$

By the definition of $\delta_{igp}$, $\delta_{igp} = 1$ if unit $i$ belongs to pair $p$, and $\delta_{igp} = 0$ if unit $i$ does not belong to pair $p$, so that the above equation reduces to

$$\sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} u_{igp} = 0, \tag{26}$$

which implies that for all $p$ the within-pair residual average is zero

$$\overline{u}_p = \frac{1}{n_p} \sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} u_{igp} = 0.$$

Equation (25) then becomes a regression with one covariate and the same residuals as in Equation (2):

$$Y_{igp} - \overline{Y}_p = \widehat{\tau}_{fe}(W_{gp} - \overline{W}_p) + u_{igp}. \tag{27}$$

Now, apply Equation (16) to obtain the PCVE of $\widehat{\tau}_{fe}$, which simplifies given that there's only one regressor $W_{gp} - \overline{W}_p$,[14] to

$$\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \frac{\left[\sum_{p=1}^{P} \left(\sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} u_{igp}(W_{gp} - \overline{W}_p)\right)^2\right]}{\left(\sum_{p=1}^{P} \sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} (W_{gp} - \overline{W}_p)^2\right)^2} \tag{28}$$

[14]The clustered variance estimator using the residuals from the deviations-from-means formula (Equation (27)) is equivalent to the clustered variance estimator including the full set of pair dummies as in Equation (2) (Cameron and Miller, 2015).

The denominator of $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ equals

$$
\begin{aligned}
\left[\sum_p \sum_g \sum_i (W_{gp} - \overline{W}_p)^2\right]^2 &= \left[\sum_p \sum_g (W_{gp} - \overline{W}_p)^2 n_{gp}\right]^2 \\
&= \left[\sum_p [T_p(1-\overline{W}_p)^2 + C_p \overline{W}_p^2]\right]^2 \\
&= \left[\sum_p \left(T_p \frac{C_p^2}{n_p^2} + C_p \frac{T_p^2}{n_p^2}\right)\right]^2 \\
&= \left[\sum_p \frac{T_p C_p}{n_p}\right]^2 \\
&= \left[\sum_p \frac{n_{1p} n_{2p}}{n_{1p} + n_{2p}}\right]^2 \\
&= \left[\sum_p (n_{1p}^{-1} + n_{2p}^{-1})^{-1}\right]^2 .
\end{aligned} \tag{29}
$$

The second and third equalities follow from the definitions of $T_p$ and $C_p$. The numerator of $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ is equal to

$$
\begin{aligned}
\sum_{p=1}^{P}\left(\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} u_{igp}(W_{gp} - \overline{W}_p)\right)^2 &= \sum_{p=1}^{P}\left(\sum_{g=1}^{2}(W_{gp} - \overline{W}_p)\sum_{i=1}^{n_{gp}} u_{igp}\right)^2 \\
&= \sum_{p=1}^{P}\left(-\overline{W}_p(SET_{p,fe} + SEU_{p,fe}) + SET_{p,fe}\right)^2 \\
&= \sum_{p=1}^{P}\left(SET_{p,fe}\right)^2 ,
\end{aligned} \tag{30}
$$

where $SET_{p,fe} + SEU_{p,fe} = \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} u_{igp} = 0$ from Equation (26). Therefore, combining Equations (28), (29) and (30),

$$
\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \frac{\sum_{p=1}^{P} SET_{p,fe}^2}{\left[\sum_{p=1}^{P}\left(n_{1p}^{-1} + n_{2p}^{-1}\right)^{-1}\right]^2} = \sum_{p=1}^{P} \omega_p^2 SET_{p,fe}^2 \left(\frac{1}{n_{1p}} + \frac{1}{n_{2p}}\right)^2 . \tag{31}
$$

**Point 4**

Consider the deviations-from-means formula of the fixed effects regression from Equation (27)

$$Y_{igp} - \overline{Y}_p = \widehat{\tau}_{fe}(W_{gp} - \overline{W}_p) + u_{igp}.$$

Apply the definition of the UCVE from Equation (17), which simplifies given that there's only one regressor $W_{gp} - \overline{W}_p$, to

$$\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe}) = \frac{\left[\sum_{p=1}^{P}\sum_{g=1}^{2}\left(\sum_{i=1}^{n_{gp}} u_{igp}(W_{gp} - \overline{W}_p)\right)^2\right]}{\left(\sum_{p=1}^{P}\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}(W_{gp} - \overline{W}_p)^2\right)^2}. \tag{32}$$

The numerator of $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$ equals

$$\begin{aligned}
\sum_{p=1}^{P}\sum_{g=1}^{2}\left(\sum_{i=1}^{n_{gp}} u_{igp}(W_{gp} - \overline{W}_p)\right)^2 &= \sum_{p=1}^{P}\sum_{g=1}^{2}(W_{gp} - \overline{W}_p)^2\left(\sum_{i=1}^{n_{gp}} u_{igp}\right)^2 \\
&= \sum_{p=1}^{P}\left((1 - \overline{W}_p)^2 SET_{p,fe}^2 + \overline{W}_p^2 SEU_{p,fe}^2\right) \\
&= \sum_{p=1}^{P} SET_{p,fe}^2\left(\frac{C_p^2}{n_p^2} + \frac{T_p^2}{n_p^2}\right) \\
&= \sum_{p=1}^{P} \frac{C_p^2 T_p^2}{n_p^2} SET_{p,fe}^2\left(\frac{1}{T_p^2} + \frac{1}{C_p^2}\right) \\
&= \sum_{p=1}^{P}(n_{1p}^{-1} + n_{2p}^{-1})^{-2} SET_{p,fe}^2\left(\frac{1}{n_{1p}^2} + \frac{1}{n_{2p}^2}\right).
\end{aligned} \tag{33}$$

The second equality follows from the definitions of $SET_{p,fe}$ and $SEU_{p,fe}$. The third equality follows from Equation (26), i.e., $SET_{p,fe} + SEU_{p,fe} = \sum_g \sum_i u_{igp} = 0$, for all $p$, so $SET_{p,fe}^2 = SEU_{p,fe}^2$, and the definitions of $T_p$ and $C_p$.

Then, combining Equations (29), (32) and (33),

$$\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P} \omega_p^2 SET_{p,fe}^2\left(\frac{1}{n_{1p}^2} + \frac{1}{n_{2p}^2}\right). \tag{34}$$

QED.

## Proof of Theorem 4.4

We start by proving a few auxiliary results that will be useful in the proof of Theorem 4.4. Note that, for all $p$,

$$\begin{aligned}
\mathbb{E}\left[|\widehat{\tau}_p|^{2+\epsilon}\right]^{1/(2+\epsilon)} &= \mathbb{E}\left[\left|\widehat{Y}_p(1)-\widehat{Y}_p(0)\right|^{2+\epsilon}\right]^{1/(2+\epsilon)} \\
&\leq \left(\mathbb{E}\left[\left|\widehat{Y}_p(1)\right|^{2+\epsilon}\right]\right)^{1/(2+\epsilon)} + \left(\mathbb{E}\left[\left|\widehat{Y}_p(0)\right|^{2+\epsilon}\right]\right)^{1/(2+\epsilon)} \\
&= \left(\mathbb{E}\left[\left|\sum_g W_{gp}\overline{y}_{gp}(1)\right|^{2+\epsilon}\right]\right)^{1/(2+\epsilon)} + \left(\mathbb{E}\left[\left|\sum_g (1-W_{gp})\overline{y}_{gp}(0)\right|^{2+\epsilon}\right]\right)^{1/(2+\epsilon)} \\
&\leq \sum_g \left(\mathbb{E}\left[\left|W_{gp}\overline{y}_{gp}(1)\right|^{2+\epsilon}\right]\right)^{1/(2+\epsilon)} + \sum_g \left(\mathbb{E}\left[\left|(1-W_{gp})\overline{y}_{gp}(0)\right|^{2+\epsilon}\right]\right)^{1/(2+\epsilon)} \\
&= \sum_g \left(\mathbb{E}[W_{gp}]\left|\overline{y}_{gp}(1)\right|^{2+\epsilon}\right)^{1/(2+\epsilon)} + \sum_g \left(\mathbb{E}[1-W_{gp}]\left|\overline{y}_{gp}(0)\right|^{2+\epsilon}\right)^{1/(2+\epsilon)} \\
&= \sum_g \left(\frac{1}{2}\left|\overline{y}_{gp}(1)\right|^{2+\epsilon}\right)^{1/(2+\epsilon)} + \sum_g \left(\frac{1}{2}\left|\overline{y}_{gp}(0)\right|^{2+\epsilon}\right)^{1/(2+\epsilon)} \\
&\leq 4\left(\frac{1}{2}M\right)^{1/(2+\epsilon)} < +\infty. \qquad (35)
\end{aligned}$$

The first equality follows from the denifition of $\widehat{\tau}_p$. The first inequality follows from Minkowski's inequality. The third line follows from the definitions of $\widehat{Y}_p(1)$ and $\widehat{Y}_p(0)$. The fourth line follows from Minkowski's inequality. The fifth line follows from $W_{gp}$ being a binary variable. The sixth line follows from Point 2 of Assumption 1. The seventh line follows from Point 1 of Assumption 3.

Using a similar reasoning, one can show that

$$\mathbb{E}\left[\left|\widehat{Y}_p(d)\right|^{2+\epsilon}\right] \leq M_1 < +\infty. \qquad (36)$$

for all $d$ and $p$ and for some $M_1 > 0$.

By Equation (35), $\mathbb{E}\left[|\widehat{\tau}_p|\right]$ is bounded uniformly in $p$, and by Point 3 of Assumption 1, $(\widehat{\tau}_p)_{p=1}^{+\infty}$ is an independent sequence of random variables, so that

$$\widehat{\tau} = \frac{1}{P}\sum_p \widehat{\tau}_p \xrightarrow{\mathbb{P}} \lim_{P\to+\infty}\frac{1}{P}\sum_p \mathbb{E}[\widehat{\tau}_p] = \lim_{P\to+\infty}\frac{1}{P}\sum_p \tau_p = \lim_{P\to+\infty}\tau \qquad (37)$$

by the SLLN in Lemma 1 in Liu (1988), the fact that almost sure convergence implies convergence in probability, and Point 2 of Assumption 3.

**Point 1**

Note that by Point 3 of Assumption 1, $\widehat{\tau}-\tau=\widehat{\tau}-\mathbb{E}[\widehat{\tau}]=\sum_p(\widehat{\tau}_p-\mathbb{E}[\widehat{\tau}_p])/P$ is a sum of independent random variables $(\widehat{\tau}_p-\mathbb{E}[\widehat{\tau}_p])_{p=1}^P$ with mean zero and with a finite variance by Equation (35). As $\sum_{p=1}^P \mathbb{E}[|\widehat{\tau}_p-\tau_p|^{2+\epsilon}/S_P^{2+\epsilon}]\to 0$ for some $\epsilon>0$ (by Point 3 of Assumption 3), then, by Lyapunov's central limit theorem, $(\widehat{\tau}-\tau)/(S_P/P)=\sum_p(\widehat{\tau}_p-\tau_p)/S_P \xrightarrow{d} \mathcal{N}(0,1)$ as $P\to\infty$, where $S_P^2=\sum_{p=1}^P \mathbb{V}(\widehat{\tau}_p)=P^2\mathbb{V}(\widehat{\tau})$. Therefore,

$$(\widehat{\tau}-\tau)/\sqrt{\mathbb{V}(\widehat{\tau})} \xrightarrow{d} \mathcal{N}(0,1). \tag{38}$$

Then,

$$\begin{aligned}
P\widehat{\mathbb{V}}_{pair}(\widehat{\tau})-P\mathbb{V}(\widehat{\tau}) &= \sum_{p=1}^P \frac{\widehat{\tau}_p^2}{P}-\widehat{\tau}^2-\sum_{p=1}^P\frac{\mathbb{V}(\widehat{\tau}_p)}{P}\\
&= \sum_{p=1}^P \frac{\widehat{\tau}_p^2}{P}-\widehat{\tau}^2-\sum_{p=1}^P\frac{\mathbb{E}[\widehat{\tau}_p^2]-\mathbb{E}[\widehat{\tau}_p]^2}{P}\\
&= \sum_{p=1}^P \frac{\widehat{\tau}_p^2-\mathbb{E}[\widehat{\tau}_p^2]}{P}-\widehat{\tau}^2+\sum_{p=1}^P\frac{\tau_p^2}{P} \qquad (39)\\
&\xrightarrow{\mathbb{P}} \lim_{P\to+\infty}\frac{1}{P}\sum_{p=1}^P(\tau_p-\tau)^2 \qquad (40)
\end{aligned}$$

The first equality follows from Equations (3) and (7). The third equality follows from $\mathbb{E}[\widehat{\tau}_p]=\tau_p$. Let's consider each of the terms in Equation (39). As $P\to\infty$, by Lemma 1 in Liu (1988), $\sum_{p=1}^P \frac{\widehat{\tau}_p^2-\mathbb{E}[\widehat{\tau}_p^2]}{P} \xrightarrow{\mathbb{P}} 0$, by Equation (35), Point 3 of Assumption 1, and the fact almost sure convergence implies convergence in probability. Then, $\widehat{\tau}^2 \xrightarrow{\mathbb{P}} \lim_{P\to+\infty}\tau^2$ by Equation (37) and the continuous mapping theorem (CMT). Equation (40) follows from these facts, and from Point 2 of Assumption 3.

Given Equation (40), Point 2 of Assumption 3, the Slutsky Lemma and the CMT, as $P\to\infty$,

$$\frac{\widehat{\tau}-\tau}{\sqrt{\widehat{\mathbb{V}}_{pair}(\widehat{\tau})}}=\frac{\widehat{\tau}-\tau}{\sqrt{\mathbb{V}(\widehat{\tau})}}\sqrt{\frac{P\mathbb{V}(\widehat{\tau})}{P\widehat{\mathbb{V}}_{pair}(\widehat{\tau})}} \xrightarrow{d} \mathcal{N}(0,\sigma_{pair}^2). \tag{41}$$

Finally, by Lemma 4.1, $\widehat{\mathbb{V}}_{pair}(\widehat{\tau})=\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$, and by Assumption 2, $\widehat{\tau}=\widehat{\tau}_{fe}$.

**Point 2**

$$
\begin{aligned}
P\widehat{\mathbb{V}}_{pop}(\widehat{\tau}) - P\mathbb{V}(\widehat{\tau}) &= \frac{1}{P}\sum_{r=1}^{R}[\widehat{\tau}_{1r}^2 - 2\widehat{\tau}_{1r}\widehat{\tau}_{2r} + \widehat{\tau}_{2r}^2] - \frac{1}{P}\sum_{p=1}^{P}\mathbb{V}(\widehat{\tau}_p) \\
&= \frac{1}{P}\sum_{p=1}^{P}\widehat{\tau}_p^2 - \frac{2}{P}\sum_{r=1}^{R}\widehat{\tau}_{1r}\widehat{\tau}_{2r} - \frac{1}{P}\sum_{p=1}^{P}[\mathbb{E}(\widehat{\tau}_p^2) - \tau_p^2] \\
&= \sum_{p=1}^{P}\frac{\widehat{\tau}_p^2 - \mathbb{E}[\widehat{\tau}_p^2]}{P} - \frac{1}{R}\sum_{r=1}^{R}\widehat{\tau}_{1r}\widehat{\tau}_{2r} + \frac{1}{P}\sum_{r=1}^{R}(\tau_{1r}^2 + \tau_{2r}^2) \\
&\xrightarrow{\mathbb{P}} \lim_{P\to+\infty}\frac{1}{P}\sum_{r=1}^{R}(\tau_{1r} - \tau_{2r})^2
\end{aligned}
\tag{42}
$$

The second equality follows from the properties of the variance. As $P \to \infty$, by Lemma 1 in Liu (1988), $\sum_{p=1}^{P}\frac{\widehat{\tau}_p^2 - \mathbb{E}[\widehat{\tau}_p^2]}{P} \xrightarrow{\mathbb{P}} 0$. Likewise, as $R = P/2 \to \infty$, by Lemma 1 in Liu (1988), $\sum_{r=1}^{R}\widehat{\tau}_{1r}\widehat{\tau}_{2r}/R - \sum_{r=1}^{R}\tau_{1r}\tau_{2r}/R \xrightarrow{\mathbb{P}} 0$, because $\mathbb{E}[|\widehat{\tau}_{1r}\widehat{\tau}_{2r}|^{1+\epsilon/2}]$ is uniformly bounded in $r$ by Equation (35) and the Cauchy-Schwarz inequality, $(\widehat{\tau}_{1r}\widehat{\tau}_{2r})_{r=1}^{+\infty}$ is a sequence of independent random variables by Point 3 of Assumption 1, and $\mathbb{E}(\widehat{\tau}_{1r}\widehat{\tau}_{2r}) = \mathbb{E}(\widehat{\tau}_{1r})\,\mathbb{E}(\widehat{\tau}_{2r}) = \tau_{1r}\tau_{2r}$. Finally, the convergence arrow follows from Point 2 of Assumption 3 and some algebra.

The result follows from Equations (38) and (42) and a reasoning similar to that used to prove Equation (41).

**Point 3**

$$
\begin{aligned}
P\widehat{\mathbb{V}}_{bsr}(\widehat{\tau}) - P\mathbb{V}(\widehat{\tau}) &= \frac{1}{2}P(\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) - \mathbb{V}(\widehat{\tau})) + \frac{1}{2}P(\widehat{\mathbb{V}}_{pop}(\widehat{\tau}) - \mathbb{V}(\widehat{\tau})) \\
&\xrightarrow{\mathbb{P}} \frac{1}{2}\lim_{P\to+\infty}\frac{1}{P}\sum_{p=1}^{P}(\tau_p - \tau)^2 + \frac{1}{2}\lim_{P\to+\infty}\frac{1}{P}\sum_{r=1}^{R}(\tau_{1r} - \tau_{2r})^2.
\end{aligned}
$$

The first equality follows from Point 3 of Lemma 4.1. The convergence arrow follows from Equations (40) and (42). The result follows from the previous display, Equation (38), and a reasoning similar to that used to prove Equation (41).

**Point 4**

By Lemma 4.2, $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = 2\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$, so given Point 1 of this theorem, the result follows.

**Point 5**

$$
\begin{aligned}
&P\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) - P\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) \\
&= \frac{2}{P}\sum_p \widehat{Y}_p(1)\widehat{Y}_p(0) - 2\frac{1}{P}\sum_p \widehat{Y}_p(1)\frac{1}{P}\sum_p \widehat{Y}_p(0) \\
&\xrightarrow{\mathbb{P}} 2\lim_{P\to+\infty}\left\{\frac{1}{P}\sum_p \mathbb{E}[\widehat{Y}_p(1)\widehat{Y}_p(0)] - \mathbb{E}[\widehat{Y}(1)]\,\mathbb{E}[\widehat{Y}(0)]\right\} \\
&= 2\lim_{P\to+\infty}\frac{1}{P}\sum_p\left\{\left(\overline{y}_p(0)-\overline{y}(0)\right)\left(\overline{y}_p(1)-\overline{y}(1)\right) - \frac{1}{2}\sum_g(\overline{y}_{gp}(0)-\overline{y}_p(0))(\overline{y}_{gp}(1)-\overline{y}_p(1))\right\}. \quad (43)
\end{aligned}
$$

The first equality follows from Equation (12). The convergence arrow follows from the fact $\mathbb{E}\left[\left|\widehat{Y}_p(1)\widehat{Y}_p(0)\right|^{1+\epsilon/2}\right]$ is bounded uniformly in $p$ by Equation (36) and the Cauchy-Schwarz inequality, from the fact $\mathbb{E}\left[\left|\widehat{Y}_p(d)\right|^{1+\epsilon/2}\right]$ is also bounded uniformly in $p$, from Point 3 of Assumption 1, from the SLLN in Lemma 1 in Liu (1988), from the CMT, and from Point 2 of Assumption 3. The last equality follows from the same steps as those used to prove Lemma 4.3. The result follows from Equations (43), (40), and (38), and a reasoning similar to that used to prove Equation (41).

**Point 6a**

$$
\begin{aligned}
&\sigma^2_{pair} \leq \sigma^2_{pop}, \\
\Leftrightarrow &\lim_{P\to+\infty}\frac{1}{R}\sum_{r=1}^{R}(\tau_{1r}-\tau_{2r})^2 \leq \lim_{P\to+\infty}\frac{1}{R}\sum_{p=1}^{P}(\tau_p-\tau)^2, \\
\Leftrightarrow &\lim_{P\to+\infty}\frac{1}{R}\sum_{r=1}^{R}(\tau_{1r}^2+\tau_{2r}^2-2\tau_{1r}\tau_{2r}) \leq \lim_{P\to+\infty}\frac{1}{R}\sum_{r=1}^{R}[\tau_{1r}^2+\tau_{2r}^2-2(\tau_{1r}+\tau_{2r})\tau+2\tau^2], \\
\Leftrightarrow\ &0 \leq \lim_{P\to+\infty}\frac{1}{R}\sum_{r=1}^{R}[2\tau_{1r}\tau_{2r}-2(\tau_{1r}+\tau_{2r})\tau+2\tau^2], \\
\Leftrightarrow\ &0 \leq \lim_{P\to+\infty}\frac{1}{R}\sum_{r=1}^{R}(\tau_{1r}-\tau)(\tau_{2r}-\tau),
\end{aligned}
$$

Then, $\sigma^2_{pair} \leq \sigma^2_{bsr} \leq \sigma^2_{pop} \Leftrightarrow \sigma^2_{pair} \leq \sigma^2_{pop}$.

Point 6b is straightforward so we do not prove it.

# E Results when randomization units do not all have the same number of observations.

In this subsection, we extend Lemma 4.2 without requiring Assumption 2.

**Lemma E.1** (Ratio of the UCVE and PCVE with fixed effects and when Assumption 2 fails)**.** *$\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \sum_p m_p \zeta_p$, where, for all $p$, $\frac{1}{2} \leq m_p = \left(\frac{n_{1p}}{n_p}\right)^2 + \left(\frac{n_{2p}}{n_p}\right)^2 \leq 1$, $\zeta_p \geq 0$ and $\sum_p \zeta_p = 1$. Therefore, $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) \in \left[\frac{1}{2}, 1\right]$.*

**Proof of Lemma E.1**

Take the ratio between $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$ and $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ from Equations (31) and (34),

$$
\begin{aligned}
\frac{\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})}{\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})} &= \frac{\sum_{p=1}^{P} \omega_p^2 SET_{p,fe}^2 \left(\frac{1}{n_{1p}^2} + \frac{1}{n_{2p}^2}\right)}{\frac{\sum_{p=1}^{P} SET_{p,fe}^2}{\left[\sum_{p=1}^{P} \left(n_{1p}^{-1} + n_{2p}^{-1}\right)^{-1}\right]^2}} \\
&= \frac{\sum_{p=1}^{P} (n_{1p}^{-1} + n_{2p}^{-1})^{-2} SET_{p,fe}^2 \left(\frac{1}{n_{1p}^2} + \frac{1}{n_{2p}^2}\right)}{\sum_{p=1}^{P} SET_{p,fe}^2} \\
&= \sum_{p=1}^{P} m_p \zeta_p,
\end{aligned}
$$

where $\zeta_p = \frac{SET_{p,fe}^2}{\sum_{p=1}^{P} SET_{p,fe}^2} \geq 0$ and $m_p = (n_{1p}^{-1} + n_{2p}^{-1})^{-2} \left(\frac{1}{n_{1p}^2} + \frac{1}{n_{2p}^2}\right)$. The second equality follows from the definition of $\omega_p$. Clearly, $\sum_{p=1}^{P} \zeta_p = 1$.

Now,

$$
\begin{aligned}
m_p &= \frac{1}{\left(\frac{1}{n_{1p}} + \frac{1}{n_{2p}}\right)^2} \left(\frac{1}{n_{1p}^2} + \frac{1}{n_{2p}^2}\right) \\
&= \frac{1}{\left(1 + \frac{n_{1p}}{n_{2p}}\right)^2} + \frac{1}{\left(\frac{n_{2p}}{n_{1p}} + 1\right)^2} \\
&= \frac{n_{2p}^2 + n_{1p}^2}{(n_{1p} + n_{2p})^2} \\
&= \left(\frac{n_{1p}}{n_p}\right)^2 + \left(\frac{n_{2p}}{n_p}\right)^2.
\end{aligned}
$$

Let's show that $\frac{1}{2} \leq m_p \leq 1$ for all $p$. Given that $n_{1p}^2 + n_{2p}^2 \leq (n_{1p} + n_{2p})^2$, then $m_p \leq 1$. Now, given that $(n_{1p} - n_{2p})^2 = n_{1p}^2 - 2n_{1p}n_{2p} + n_{2p}^2 \geq 0$, then $2n_{1p}^2 + 2n_{2p}^2 \geq (n_{1p} + n_{2p})^2$, so that $m_p \geq \frac{1}{2}$.

Then, the ratio of $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$ to $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ is in $[1/2, 1]$ since $m_p \in [1/2, 1]$ for all $p$, and $\sum_p \zeta_p = 1$. **QED.**

Lemma E.1 shows that $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ is a weighted average across pairs of the sum of squared randomization-unit shares within pairs. Accordingly, this ratio is bounded between a half and one. Figure 1 plots this ratio when $n_{1p}/n_{2p}$ is constant across pairs. $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ is very close to 1/2 when $n_{1p}/n_{2p}$ is included between 0.5 and 2, meaning that the first randomization unit has between half and twice as many observations as the second one. For instance, if in every pair, one randomization unit has twice as many observations as the other, then the ratio of the two variances is equal to 5/9. Moreover, the fact that $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ is a weighted average across pairs implies that even if there is a pair with randomization units that are highly unbalanced, this ratio will still be close to 1/2 if other pairs are balanced. Overall, Lemma E.1 shows that Lemma 4.2 still approximately holds when randomization units in each pair have different numbers of observations, unless they have an extremely unbalanced number of observations.

Figure 1: Ratio of Randomization-Unit-Clustered and Pair-Clustered Variance Estimator in Regressions with Paired Fixed Effects