At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?

Clément de Chaisemartin, Jaime Ramirez-Cuellar

At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?*

Clément de Chaisemartin†

Jaime Ramirez-Cuellar‡

September 15, 2022

Abstract

In clustered and paired experiments, to estimate treatment effects, researchers often regress their outcome on the treatment and pair fixed effects, clustering standard errors at the unit-of-randomization level. We show that even if the treatment has no effect, a 5%-level t-test based on this regression will wrongly conclude that the treatment has an effect up to 16.5% of the time, an error rate much larger than the researcher's 5% target. To achieve their targeted error rate, researchers should instead cluster standard errors at the pair level. Using simulations, we show that similar results apply to clustered experiments with small strata.

*We are very grateful to Antoine Deeb, Jake Kohlhepp, David McKenzie, 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.

†Sciences Po.

‡Microsoft Research Redmond.

1 Introduction

In this paper, we show that a statistical test commonly used by researchers analyzing a certain type of randomized controlled trials (RCTs) has a much larger error rate than previously thought. We then suggest a simple fix.

The type of RCTs this paper applies to are clustered RCTs with large clusters, meaning that there are more than 10 observations per randomization unit, and where the treatment is assigned within pairs of units, or within small strata of less than 10 units. For instance, our paper would apply to an RCT where the researcher pairs some villages, randomly assigns one village per pair to the treatment, and estimates a regression at the villager rather than at the village level, with more than 10 villagers per village. Our paper would also apply to an RCT where the researcher groups villages into small strata of six villages and randomly assigns two villages per pair to the treatment. While such paired or small-strata RCTs with large clusters do not account for the majority of RCTs conducted in economics, it seems that they are still fairly common. We surveyed the universe of published editions of the *American Economic Journal: Applied Economics* (*AEJ Applied*) from 2014 to 2018, and found that this type of RCT represents 20% of the RCTs published by the journal during that period.

In paired or small-strata RCTs with large clusters, researchers usually estimate the treatment effect by regressing their outcome on the treatment and pair or stratum fixed effects, "clustering" their standard errors at the unit-of-randomization level, namely, at the village level in our example.[1] Then, they typically use the 5%-level t-test based on this regression to assess if the treatment has an effect. As any statistical test, this t-test may lead them to commit a type 1 error: even if the treatment does not have an effect, this t-test may lead them to wrongly conclude that the treatment has an effect. When using a 5%-level test, researchers hope that the probability that this would happen, the so-called error rate of the test, is less than 5%. We show that the error rate of this t-test may in fact be much larger than the researcher's 5% target.

We start by considering paired RCTs with large clusters. There, we show that even if the treatment has no effect, this 5%-level t-test will wrongly conclude that the treatment has an effect up to 16.5% of the time, an error rate more than three times larger than the targeted one. We then show that to achieve the desired 5% error rate, researchers should instead cluster their standard

[1] Throughout this paper, clustered standard errors refer to the estimators proposed by Liang and Zeger (1986), which are routinely implemented in standard statistical packages. Hereafter, we write "clustered RCT" to refer to the RCT design, and "clustered standard errors" to refer to the standard errors researchers use.

errors at the pair level. Finally, we revisit 371 regressions from the paired RCTs in our survey, and find that clustering at the pair rather than at the randomization-unit level diminishes the number of effects that are significant at the 5% level by one third.

The intuition underlying our results is rather simple. In regression analysis, clustered standard errors are reliable when the regression's dependent and independent variables are uncorrelated across clusters (Cameron and Miller, 2015). Therefore, in a paired RCT, clustering at the unit level relies on the assumption that the regression's independent variable, the treatment, is uncorrelated across all randomization units. However, the treatments of the two units in the same pair are perfectly negatively correlated: if unit A is treated, then unit B must be untreated, and vice versa. This is the reason why unit-clustered standard errors are unreliable, and the error rate of the t-test based on unit-clustered standard errors differs from that targeted by the researcher. On the other hand, clustering at the pair level only relies on the assumption that units' treatments are uncorrelated across pairs, which is true by design in paired experiments. This is the reason why pair-clustered standard errors are reliable, and the error rate of the t-test based on pair-clustered standard errors is equal to the error rate targeted by the researcher.

Our recommendations apply only to paired and clustered RCTs with large clusters. If the RCT is nonclustered, the 5%-level t-test based on unit-clustered standard errors has a 5% error rate, as targeted, after the degrees-of-freedom (DOF) adjustment automatically implemented in most statistical software. On the other hand, after this DOF adjustment the error rate of the 5%-level t-test based on pair-clustered standard errors is lower than 5%. Therefore, to achieve their targeted error rate in nonclustered paired RCTs, researchers should either use DOF-adjusted unit-clustered standard errors, or non-DOF-adjusted pair-clustered standard errors. In clustered RCTs, the same applies to regressions estimated at the unit-of-randomization level rather than at the observation level. Finally, in clustered RCTs with strictly fewer than 10 observations per randomization unit, our recommendation is to use pair-clustered standard errors, without the DOF adjustment. Figure 1 below summarizes our recommendations for applied researchers, depending on whether their RCT is clustered and on the number of observations per randomization unit.

Then, we turn to small-strata RCTs with large clusters. Using simulations, we show that our results for paired designs extend to this case. Intuitively, there as well the treatments of units in the same stratum are negatively correlated, so this correlation should be accounted for. Therefore, our simulations show that in those designs too, the error rate of the 5%-level t-test based on unit-clustered standard errors is larger than 5%. The difference between the t-test's actual and targeted

error rates diminishes when the number of units per strata increases. Again, this makes intuitive sense: the larger the strata, the lower the correlation of the treatments of units belonging to the same stratum. For instance, with five units per strata, the error rate of the 5% level t-test based on unit-clustered standard errors is equal to 7.9%. With 10 units per strata, this error rate is equal to 6.2%. With more than 10 units per strata, this error rate is lower than 6%, and becomes very close to the targeted 5% error rate. This is why—though we acknowledge it is somewhat arbitrary—we use a threshold of 10 units per strata to define an RCT with "small" strata. Our simulations also show that the error rate of the 5%-level t-test based on strata-clustered standard errors is very close to 5%. Accordingly, in small-strata RCTs with large clusters, we recommend that researchers cluster their standard errors at the strata level. If the RCT has too few strata to cluster at that level, researchers could use randomization inference, or the standard-error estimator in Section 9.5.1 of Imbens and Rubin (2015), provided each stratum has at least two treated and two control units.

Related literature

Our paper is related to several papers that predate ours. Imai et al. (2009) show that, when units all have the same number of observations, pair-clustered standard errors are reliable in clustered and paired RCTs.[2] With respect to their paper, we show that this result still holds when units have varying numbers of observations, thus justifying pair-level clustering under more realistic assumptions. Moreover, while they focus on finite-sample results, we present large-sample results for t-tests based on pair-clustered standard errors.

Bruhn and McKenzie (2009) use simulations to study, in nonclustered paired RCTs, the error rate of a t-test without clustering, which is equivalent to unit-clustering when the RCT is nonclustered. They show that this error rate is equal to the targeted error rate. This result may appear to conflict with ours, but this apparent discrepancy comes from the DOF adjustment embedded in most statistical software. In nonclustered RCTs, the regression with pair fixed effects has one fixed effect for each pair of observations. Accordingly, it has approximately half as many regressors as observations, so the DOF adjustment amounts to multiplying nonclustered standard errors by approximately $\sqrt{2}$, which, we show, makes them almost equivalent to the non-DOF-adjusted pair-clustered standard errors. This is why the error rate of DOF-adjusted nonclustered t-tests is equal to the targeted error rate in nonclustered paired RCTs.

[2]Imai (2008) show similar results in nonclustered and paired RCTs.

Abadie et al. (2017) examine the appropriate level of clustering in regression analysis. They define a cluster as a group of units whose treatments are positively correlated. Their results apply to the case where the assignment is fully clustered (all units in the same cluster have the same treatment), and to the case where the assignment is probabilistically clustered (units in the same cluster have positively correlated treatments). Their Corollary 1 states that standard errors need to account for clustering whenever units' treatments are positively clustered within clusters. Our results are consistent with theirs. The main difference is that we consider a case in which units' treatments are negatively correlated within clusters (the pairs in our paper), which is not something they consider. However, our papers share a common theme: that standard errors need to account for clustering when treatments are correlated within some groups of units.

Athey and Imbens (2017) and Bai et al. (2021) study the error rate of t-tests based on unit-clustered standard errors in paired experiments, when pair fixed effects are not included in the regression. They both show that without pair fixed effects, the test's error rate is lower than the targeted error rate. We instead show that when pair fixed effects are included in the regression, the error rate of this test becomes larger than the targeted error rate. In our survey of paired experiments, we find that including pair fixed effects in the regression is a much more common practice than not including those fixed effects.

The rest of this paper is organized as follows. Section 2 presents our survey of paired and small-strata RCTs in economics. Section 3 introduces our main theoretical results. Section 4 presents our simulation study. Section 5 presents our empirical application. Section 6 briefly discusses various extensions of our baseline results, which are fully developed in our Web Appendix. Section 7 concludes. Throughout the paper, a unit refers to a randomization unit (e.g., a village), while an observation refers to the level at which the regression is estimated (e.g., a villager).

2 Survey of Paired and Small-Strata Experiments in Economics

We searched the 2014–2018 issues of the *AEJ Applied* for clustered and paired RCTs, and for clustered and stratified RCTs with 10 or fewer units per strata. 50 field-RCT papers were published over that period. Three RCTs were clustered and relied on a paired randomization for all of their analysis. One RCT was clustered and relied on a paired randomization for part of its analysis. Seven RCTs were clustered and used a stratified design with, on average, 10 or fewer units per strata. 10 of those 11 RCTs have, on average, more than 10 observations per randomization unit,

while one stratified RCT has 5.8 observations per randomization unit. Overall, 11 (22%) of the 50 field RCTs published by the *AEJ Applied* over that period are clustered paired or small-strata RCTs, and 10 (20%) also have large clusters.

To increase our sample of paired RCTs, we also searched the AEA's registry website (`https://www.socialscienceregistry.org`). We looked at all completed projects, whose randomization method includes the word "pair" and that either have a working or a published paper. We conducted that search on January 9th 2019, and found four more clustered and paired RCTs. All of them have, on average, more than 10 observations per randomization unit. Combining our two searches, we found 15 clustered paired or small-strata RCTs. The list is in Table 5 in the Web Appendix.

We now give descriptive statistics on our sample of RCTs. Across the eight paired RCTs, the median number of pairs is 27, the median number of observations per unit is 112, and units have on average more than 10 observations in all RCTs. To estimate the treatment effect, six articles include pair fixed effects in all their regressions, one article includes pair fixed effects in some but not all their regressions, and one article does not include pair fixed effects in any regression. All articles cluster standard errors at the unit level.

Across the seven small-strata RCTs, the median number of units per strata is 7, the median number of strata is 48, the median number of observations per unit is 26, and units have more than 10 observations on average in all but one RCT. To estimate the treatment effect, six articles include stratum fixed effects in all their regressions, and one article does not include stratum fixed effects in any regression. All articles cluster standard errors at the unit level.

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

3 Theoretical Results

3.1 Setup

We consider a population of $2P$ units. Unlike Abadie and Imbens (2008) and Bai et al. (2021), we do not assume that the units are an independent and identically distributed (i.i.d.) sample drawn from a superpopulation. Instead, that population is fixed, and its characteristics are not random. Our survey suggests that this modeling framework—similar to that in Neyman (1923) and Abadie et al. (2020)—is applicable to the majority of paired- and small-strata RCTs in our survey. Units

are drawn from a larger population in only one of those fifteen RCTs.[3] In all the other RCTs, the sample is a convenience sample, consisting of volunteers to receive the treatment, or of units located in areas where conducting the research was easier. When the units are an i.i.d. sample drawn from a super-population, our results still hold, conditional on the sample.

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 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. When $n_{gp} > 1$ for at least some units, the RCT is clustered; when $n_{gp} = 1$ for all units, the RCT is nonclustered.

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. $(W_{1p}, W_{2p})_{p=1}^{P}$ is jointly independent across p.

Point 1 requires that in each pair, one of the two units is treated. Point 2 requires that the two units have the same probability of being treated. Point 3 requires that the treatments be independent across pairs. Assumption 1 is typically satisfied by design in paired experiments.

Let $y_{igp}(1)$ and $y_{igp}(0)$ represent the potential outcomes of observation i in unit g and pair p with and without the treatment, respectively. We follow the randomization-inference literature (see Abadie et al., 2020) and assume that potential outcomes are fixed.[4] The observed outcome is $Y_{igp} = y_{igp}(1)W_{gp} + y_{igp}(0)(1 - W_{gp})$. Our target parameter is the average treatment effect (ATE)

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

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

[4] In a previous version of this paper, we allowed potential outcomes to be stochastic. Having stochastic potential outcomes does not change our main results; see de Chaisemartin and Ramirez-Cuellar (2020).

We consider two estimators of τ. The first estimator $\widehat{\tau}$ is the OLS estimator from the regression of the observed outcome Y_{igp} on a constant and W_{gp}:

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

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 $(\delta_{ig1}, \ldots, \delta_{igP})$:

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

3.2 Properties of Unit- and Pair-Clustered Variance Estimators

We study the variance estimators of $\widehat{\tau}$ and $\widehat{\tau}_{fe}$, when the regression is clustered at either the pair level or the unit level. The clustered-variance estimators we study are those proposed in Liang and Zeger (1986). Lemma C.1 in Web Appendix C gives simple expressions of $\widehat{\mathbb{V}}_{pair}(\widehat{\tau})$ and $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$, the pair-clustered variance estimators (PCVEs) of $\widehat{\tau}$ and $\widehat{\tau}_{fe}$, and of $\widehat{\mathbb{V}}_{unit}(\widehat{\tau})$ and $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$, the unit-clustered variance estimators (UCVEs) of $\widehat{\tau}$ and $\widehat{\tau}_{fe}$.

We now present our main results, that are derived under the following assumption.

Assumption 2. There is a strictly positive integer N such that for all p, $n_{1p} = n_{2p} = N$.

Assumption 2 requires that all units have the same number of observations. Let

$$\widehat{\tau}_p = \sum_g \left[W_{gp} \frac{1}{n_{gp}} \sum_i Y_{igp} - (1 - W_{gp}) \frac{1}{n_{gp}} \sum_i Y_{igp} \right]$$

denote the difference between the average outcome of treated and untreated observations in pair p. Under Assumption 2, one can show that

$$\widehat{\tau} = \widehat{\tau}_{fe} = \sum_{p=1}^{P} \frac{\widehat{\tau}_p}{P},$$

that both estimators are unbiased for the ATE, and that

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

Let $\tau_p \equiv \frac{1}{n_p} \sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} [y_{igp}(1) - y_{igp}(0)]$ be the ATE in pair p. For all $d \in \{0, 1\}$, let $\overline{y}_{gp}(d) \equiv$

$\frac{1}{n_{gp}} \sum_i y_{igp}(d)$, $\overline{y}_p(d) \equiv \frac{1}{2} \sum_g \overline{y}_{gp}(d)$, and $\overline{y}(d) \equiv \sum_p \overline{y}_p(d)/P$, respectively, denote the average outcome with treatment d in pair p's unit g, in pair p, and in the entire population.

Lemma 3.1.

1. *If Assumptions 1 and 2 hold, then* $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = \widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$*, and*

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

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

3. *If Assumptions 1 and 2 hold, then*

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

Proof. See Web Appendix A. □

Point 1 of Lemma 3.1 shows that the PCVEs without and with pair fixed effects are equal, and that after a DOF correction, their expectation is at least as large as the variance of $\widehat{\tau}$. If the treatment effect is heterogeneous across pairs, $\frac{1}{P(P-1)}\sum_{p=1}^{P}(\tau_p - \tau)^2 > 0$ so the inequality is strict: the PCVEs are upward-biased estimators for the variance of $\widehat{\tau}$. If the treatment effect does not vary across pairs, the inequality becomes an equality: the PCVEs are unbiased for the variance of $\widehat{\tau}$.[5] Building upon Point 1 of Lemma 3.1, in the Web Appendix we show that when the number of pairs grows, $(\widehat{\tau} - \tau)/\widehat{\mathbb{V}}_{pair}(\widehat{\tau})$ and $(\widehat{\tau}_{fe} - \tau)/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$, the t-statistics of the difference-in-means and fixed-effects estimators using the PCVEs, both converge 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 (see Point 1 of Theorem B.1). Comparing those t-statistics to critical values of a standard normal leads to a test with an error rate at most equal to the researcher's target. For instance, if the average treatment effect τ is equal to zero, by comparing $\left|\widehat{\tau}/\sqrt{\widehat{V}_{pair}(\widehat{\tau})}\right|$ to 1.96, one would wrongly conclude that $\tau \neq 0$ at most 5% of the time, as desired.

[5]The displayed equation in Point 1 is almost identical to Proposition 1 in Imai et al. (2009), up to a DOF adjustment. We restate that result from their paper for completeness.

On the other hand, Point 2 of Lemma 3.1 shows that the UCVE with pair fixed effects is equal to a half of the PCVEs. Combined with Point 1 of Lemma 3.1, this implies that the UCVE with pair fixed effects may severely underestimate the variance of $\widehat{\tau}$: if the treatment effect is constant across pairs, its expectation is equal to half of the variance of $\widehat{\tau}$. Building upon Point 2 of Lemma 3.1, in the Web Appendix we show that when the number of pairs grows, $(\widehat{\tau}_{fe} - \tau)/\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$, 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 (see Point 2 of Theorem B.1). Therefore, comparing that t-statistic to critical values of a standard normal may yield a test with a substantially larger error rate than the researcher's target. For instance, if the average treatment effect τ is equal to zero and the treatment effect is homogenous across pairs, by comparing $\left|\widehat{\tau}_{fe}/\sqrt{\widehat{V}_{unit}(\widehat{\tau}_{fe})}\right|$ to 1.96, one would wrongly conclude that $\tau \neq 0$ 16.5% of the time, an error rate more than three times larger than the researcher's target.

With heterogeneous treatment effects across pairs, the error rates of the t-tests using the PCVEs may be lower than the researcher's target, while the error rate of the t-test using the UCVE with pair fixed effects may be equal to that target. However, in practice we do not know if the treatment effect is constant or heterogeneous, and it is common to require that a test have an error rate no larger than some target uniformly across all possible data-generating processes. The t-tests using the PCVEs satisfy that property, unlike the t-test using the UCVE with pair fixed effects.

Finally, Point 3 of Lemma 3.1 shows that without pair fixed effects, the expectation of the difference between the UCVE and PCVE 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 the difference between those two terms 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. Consequently, the expectation of the difference between the UCVE and the PCVE should often be positive. Then, it follows from Point 1 of Lemma 3.1 that the UCVE without pair fixed effects is a more upward-biased estimator of the variance of $\widehat{\tau}$ than the PCVEs, and that it remains upward-biased even if the treatment effect is constant across pairs. Finally, building upon Point 3 of Lemma 3.1, in the Web Appendix we show that the error rate of the t-statistic of the difference-in-means estimator using the UCVE is lower and further away from the researcher's target than the error rate of the t-test making use of the PCVEs (see Point 3 of Theorem B.1).

Intuitively, the UCVEs are biased because clustering at the unit level does not account for the perfect negative correlation of the treatments of the two units in the same pair. Cluster-robust standard errors rely on the assumption that observations' outcomes and treatments are uncorrelated across clusters (see Cameron and Miller, 2015). This assumption is violated when one clusters at the unit level, but it holds when one clusters at the pair level.

The direction of the bias of the UCVE depends on whether pair fixed effects are included in the regression. When pair fixed effects are not included in the regression, the UCVE will in general overestimate the variance of $\widehat{\tau}$. This result may be relatively intuitive. With positive correlations between observations, as is often the case with time-series data, the variance of an estimator is usually larger than what it would be without those correlations. Then, one would expect that negative correlations would reduce an estimator's variance. This is indeed what we find in Point 3 of Lemma 3.1: the UCVE, which estimates $\widehat{\tau}$'s variance as if the treatments of two units in the same pair were not negatively correlated, is larger than needed.

On the other hand, when pair fixed effects are included in the regression, the UCVE may underestimate the variance of $\widehat{\tau}$. This result is less intuitive. It comes from the fact that with pair fixed effects in the regression, the sample residuals u_{igp} are by construction uncorrelated with the pair fixed effects, which implies that for every p, the sum of the residuals in pair p is zero:

$$\sum_{i,g} u_{igp} = 0.$$

Splitting the summation between $g = 1$ and $g = 2$, using the fact that under Assumption 2 units 1 and 2 have the same number of observations, and letting $\overline{u}_{g,p}$ denote the average residuals of observations in unit g of pair p, the previous display implies that $\overline{u}_{1,p} = -\overline{u}_{2,p}$, which in turn implies that $(\overline{u}_{1,p})^2 = (\overline{u}_{2,p})^2$: by construction, the squares of the average residuals are equal in the treated and control units of each pair. Now, one can show that with pair fixed effects, the UCVE is proportional to

$$\frac{1}{(2P)^2} \sum_{p=1}^{P} \sum_{g=1}^{2} (\overline{u}_{g,p})^2 ,$$

the sum, across all units, of their average squared residuals, divided by the number of units squared. Accordingly, $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$ treats $(\overline{u}_{1,p})^2$ and $(\overline{u}_{2,p})^2$ as if they were independent to estimate the variance

of $\widehat{\tau}_{fe}$, while they are equal to each other. Instead, the PCVE is proportional to

$$\frac{1}{P^2}\sum_{p=1}^{P}\left(\overline{u}_{1,p}\right)^2.$$

$\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ uses only one squared-residual per pair to estimate the variance of $\widehat{\tau}_{fe}$.

As Section 5 below shows, our recommendation of using the PCVE rather than the UCVE in clustered-paired RCTs and regressions with pair fixed effects leads to a significant reduction in the number of effects that are significant at the 5% level in the published papers we revisit. One may then wonder whether our results contradict those in Bai (2019), who shows that pairing is the optimal RCT design to maximize statistical precision.[6] The short answer is that our findings do not contradict his important result. Bai (2019) shows that the RCT design that minimizes $\mathbb{V}(\widehat{\tau})$, and therefore the mean-squared error of $\widehat{\tau}$, is a specific paired design. We do not derive any new result on $\mathbb{V}(\widehat{\tau})$, so our results have no bearing on his. Instead, our main result is to show that $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$, a commonly used variance estimator in paired experiments, can be severely downward-biased. Instead, we recommend using another estimator, $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$, that is not downward biased, and leads to a t-test with an error rate no larger than the researcher's target when the treatment does not have an effect. Using $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ instead of $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$, researchers will conclude less often that the treatments they consider have an effect, but comparing the power of those two t-tests is not a fair comparison: the former test has an error rate no larger than the researcher's target when the treatment does not have an effect, unlike the latter one. Overall, while paired RCTs may not be as powerful as the use of a spuriously-low variance estimator had led researchers to believe, they remain a very powerful RCT design, the one that leads to the lowest mean-squared error of $\widehat{\tau}$.

There is only one case where our results could imply that other designs might be preferable to paired RCTs, though further research is needed to validate or invalidate this conjecture. In our simulations, we find that with fewer than 20 pairs, t-tests based on the PCVE become less reliable: with fewer than 40 units, using the PCVE in paired RCTs may lead to invalid inference. Thus, it may be preferable to run a more coarsely stratified RCT with at least four units per strata, and use, for example, the variance estimator proposed in Section 6.1 of Athey and Imbens (2017) for stratified RCTs. However, to our knowledge, the validity of this alternative inference procedure has not been assessed yet with a small number of units in the RCT. Note that in the (admittedly small)

[6] Bai (2019) studies this question in nonclustered RCTs. The optimal design in clustered RCTs has not been derived yet, though we conjecture that the result in Bai (2019) carries through to clustered RCTs where units all have the same number of observations.

sample of eight paired RCTs in our survey, one has five pairs and another one has 14 pairs. All the other RCTs are close to the 20-pair "threshold" (one has 19 pairs), or above it. Accordingly, while paired RCTs with far fewer than 20 pairs are not a rarity, they do not seem to be common either.

3.3 Accounting for Degrees-of-Freedom Adjustments

The clustered-variance estimators we study are those proposed in Liang and Zeger (1986). Typically, statistical software report DOF-adjusted versions of those estimators. For instance, in Stata the default adjustment is to multiply the Liang and Zeger estimator by $[(n-1)/(n-k)] \times [G/(G-1)]$, where n is the sample size, k the number of regressors, and G the number of clusters (see StataCorp, 2017). This DOF adjustment is implemented when one uses the regress or areg command, not when one uses the xtregress command (see Cameron and Miller, 2015).[7] In R, if the researcher uses the sandwich package, the default DOF adjustment when declaring a cluster variable is the same as in Stata, namely, $[(n-1)/(n-k)] \times [G/(G-1)]$. $G/(G-1)$ is close to 1, so the important term in the DOF adjustment is $(n-1)/(n-k)$.

In regressions without pair fixed effects, there are only two regressors (the constant and the treatment), so $(n-1)/(n-k) = (n-1)/(n-2)$. This quantity is close to 1, so the DOF adjustment leaves the UCVE and the PCVE almost unchanged. Accordingly, in regressions without pair fixed effects, the guidance we derived in the previous section also applies to the DOF-adjusted UCVE and PCVE: the former estimator should not be used, while the latter estimator can be used.

On the other hand, in regressions with pair fixed effects, the DOF adjustment may affect the UCVE and the PCVE more substantially. When the paired RCT is not clustered, the regression has $2P$ observations and $P+1$ regressors, so $(n-1)/(n-K) = (2P-1)/(P-1) \approx 2$: the DOF-adjusted UCVE is twice as large as the non-DOF-adjusted UCVE. This fact and Point 2 of Lemma 3 imply that in nonclustered RCTs, the DOF-adjusted UCVE with pair fixed effects is almost equal to the non-DOF-adjusted PCVE with pair fixed effects and has the same desirable properties. On the other hand, the DOF-adjusted PCVE with pair fixed effects is now about twice as large as the non-DOF-adjusted PCVE with pair fixed effects, so this estimator is upward-biased even under constant treatment effect. Overall, in nonclustered paired RCTs and regressions with pair fixed effects, the guidance we derived in the previous section no longer applies to the DOF-adjusted UCVE and PCVE: the former estimator can be used, while the latter estimator should not be used.

[7] Three of the four papers we revisit in Section 5 use the regress or areg command; one uses the xtreg command.

When the paired RCT is clustered and the regression has pair fixed effects, the regression has $2P\overline{n}_u$ observations and $P+1$ regressors, where $\overline{n}_u$ denotes the average number of observations across all units. Accordingly, $(n-1)/(n-K) = (2P\overline{n}_u - 1)/(2P\overline{n}_u - (P+1)) \approx 2\overline{n}_u/(2\overline{n}_u - 1)$. This quantity is decreasing in $\overline{n}_u$: the larger the average number of observations across units, the smaller the DOF adjustment. Simulations shown in Panel D of Table 1 show that with $\overline{n}_u = 5$, the error rate of a t-test based on the DOF-adjusted UCVE is still considerably larger than the researcher's target: unlike what happens in nonclustered experiments, the DOF-adjustment is not sufficient to ensure the error rate of this t-test is equal to the researcher's target. The same panel also shows that with $\overline{n}_u = 5$, the error rate of a t-test based on the DOF-adjusted PCVE is slightly below the researcher's target, even under constant treatment effects. When $\overline{n}_u = 10$, simulations shown in Panel C of Table 1 show that the error rate of a t-test based on the DOF-adjusted PCVE is now very close to the researcher's target. Overall, in clustered paired RCTs with more than 10 observations per unit, the guidance we derived in the previous section also applies to the DOF-adjusted UCVE and PCVE: the former estimator should not be used, while the latter estimator can be used. In clustered paired RCTs with strictly fewer than 10 observations per unit, we recommend using the PCVE without the DOF adjustment, which is in line with a recommendation in Cameron and Miller (2015) in a different context. In Stata, the xtregress command computes this estimator.

3.4 Should Pair Fixed Effects Be Included in the Regression?

Though this paper is primarily concerned with the estimation of the variance of treatment effect estimators, our recommendations crucially depend on whether pair fixed effects are included in the regression. In this section, we discuss the pros and cons of including such pair fixed effects. (Our paper does not bring any new result to this longstanding discussion; we rely on earlier results, sometimes specializing them to the case of paired RCTs.)

In nonclustered experiments, or in clustered experiments where in each pair the two randomization units have the same number of observations ($n_{1p} = n_{2p}$ for all p), if no randomization unit attrits from the sample, adding pair fixed effects to the regression leaves the treatment coefficient unchanged: $\widehat{\tau} = \widehat{\tau}_{fe}$. Because $\widehat{\tau}$ and $\widehat{\tau}_{fe}$ are equal, their variances are also equal: adding pair fixed effects to the regression does not lead to any precision gain in nonclustered paired experiments or in clustered experiments where in each pair the two randomization units have the same number of observations. When there is attrition, $\widehat{\tau}$ and $\widehat{\tau}_{fe}$ will differ: $\widehat{\tau}_{fe}$ will only leverage observations from pairs where both randomization units are observed, while $\widehat{\tau}$ will also leverage observations from

pairs where only one of the randomization units is observed. King et al. (2007) argue in favor of dropping pairs with one attriting unit, while Bai (2019) shows they can be kept. At any rate, even if one would prefer to drop those pairs, one can simply do so before running the regression, rather than running the regression with pair fixed effects in the full sample. Overall, in nonclustered experiments, or in clustered experiments where in each pair the two randomization units have the same number of observations, there is no strong argument for or against adding pair fixed effects to the regression.

In clustered experiments where there are pairs where the two randomization units have different numbers of observations ($n_{1p} \neq n_{2p}$ for some p), adding pair fixed effects to the regression may change the treatment coefficient: $\widehat{\tau} \neq \widehat{\tau}_{fe}$. In such cases, one can show that $\widehat{\tau}$, the standard difference in means estimator, converges toward our target parameter τ, the average treatment effect, when the number of pairs goes to infinity. $\widehat{\tau}_{fe}$ on the other hand does not converge toward τ: one can show that it converges toward a parameter that has sometimes been called a variance weighted average (see Angrist and Pischke, 2008) of the average treatment effect in each pair.[8] This parameter may differ from τ if the treatment effect varies across pairs. Accordingly, unlike $\widehat{\tau}$, $\widehat{\tau}_{fe}$ may be biased for τ, even asymptotically. On the other hand, the variance of $\widehat{\tau}_{fe}$ is often lower than that of $\widehat{\tau}$ (see Imai et al., 2009). Overall, if one is primarily interested in consistently estimating τ, pair fixed effects should not be included in the regression. If one wants to use the most precise estimator, pair fixed effects should be included in the regression.

Figure 1 summarizes our recommendations for practitioners, regarding whether pair fixed effects should be included in the regression, and regarding which variance estimator one should use.

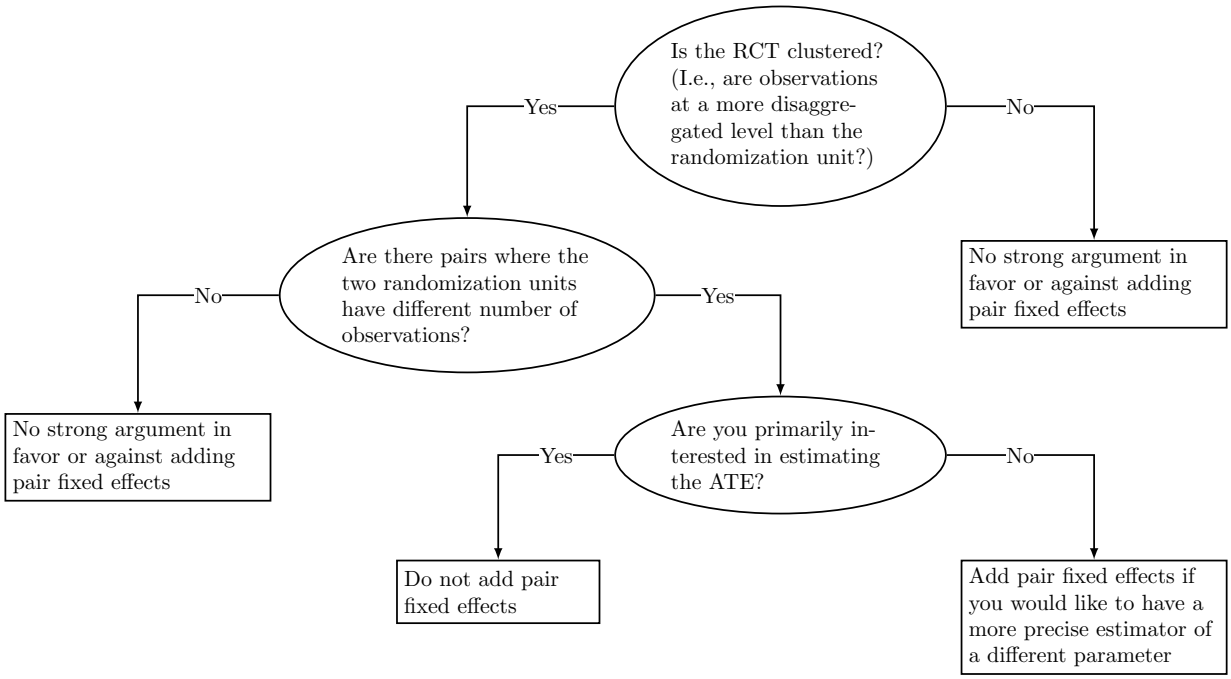
4 Simulations Using Real Data

We perform Monte-Carlo simulations using a real data set. We use the data from the microfinance RCT in Crépon et al. (2015b). 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. The number of observations varies substantially across units: the average number of villagers per village is 34.1, with a standard deviation of 9.2, a minimum of 13 and a maximum of 58.

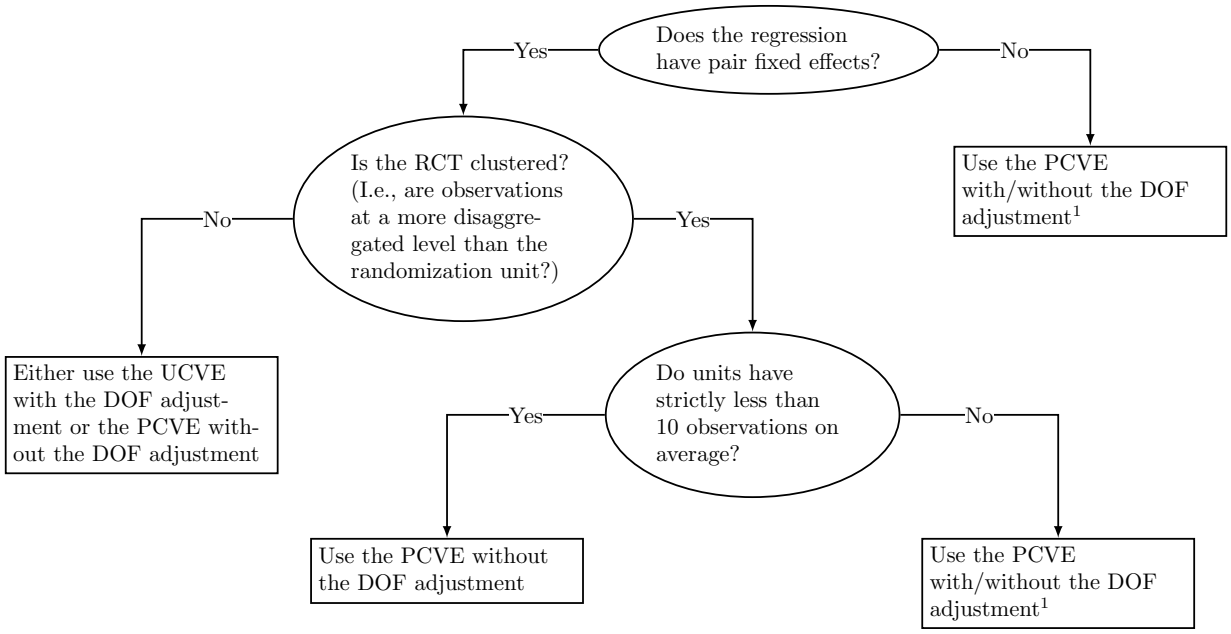
[8] Specifically, let τ_{gp} denote the average treatment effect in unit g of pair p. One can show that $\widehat{\tau}_{fe}$ is consistent for a weighted average, across pairs, of $1/2(\tau_{1p} + \tau_{2p})$, where pairs in which the numbers of observations of the two units are close receive more weight than pairs where the numbers of observations are different.

Figure 1: Recommendations for Practitioners

(a) Decision 1: Should the regression include pair fixed effects? (See Section 3.4 for more details.)



(b) Decision 2: Which variance estimator should one use? (See Section 3.3 for more details.)



Notes: UCVE = unit-clustered variance estimators. PCVE = pair-clustered variance estimators. DOF = degrees of freedom. [1]In these cases, the PCVE with and without the DOF adjustment are very similar.

In the paper, the authors report the effect of the microfinance intervention on 82 outcome variables.[9] For each outcome, we construct potential outcomes assuming no treatment effect, i.e., $y_{igpk}(0) = y_{igpk}(1) = Y_{igpk}$, where Y_{igpk} is the value of outcome k for villager i in village g and pair p. We then simulate 1000 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. We estimate regressions with and without pair fixed effects, clustering at the pair level and at the village level. Thus, we obtain four t-statistics, and four 5% level t-tests. Importantly, those t-tests are based on Stata's regress command, so they make use of DOF-adjusted variance estimators. The estimated error rate of each t-test is the percentage of times, across the 82,000 regressions (82 outcomes $\times$ 1000 simulations), that the t-statistic is greater in absolute value than 1.96, meaning that the test leads the researcher to wrongly conclude that the treatment has an effect. Because the data is generated with a constant treatment effect of zero, these error rates should be equal to 5% if the tests are valid.

Column (1) of Panel A of Table 1 shows the results using the authors' actual data set, with 81 pairs and villages' actual number of villagers. The error rates of the t-tests using pair-clustered variance estimators (PCVEs) are close to 5%, irrespective of whether pair fixed effects are included in the regression. On the other hand, when the unit-clustered variance estimator (UCVE) is used with pair fixed effects, the error rate of the t-test is equal to 17.4%, very close to the 16.5% error rate predicted by Point 2 of Theorem B.1. Finally, the error rate of the t-test with the UCVE and no pair fixed effects is equal to 1.4%, well below 5%. Columns (2), (3), and (4) show that we obtain similar results if we use a random sample of 40, 30, and 20 pairs. With fewer than 20 pairs, the PCVE becomes downward-biased. One may then have to use randomization inference tests.

Panel B (resp. C) of Table 1 shows the error rates of the four t-tests, in a data set where villages all have 20 (resp. 10) villagers. In each village, the villagers are a random sample from the village's population, that does not vary across simulations.[10] Results are similar to Panel A.

Panel D shows the error rates of the four t-tests, in a data set where villages all have 5 villagers. Again, the error rate of the t-test with the PCVE and no pair fixed effects is close to 5%. On the other hand, the error rate of the t-test with the PCVE and pair fixed effects is now below 5%. As discussed in Section 3.3, this is due to the fact that the DOF-adjustment is not negligible anymore

[9] Across the 82 outcomes, the median intracluster correlation coefficient is 0.063 at the village level, and 0.054 at the pair level.

[10] Some villages have fewer than 20 villagers. For a village with, for example, 13 villagers, we draw 7 villagers from the village's population and add them to the original villagers.

with 5 villagers per village. The error rate of the t-test with the UCVE and pair fixed effects is still much higher than 5%, but less so than in Panel A. Finally, the error rate of the t-test with the UCVE and no pair fixed effects is still well below 5%, though less so than in Panel A. These simulations justify the guidance above: in clustered RCTs with strictly fewer than 10 observations per unit, one should either use the PCVE without pair fixed effects, with or without the DOF adjustment, or the PCVE with pair fixed effects without the DOF adjustment.

Finally, Panel E of Table 1 shows the error rates of the four t-tests, in a data set where a quarter of the villages have five villagers, a quarter have 10 villagers, a quarter have 20 villagers, and a quarter have their actual number of villagers. In Columns (1) and (2), results are fairly similar to those in Panel A. In Columns (3) and (4), the error rates of the t-tests using the PCVEs are larger than 5% (though much less so than the t-test using the UCVE with pair fixed effects). This is related to the results in Carter et al. (2017), who find that when clusters have very heterogeneous sizes, one needs a larger number of clusters to ensure that asymptotic distributions yield accurate approximations of the finite-sample distribution of cluster-robust t-statistics. Note that this phenomenon is absent in Panel A, while village sizes are already fairly heterogeneous in those simulations. In applications where units have very heterogeneous numbers of observations, researchers may need to perform their own simulations to assess whether t-tests using the PCVEs can be used.

5 Application

In this section, we revisit the paired RCTs in our survey. The data used in four of those papers is publicly available (Beuermann et al., 2015b; Bruhn et al., 2016b; Crépon et al., 2015b; Glewwe et al., 2016b). Those four papers used a clustered RCT, and all have more than 10 observations per randomization unit (across the four papers, the lowest average number of observations per randomization unit is 21.5). The authors estimated the effect of the treatment in 294 regressions, clustering at the unit level. In Panel A of Table 2, we re-estimate those regressions, clustering at the pair level, and including the same controls as the authors. In the 240 regressions with pair fixed effects, the average of the unit-clustered variance estimator (UCVE) divided by the pair-clustered variance estimator (PCVE) is equal to 0.548. The UCVE divided by the PCVE is not always exactly equal to 1/2, because Assumption 2 is not always satisfied, but these fractions all are quite close to 1/2, as predicted by Lemma G.4. The authors originally found that the treatment has a

Table 1: Error Rates of T-tests, in Simulations Based on Crépon et al. (2015a)

Clustering level	Pair Fixed Effects	5% level t-test error rate			
		With 81 pairs	With 40 pairs	With 30 pairs	With 20 pairs
		(1)	(2)	(3)	(4)
Panel A: Actual village sizes					
Pair	Yes	0.0500	0.0526	0.0527	0.0559
Pair	No	0.0518	0.0544	0.0541	0.0572
Unit	Yes	0.1709	0.1795	0.1802	0.1840
Unit	No	0.0132	0.0178	0.0178	0.0192
Panel B: All villages have 20 villagers					
Pair	Yes	0.0465	0.0504	0.0510	0.0537
Pair	No	0.0496	0.0539	0.0541	0.0562
Unit	Yes	0.1642	0.1682	0.1741	0.1737
Unit	No	0.0162	0.0213	0.0190	0.0259
Panel C: All villages have 10 villagers					
Pair	Yes	0.0426	0.0439	0.0464	0.0454
Pair	No	0.0490	0.0499	0.0538	0.0498
Unit	Yes	0.1598	0.1513	0.1622	0.1686
Unit	No	0.0223	0.0235	0.0249	0.0265
Panel D: All villages have 5 villagers					
Pair	Yes	0.0361	0.0353	0.0371	0.0363
Pair	No	0.0480	0.0480	0.0502	0.0477
Unit	Yes	0.1382	0.1372	0.1367	0.1421
Unit	No	0.0240	0.0261	0.0313	0.0284
Panel E: Heterogeneous village sizes					
Pair	Yes	0.0514	0.0573	0.0597	0.0673
Pair	No	0.0554	0.0602	0.0623	0.0692
Unit	Yes	0.1764	0.1753	0.1775	0.1797
Unit	No	0.0184	0.0225	0.0228	0.0271

Notes: Table 1 reports the error rates of four 5% level t-tests in Crépon et al. (2015a). 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 pair fixed effects in the regression, and clustering standard errors at the village or at the pair level. All t-tests are based on Stata's regress command, so they make use of DOF-adjusted variance estimators. The error rate of each test is the percent of times, across the 82,000 regressions (82 outcomes $\times$ 1000 replications), that the test leads the researcher to wrongly conclude that the treatment has an effect. Column (1) (resp. (2), (3), (4)) shows the results using the original sample of 81 pairs (resp. a fixed sample of 40, 30, 20 randomly selected pairs). In Panel A, villages all have their actual number of villagers. In Panel B (resp. C, D), each village has 20 (resp. 10, 5) villagers, that are a fixed random sample from the village's population. In Panel E, 1/4 of villages have 5 villagers, 1/4 have 10 villagers, 1/4 have 20 villagers, and 1/4 have their actual number of villagers.

5%-level significant effect in 110 regressions. Using the PCVE, we find significant effects in just 74 regressions. In the 54 regressions without pair fixed effects, the UCVE is on average 1.18 times larger than the PCVE. The authors originally found 31 significant effects, we find 36 significant effects using the PCVE.

The data used in the remaining four papers is not publicly available. Three of those papers estimated 131 regressions with pair fixed effects, clustering standard errors at the unit level.[11] For those regressions, we multiply the UCVE by the average value of of the PCVE divided by the UCVE found in Panel A of Table 2 to predict the value of the PCVE. Panel B of Table 2 shows that while the authors originally found a 5%-level significant effect in 51 regressions, we find significant effects in just 34 regressions. The fourth paper estimated regressions only without pair fixed effects. Because without fixed effects, the value of the PCVE divided by the UCVE can vary significantly across regressions, we do not try to predict the PCVEs of that paper.

Table 2: Using Unit- or Pair-Level Clustered Variance Estimators in Paired RCTs

	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		51	34	131

Notes: The table shows the effect of using pair-clustered variance estimators (PCVE) rather than unit-level clustered variance estimators (UCVE) in seven of the paired RCTs 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 pair 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.

6 Extensions

In our Web Appendix, we consider various extensions. In Appendix E, we present simulations showing that our results for paired RCTs extend to stratified RCTs with few units per strata.

[11] Across these papers, the lowest number of observations per randomization unit is 99.0.

Assumption 2, which requires that all units have the same number of observations, allows us to derive the stark results in Lemma 3.1. Under Assumption 2, $\widehat{\tau}$ and $\widehat{\tau}_{fe}$ are equal, but the UCVE drastically changes when one adds pair fixed effects to the regression. This is obviously undesirable: the two estimators are equal, their variances are equal, so their variance estimators should not be drastically different. In practice, however, Assumption 2 often fails. In that case, we show in Section G of the Web Appendix that our main conclusions still hold. Without that assumption, the PCVEs remain upward-biased in general and unbiased if the treatment effect is homogeneous across pairs. On the other hand, the UCVE with pair fixed effects may still be downward-biased. Specifically, Point 2 of Lemma 3.1 still holds if the number of observations per unit varies across pairs, as long as the two units in a pair have the same number of observations. If the number of observations per unit varies within pairs, Point 2 of Lemma 3.1 still approximately holds, unless units in the same pair have very heterogeneous numbers of observations. Indeed, Lemma G.4 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 unit has between half and twice as many observations as the second one.

In Appendix D, we study two alternatives to the PCVE. With heterogeneous treatment effects across pairs, the PCVE overestimates the variance of the treatment effect estimator. To increase power, one may want to use an unbiased estimator of that variance. We study two alternatives, the pair-of-pairs estimator proposed by Abadie and Imbens (2008), and a variance estimator proposed by Bai et al. (2021).[12] Both are unbiased, or at least consistent, when units are an i.i.d. sample drawn from a superpopulation. In the set-up we consider, where units are a convenience sample, we show that those two estimators are upward-biased, like the PCVE. They are less upward-biased than the PCVE when the treatment effect is less heterogeneous within than between pairs of pairs, and more upward-biased otherwise. We compute the three estimators in the regressions in our survey, and find that they are on average equivalent, so it does not seem one can expect large power gains from using those two alternative estimators. Moreover, simulations based on the data from Crépon et al. (2015a) show that t-tests using those two estimators have a drawback relative to the t-test using the PCVE. The corresponding t-statistics are approximately normally distributed only if the sample has more than a couple hundred pairs. On the other hand, the t-test based on

[12]Other alternatives have been proposed. For instance, Fogarty (2018) proposes to use covariates that predict the treatment effect heterogeneity across pairs to form a less-upward-biased estimator than the pair-clustered one. We do not consider this estimator, merely because it lends itself less easily to the automatic replication exercise we conduct: in each application, one has to determine the relevant covariates to include, based on context-specific knowledge.

the PCVE is approximately normally distributed with as few as 20 pairs.

7 Conclusion

In paired or small-strata RCTs with large clusters, researchers usually estimate the treatment effect by regressing their outcome on the treatment and pair or stratum fixed effects, "clustering" their standard errors at the unit-of-randomization level. Then, they typically use the 5%-level t-test based on this regression to determine if the treatment has an effect or not. As any statistical test, this t-test may lead them to commit a type 1 error. Specifically, it may lead them to wrongly conclude that the treatment has an effect, while the truth is that the treatment does not have an effect. But when using a 5%-level test, their hope is that the probability that this would happen, the so-called error rate of the test, is no larger than 5%. We show that unfortunately, the error rate of this t-test may be much larger than the researcher's 5% target. We then show that to achieve their desired error rate, researchers should cluster their standard errors at the pair or at the strata level, rather than at the unit-of-randomization level. Clustering at the pair rather than at the unit level in a sample of 371 regressions from published paired RCTs reduces the number of significant effects by 1/3.

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.

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. (2019), 'Optimality of matched-pair designs in randomized controlled trials', Available at SSRN 3483834 .

Bai, Y., Romano, J. P. and Shaikh, A. M. (2021), 'Inference in experiments with matched pairs', Journal of the American Statistical Association pp. 1–37.

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. (2015a), 'One laptop per child at home: Short-term impacts from a randomized experiment in peru', American Economic Journal: Applied Economics **7**(2), 53–80.

Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O. and Cruz-Aguayo, Y. (2015b), 'Replication data for: 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. (2016a), '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., Leão, L. d. S., Legovini, A., Marchetti, R. and Zia, B. (2016b), 'Replication data for: 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.

Carter, A. V., Schnepel, K. T. and Steigerwald, D. G. (2017), 'Asymptotic behavior of at-test robust to cluster heterogeneity', Review of Economics and Statistics **99**(4), 698–709.

Crépon, B., Devoto, F., Duflo, E. and Parienté, W. (2015a), '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.

Crépon, B., Devoto, F., Duflo, E. and Parienté, W. (2015b), 'Replication data for: 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 .

Fogarty, C. B. (2018), 'On mitigating the analytical limitations of finely stratified experiments', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **80**(5), 1035–1056.

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

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

Imai, K. (2008), 'Variance identification and efficiency analysis in randomized experiments under the matched-pair design', Statistics in Medicine **27**(24), 4857–4873.

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.

King, G., Gakidou, E., Ravishankar, N., Moore, R. T., Lakin, J., Vargas, M., Téllez-Rojo, M. M., Hernández Ávila, J. E., Ávila, M. H. and Llamas, H. H. (2007), 'A "politically robust" experimental design for public policy evaluation, with application to the mexican universal health insurance program', Journal of Policy Analysis and Management **26**(3), 479–506.

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.

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.

For Online Publication

A Proof of Lemma 3.1

We first introduce some notation. 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. Let $SET_p = \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} W_{gp}\epsilon_{igp}$ and $SEU_p = \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}(1-W_{gp})\epsilon_{igp}$ respectively be the sum of the residuals ϵ_{igp} for the treated and untreated observations in pair p.

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

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

Remember that $\widehat{\tau}_p = \sum_{g=1}^{2}\left[W_{gp}\sum_{i=1}^{n_{gp}}\frac{Y_{igp}}{n_{gp}} - (1-W_{gp})\sum_{i=1}^{n_{gp}}\frac{Y_{igp}}{n_{gp}}\right]$ is the difference between the average outcome of treated and untreated observations in pair p. It follows from, e.g., Equation (3.3.7) in Angrist and Pischke (2008) and a few lines of algebra that

$$\widehat{\tau}_{fe} = \sum_{p=1}^{P}\omega_p\widehat{\tau}_p, \qquad \text{where} \quad \omega_p = \frac{\left(n_{1p}^{-1} + n_{2p}^{-1}\right)^{-1}}{\sum_{p'=1}^{P}\left(n_{1p'}^{-1} + n_{2p'}^{-1}\right)^{-1}}.$$

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

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

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

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

Then,

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

The first equality follows from Point 1 of Lemma C.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,

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

where the first equality follows from Equation (37) in the proof of Lemma C.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

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

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 τ and $\tau_p = \mathbb{E}[\widehat{\tau}_p]$,

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

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

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

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

QED.

Point 2

The result directly follows from Points 3 and 4 of Lemma C.1 and the fact that $n_{1p} = n_{2p} = n_p/2$ under Assumption 2.

QED.

Point 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 $\widehat{Y}(d) \equiv \sum_p \widehat{Y}_p(d)/P$, for $d \in \{0,1\}$.

$$\mathbb{E}[\widehat{Y}_p(1)] = \mathbb{E}\left[\sum_g W_{gp}\overline{y}_{gp}(1)\right] = \frac{1}{2}\sum_g \overline{y}_{gp}(1) = \overline{y}_p(1). \tag{8}$$

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

$$\mathbb{E}[\widehat{Y}_p(0)] = \mathbb{E}[\overline{y}_p(0)] \tag{9}$$

$$\mathbb{E}[\widehat{Y}(d)] = \overline{y}(d), \quad \text{for } d \in \{0,1\}. \tag{10}$$

Then, one has

$$\begin{aligned}
\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) - \widehat{\mathbb{V}}_{pair}(\widehat{\tau}) &= \frac{8}{n^2}\sum_p SET_p SEU_p \\
&= \frac{8}{n^2}\sum_p \left(\sum_g W_{gp}\sum_i (y_{igp}(1) - \widehat{Y}(1))\right)\left(\sum_g (1-W_{gp})\sum_i (y_{igp}(0) - \widehat{Y}(0))\right) \\
&= \frac{8}{n^2}\sum_p \frac{n_p^2}{4}\left(\sum_g W_{gp}\sum_i \frac{y_{igp}(1)}{n_{gp}} - \widehat{Y}(1)\right)\left(\sum_g (1-W_{gp})\sum_i \frac{y_{igp}(0)}{n_{gp}} - \widehat{Y}(0)\right) \\
&= \frac{2}{P^2}\sum_p \widehat{Y}_p(1)\widehat{Y}_p(0) - \frac{2}{P}\widehat{Y}(1)\widehat{Y}(0)
\end{aligned} \tag{11}$$

The first equality follows from Points 1 and 2 of Lemma C.1 and Assumption 2. The second equality follows from the definitions of SET_p, SEU_p, and ϵ_{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 (11),

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

The first equality follows from adding and subtracting $\frac{2}{P}\,\mathbb{E}[\widehat{Y}(1)]\,\mathbb{E}[\widehat{Y}(0)]$ and $\frac{2}{P^2}\sum_p \mathbb{E}[\widehat{Y}_p(1)]\,\mathbb{E}[\widehat{Y}_p(0)]$, and from Equations (8), (9) and (10). The third equality follows from Point 3 of Assumption 1. Therefore,

$$\frac{P}{P-1}\,\mathbb{E}\left[\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) - \widehat{\mathbb{V}}_{pair}(\widehat{\tau})\right] = \frac{2}{P^2}\sum_p \Big(\mathrm{Cov}(\widehat{Y}_p(1), \widehat{Y}_p(0))\Big) + \frac{2}{P(P-1)}\sum_p (\overline{y}_p(0) - \overline{y}(0))(\overline{y}_p(1) - \overline{y}(1)). \tag{12}$$

Finally,

$$\begin{aligned}
\mathrm{Cov}\left(\widehat{Y}_p(1), \widehat{Y}_p(0)\right) &= \mathbb{E}[\widehat{Y}_p(1)\widehat{Y}_p(0)] - \mathbb{E}[\widehat{Y}_p(1)]\,\mathbb{E}[\widehat{Y}_p(0)] \\
&= \left(\frac{1}{2}\overline{y}_{1p}(1)\overline{y}_{2p}(0) + \frac{1}{2}\overline{y}_{2p}(1)\overline{y}_{1p}(0)\right) - \left(\frac{1}{2}\sum_g \overline{y}_{gp}(1)\right)\left(\frac{1}{2}\sum_g \overline{y}_{gp}(0)\right) \\
&= \frac{1}{4}\overline{y}_{1p}(1)\overline{y}_{2p}(0) + \frac{1}{4}\overline{y}_{2p}(1)\overline{y}_{1p}(0) - \frac{1}{4}\overline{y}_{1p}(1)\overline{y}_{1p}(0) - \frac{1}{4}\overline{y}_{2p}(1)\overline{y}_{2p}(0) \\
&= \frac{1}{4}\left(\overline{y}_{1p}(1) - \overline{y}_{2p}(1)\right)\left(\overline{y}_{2p}(0) - \overline{y}_{1p}(0)\right) \\
&= -\frac{1}{2}\sum_g \left(\overline{y}_{gp}(0) - \overline{y}_p(0)\right)\left(\overline{y}_{gp}(1) - \overline{y}_p(1)\right)
\end{aligned} \tag{13}$$

The second equality follows from Points 1 and 2 of Assumption 1, and Equations (8) and (9). The third, fourth, and fifth equalities follow after some algebra. The result follows plugging Equation (13) into (12).

QED.

B Large sample results for the pair- and unit-clustered variance estimators

In this section, we present the large sample distributions of the t-tests attached to the four variance estimators we considered in Section 3. Let

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

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

Assumption 3.

1. For every d, g and p, there is a constant M such that $\left|\overline{y}_{gp}(d)\right| < M < +\infty$.
2. When $P \to +\infty$, $\frac{1}{P}\sum_p \tau_p$, $\frac{1}{P}\sum_p(\tau_p - \tau)^2$, 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 $(\widehat{\tau}_p)_{p=1}^{+\infty}$.

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

1. $(\widehat{\tau} - \tau)/\sqrt{\widehat{\mathbb{V}}_{pair}(\widehat{\tau})} = (\widehat{\tau}_{fe} - \tau)/\sqrt{\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})} \xrightarrow{d} \mathcal{N}(0, \sigma_{pair}^2)$. $\sigma_{pair}^2 \leq 1$, *and if* $\tau_p = \tau$ *for every* p, $\sigma_{pair}^2 = 1$.
2. $(\widehat{\tau}_{fe} - \tau)/\sqrt{\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})} \xrightarrow{d} \mathcal{N}(0, 2\sigma_{pair}^2)$.
3. $(\widehat{\tau} - \tau)/\sqrt{\widehat{\mathbb{V}}_{unit}(\widehat{\tau})} \xrightarrow{d} \mathcal{N}(0, \sigma_{unit}^2)$.
4. $\sigma_{unit}^2 \leq \sigma_{pair}^2$ *if and only if* $\Delta_{cov,P}$ *converges towards a positive limit.*

Proof. See Web Appendix H.

Point 3 is related to Theorem 3.1 in Bai et al. (2021), who show that when $n_{gp} = 1$, the t-test in Point 3 under-rejects. The asymptotic variance we obtain is different from theirs, because our results are derived under different assumptions. For instance, we assume a fixed population, while Bai et al. (2021) 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.

C Clustered variance estimators

Lemma C.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 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* $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P} \omega_p^2 \left(\widehat{\tau}_p - \widehat{\tau}_{fe} \right)^2$.

4. *The UCVE of* $\widehat{\tau}_{fe}$ *is* $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P} \omega_p^2 \left(\widehat{\tau}_p - \widehat{\tau}_{fe} \right)^2 \left(\left(\frac{n_{1p}}{n_p} \right)^2 + \left(\frac{n_{2p}}{n_p} \right)^2 \right)$.

Proof. See Web Appendix H. □

D Variance estimators that rely on pairs of pairs

We also study two other estimators of $\mathbb{V}(\widehat{\tau})$. Those estimators have been proposed in the one-observation-per-unit special case, but it is straightforward to extend them to the case where all units have the same number of observations, as stated in Assumption 2.[13]

The first alternative estimator we consider is a slightly modified version of the pairs-of-pairs (POP) 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 Subsection D.4 below 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 POP 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 POPs, 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 covariate is used. In our application in subsection D.4, we use the baseline outcome to match pairs into POPs, 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 D.1, we will argue below that the baseline outcome should often be a good choice to match pairs into POPs. The variable one uses to form POPs 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 lead to incorrect inference.

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

There are two differences between the POPVE 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 is matched to pairs 1 and 3 in Abadie and Imbens (2008), while $\widehat{\mathbb{V}}_{pop}(\widehat{\tau})$ matches pair 1 to pair 2 and pair 3 to pair 4. Matching without replacement makes the properties of $\widehat{\mathbb{V}}_{pop}(\widehat{\tau})$ easier to analyze.

The second alternative variance estimator we consider is that proposed by Bai et al. (2021) 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. (2021) propose another variance estimator in their Equation (28). 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.

D.1 Finite-sample results

Let $\tau_{\cdot r} = \frac{1}{2}(\tau_{1r} + \tau_{2r})$ denote the average treatment effect in POP r.

Lemma D.1. *If Assumptions 1 and 2 hold and P is even,*

1. $\mathbb{E}\left[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})\right] = \mathbb{V}(\widehat{\tau}) + \frac{1}{P^2}\sum_{r=1}^{R}(\tau_{1r} - \tau_{2r})^2$.
2. $\widehat{\mathbb{V}}_{brs}(\widehat{\tau}) = \frac{1}{2}\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) + \frac{1}{2}\widehat{\mathbb{V}}_{pop}(\widehat{\tau})$.
3. *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 Web Appendix H. □

Point 1 of Lemma D.1 shows that the POPVE is upward biased in general, and unbiased if the treatment effect is constant within POP. The less treatment effect heterogeneity within POP, the less upward biased the POPVE. An important practical consequence of Point 1 is that the variable used to form POPs 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., 2016a).

Point 1 of Lemma D.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 conditional variance of $\widehat{\tau}$.[14] The fact that the POPVE is upward biased in Lemma D.1 and consistent in their Theorem 1 is because we do not assume that the experimental units are an i.i.d. sample from a super population. The intuition is the following. In Abadie and Imbens (2008), when the number of units grows, the covariates X_i on which pairing is based become equal to the same value x for units in the same POP: with an infinity of units, each unit can be matched to another unit with the same X_i, and each pair can be matched to another pair with the same X_i. Then, asymptotically those units are an i.i.d. sample drawn from the super-population conditional on $X_i = x$, and they all have the same expectation of their treatment effect. Treatment effect heterogeneity within POPs, the source of the POPVE's upward bias in Lemma D.1, vanishes asymptotically. On the other hand, with a convenience sample, units in the same POP may have asymptotically the same covariates, but they could still have different treatment effects, because they are not i.i.d. draws from a superpopulation.

Point 2 shows that the BRSVE is equal to the average of the PCVE and POPVE. Then, it follows from Point 1 of Lemma 3.1 and Point 1 of Lemma D.1 that $\frac{P}{P-1}\widehat{\mathbb{V}}_{brs}(\widehat{\tau})$ is upward biased. Point 2 is related to Lemma 6.4 and Theorem 3.3 in Bai et al. (2021), 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 upward biased in Lemma D.1 and consistent in Bai et al. (2021) 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 3 shows that if the treatment effect varies less within than across POPs, the

[14] In our setting, the covariates are assumed to be fixed, so the fact that we consider the unconditional variance of $\widehat{\tau}$ while they consider its conditional variance does not explain the difference between our results.

POPVE is less upward biased than the degrees-of-freedom-adjusted PCVE and BRSVE, and the BRSVE is less upward biased than the degrees-of-freedom-adjusted PCVE. A sufficient condition to have that the treatment effect varies less within than across POPs is $\frac{1}{R}\sum_{r=1}^{R}(\tau_{1r}-\tau)(\tau_{2r}-\tau)\geq 0$, meaning that the treatment effects of the two pairs in the same POP are positively correlated.

D.2 Large-sample results

Assumption 4. When $P\to+\infty$, $\frac{1}{P}\sum_r(\tau_{1r}-\tau_{2r})^2$ converges towards a finite limit.

Let

$$\sigma^2_{pop}=\lim_{P\to+\infty}\frac{P\mathbb{V}(\widehat{\tau})}{P\mathbb{V}(\widehat{\tau})+\frac{1}{P}\sum_r(\tau_{1r}-\tau_{2r})^2},$$
$$\sigma^2_{brs}=\lim_{P\to+\infty}\frac{P\mathbb{V}(\widehat{\tau})}{P\mathbb{V}(\widehat{\tau})+\frac{1}{2P}\sum_r(\tau_{1r}-\tau_{2r})^2+\frac{1}{2P}\sum_p(\tau_p-\tau)^2},$$

where Assumptions 3 and 4 ensure the limits in the previous display exist.

Theorem D.2. *(t-stats' asymptotic behavior) Under Assumptions 1, 2, 3, and 4,*

1. $(\widehat{\tau}-\tau)/\sqrt{\widehat{\mathbb{V}}_{pop}(\widehat{\tau})}\xrightarrow{d}\mathcal{N}(0,\sigma^2_{pop})$. $\sigma^2_{pop}\leq 1$, *and if* $\tau_{1r}=\tau_{2r}$ *for every* r, $\sigma^2_{pop}=1$.
2. $(\widehat{\tau}-\tau)/\sqrt{\widehat{\mathbb{V}}_{brs}(\widehat{\tau})}\xrightarrow{d}\mathcal{N}(0,\sigma^2_{brs})$. $\sigma^2_{brs}\leq 1$, *and if* $\tau_p=\tau$ *for every* p, $\sigma^2_{brs}=1$.
3. $\sigma^2_{pair}\leq\sigma^2_{brs}\leq\sigma^2_{pop}$ *if and only if* $0\leq\lim_{P\to+\infty}\frac{1}{R}\sum_{r=1}^{R}(\tau_{1r}-\tau)(\tau_{2r}-\tau)$.

Proof. See Web Appendix H.

Points 1 and 2 of Theorem D.2 show that when the number of pairs grows, the t-statistic using the POPVE and BRSVE, respectively, 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 under-reject. Point 3 shows that whenever there is a positive correlation between the treatment effects of the two pairs in the same POP, the t-test using the POPVE under-rejects less than that using the BRSVE, which itself under-rejects less than that using the PCVE.

D.3 Simulations

For 26 of the 82 regressions in Crépon et al. (2015a), 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 error rate of each t-test is shown in Table 3 below. The error rate of the t-test using the PCVE is close to 5% with as few as 20 pairs. On the other hand, the error rates of the t-tests using the POPVE and BRSVE are larger than 5%, even with 80 pairs. Accordingly, we run simulations again, duplicating the random sample of 80 pairs twice to have 160 pairs. The error rate of the t-test using the BRSVE is now close to 5%, but the error rate of the t-test using the POPVE is still larger than 5%. With a sample of 320 pairs obtained by duplicating the random sample of 80 pairs four times, all tests have error rates close to 5%. 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 $\widehat{\mathbb{V}}_{pop}(\tau)$ over-rejects, 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 rejects more often. With 160 and 320 pairs, this phenomenon becomes less pronounced. Overall, the asymptotic approximations in Points 1 and 2 of Theorem D.2 seem to hold only with a large number of pairs, contrary to that in Point 1 of Theorem B.1.

Table 3: Simulations with data aggregated at village-level to compute $\widehat{\mathbb{V}}_{pop}$ and $\widehat{\mathbb{V}}_{brs}$

Variance estimator	5% level t-test error rate			
	With 20 pairs	With 80 pairs	With 160 pairs	With 320 pairs
PCVE	0.0499	0.0508	0.0509	0.0502
POPVE	0.1306	0.0833	0.0647	0.0565
BRSVE	0.0807	0.0624	0.0569	0.0526

Table 3 reports the error rates of three 5% level t-tests in Crépon et al. (2015a), aggregating 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 pair-clustered variance estimator (PCVE), the pairs-of-pairs variance estimator (POPVE) in Abadie and Imbens (2008), the variance estimator in Bai et al. (2021) (BRSVE), and the three corresponding t-tests. The error rate of each test is the percent of times, across the 78,000 regressions (26 outcomes $\times$ 3,000 replications), that the test leads the researcher to wrongly conclude that the treatment has an effect. 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).

D.4 Application

For 152 of the 294 regressions in Panel A of Table 2, the baseline outcome is available in the data set, so we can compute the POPVE and BRSVE. 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 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 1.026. The BRSVE divided by the PCVE is on average equal to 1.014.[15] In those regressions, the POPVE and BRSVE do not lead to power gains.

E Extension: stratified experiments with few units per strata

In this section, we perform Monte-Carlo simulations to assess how our results in Section 3 extend to stratified RCTs where the number of units per strata is larger than two, but still fairly small. Three main findings emerge. First, the error rates of t-tests using stratum-clustered standard errors are equal to 5%. Second, the error rates of t-tests using standard errors clustered at the unit level are larger than 5% in regressions with stratum fixed effects, but decrease as the number of units per strata increases. With 5 units per strata, and averaging across Panels A to D of Table 4 below, the error rate of a 5% level test with UCVE and stratum fixed effects is around 7.9%, while with 10 units per strata this error rate is around 6.2%. Finally, the error rates of t-tests using standard errors clustered at the unit level tend to be lower than 5% in regressions without stratum fixed effects.

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

$$Y_{igp} = W_{gp} y_{igp}(1) + (1 - W_{gp}) y_{igp}(0) + \gamma_p, \qquad i = 1, \ldots, n_{gp};\ g = 1, \ldots, G;\ p = 1, \ldots, P, \quad (14)$$

where $y_{igp}(1)$ and $y_{igp}(0)$ are independent and both follow a $\mathcal{N}(0,1)$ distribution, $\{\gamma_p\}_p \sim$ iid $\mathcal{N}(0, \sigma_\gamma^2)$,

[15] The variance estimator in Equation (28) of Bai et al. (2021) is also on average higher than the PCVE.

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 γ_p do not vary across simulations.

Each stratum has G 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 the number of strata P be equal to 100.

We compute t-tests based on unit- and stratum-clustered standard errors in regressions of the outcome on the treatment with and without stratum fixed effects. We perform 10,000 simulations for each DGP. Table 4 presents the error rates of the t-tests in each DGP.

t-tests using stratum-clustered standard errors achieve error rates close to 5% for all data configurations (as in Table 1, with $n_{gp} = 5$, the $t-$test using the PCVE with stratum fixed effects under-rejects slightly, due to the DOF-adjustment). In contrast, t-tests based on unit-clustered standard errors in regressions with stratum fixed effects overreject the true null of no treatment effect. These results are in line with Points 1 and 2 of Theorem B.1, which covered the special case where $G = 2$. t-tests based on unit-clustered standard errors in regressions with stratum fixed effects over-reject 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 3.1, 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 stratum fixed effects have error rates close to 5 %. When $\sigma_\eta = 0$, there is no between and within strata heterogeneity in $\overline{y}_{gp}(0)$, so it follows from Point 3 of Theorem B.1 that in the special case where $G = 2$, t-tests based on unit-clustered standard errors in regressions without stratum fixed effects have error rates close to 5%. 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 stratum fixed effects have error rates lower than 5%, because there is now between-strata heterogeneity in $\overline{y}_{gp}(0)$. We obtain similar results with five observations per unit (Panels C and D).

Table 4: Error rates of t-test in simulated stratified RCTs 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	0.0311	0.0462	0.0607	0.0552	0.0533	0.0566	0.0562	0.0471	0.0490
UCVE with FE	0.1655	0.1096	0.0937	0.0852	0.0734	0.0710	0.0684	0.0631	0.0624
SCVE without FE	0.0509	0.0495	0.0576	0.0554	0.0532	0.0520	0.0553	0.0505	0.0522
SCVE with FE	0.0503	0.0488	0.0571	0.0553	0.0532	0.0518	0.0551	0.0505	0.0521
$\frac{\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})}{\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})}$	0.7053	0.8168	0.8694	0.8976	0.9187	0.9326	0.9420	0.9488	0.9557
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	0.1750	0.1178	0.0808	0.0804	0.0768	0.0655	0.0663	0.0678	0.0654
SCVE without FE	0.0563	0.0547	0.0449	0.0536	0.0564	0.0477	0.0538	0.0545	0.0543
SCVE with FE	0.0557	0.0541	0.0446	0.0535	0.0564	0.0473	0.0535	0.0543	0.0542
$\frac{\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})}{\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})}$	0.7053	0.8165	0.8692	0.8984	0.9177	0.9314	0.9412	0.9488	0.9552
Panel C. iid standard normal potential outcomes and $n_{gp} = 5$									
UCVE without FE	0.0853	0.0431	0.0428	0.0536	0.0540	0.0526	0.0553	0.0472	0.0491
UCVE with FE	0.1495	0.0951	0.0860	0.0736	0.0724	0.0632	0.0658	0.0612	0.0601
SCVE without FE	0.0556	0.0497	0.0511	0.0535	0.0546	0.0508	0.0573	0.0521	0.0510
SCVE with FE	0.0425	0.0432	0.0458	0.0488	0.0521	0.0482	0.0530	0.0493	0.0491
$\frac{\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})}{\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})}$	0.7053	0.8166	0.8702	0.8981	0.9188	0.9312	0.9414	0.9486	0.9544
Panel D. Stratum-level shock affecting potential outcomes and $n_{gp} = 5$									
UCVE without FE	0.0213	0.0166	0.0174	0.0178	0.0141	0.0200	0.0141	0.0190	0.0133
UCVE with FE	0.1475	0.1027	0.0858	0.0781	0.0626	0.0617	0.0668	0.0631	0.0588
SCVE without FE	0.0503	0.0544	0.0538	0.0546	0.0481	0.0478	0.0534	0.0545	0.0520
SCVE with FE	0.0392	0.0476	0.0489	0.0507	0.0440	0.0438	0.0508	0.0519	0.0498
$\frac{\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})}{\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})}$	0.7053	0.8168	0.8701	0.8991	0.9183	0.9311	0.9408	0.9485	0.9544

The table shows the error rates 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 units per strata from two to ten ($G = 2, \ldots, 10$); we vary the number of observations per 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 Panels A and C, and a model with an additive stratum-level shock affecting both potential outcomes in Panel B and D. UCVE and SCVE stand for unit- and stratum-clustered variance estimators, respectively. FE stands for stratum fixed effects. $\frac{\widehat{s.e.}_{unit}(\widehat{\tau}_{fe})}{\widehat{s.e.}_{strat}(\widehat{\tau}_{fe})}$ is the average across simulations of the ratio of standard errors clustering at the unit and stratum levels in regressions with stratum fixed effects.

F Articles in our survey of paired or small strata experiments

Table 5: Paired RCTs and stratified RCTs with small strata

Reference	Search source
Paired RCTs	
Ashraf et al. (2006)	AEA registry
Banerjee et al. (2015)	*AEJ: Applied*
Crépon et al. (2015a)	*AEJ: Applied*
Beuermann et al. (2015a)[1]	*AEJ: Applied*
Fryer Jr et al. (2016)	AEA registry
Glewwe et al. (2016a)	AEA registry
Bruhn et al. (2016a)	*AEJ: Applied*
Fryer Jr (2017)	AEA registry
Small-strata RCTs	
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 clustered and paired RCTs, or clustered and stratified RCTs 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 clustered and paired RCT and seven had conducted a clustered and stratified RCT with ten units or less per strata. We also searched the AEA's registry website for RCTs (`https://www.socialscienceregistry.org`). We looked at all completed projects, whose randomization method included the word "pair" and that had either a working or a published paper. Thus, we found four more papers that had conducted a clustered and paired RCT. Beuermann et al. (2015a) 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.

G Results when the number of observations varies across units

In this section, we extend some of the results in Section 3 to instances where units may have different numbers of observations, as is often the case in practice.

G.1 Upward bias of the pair-clustered variance estimator (PCVE)

In this subsection, we show that when units have different numbers of observations, our recommendation of using the PCVE still applies.

When units have different numbers of observations, there are several estimators of the treatment effect one may consider. $\widehat{\tau}$, the standard difference in means estimator, is such that

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

where T and C respectively denote the total number of treated and control observations. When the number of observations varies across units, T and C are stochastic. For instance, assume one has two pairs. In pair 1, units 1 and 2 both have 1 observation, but in pair 2 unit 1 has 1 observations while unit 2 has 2 observations. Then, T is equal to 2 with probability 1/2, and to 3 with probability 1/2. These stochastic denominators in $\widehat{\tau}$ make it impossible to derive a closed-form expression of its expectation and variance. One can still show that when the number of pairs goes to infinity, $\widehat{\tau}$ converges toward τ, the average treatment effect, and one could also use the delta method to show that $\widehat{\tau}$ is asymptotically normal and derive its asymptotic variance. However, throughout the paper we have focused on estimators' finite sample variances.

Therefore, instead of $\widehat{\tau}$ we consider another, closely related estimator, whose expectation and variance are straightforward to derive even when the number of observations varies across units, and which is unbiased for a causal effect that differs from τ but that is still relatively natural (see Imai et al. (2009) for closely related discussions). Let $\widetilde{\tau}$ denote the coefficient of W_{gp} in the weighted OLS regression of Y_{igp} on a constant and W_{gp}, with weights $V_{gp} = n_p/n_{gp}$.[16] Let $\widetilde{\alpha}$ be the intercept in that regression. One can show that

$$\widetilde{\tau} = \frac{1}{P}\sum_{p}\frac{n_p}{\overline{n}}\sum_{g}(W_{gp}\overline{Y}_{gp} - (1-W_{gp})\overline{Y}_{gp}) = \frac{1}{P}\sum_{p}\frac{n_p}{\overline{n}}\widehat{\tau}_p, \tag{15}$$

where $\bar{n} = n/P$. Under Assumption 2, $\widetilde{\tau} = \widehat{\tau}$. Hence, $\widetilde{\tau}$ generalizes $\widehat{\tau}$ to the case where the number

[16] Specifically, the intercept $\widetilde{\alpha}$ and $\widetilde{\tau}$ are such that $(\widetilde{\alpha}, \widetilde{\tau}) = \text{argmin}_{\alpha,\tau}\sum_p\sum_g\sum_i V_{gp}(Y_{igp} - \alpha - \tau W_{gp})^2$.

of observations varies across units. $\widetilde{\tau}$ is also one of the estimators considered by Imai et al. (2009), though the fact $\widetilde{\tau}$ can be obtained by weighted least squares is not noted therein.

$\widetilde{\tau}$ is generally not unbiased for τ, unless in every pair, the two units have the same number of observations, i.e., $n_{1p} = n_{2p}$ for all p (Imai et al., 2009). On the other hand, $\widetilde{\tau}$ is unbiased for

$$\tau^* = \frac{1}{P} \sum_p \frac{n_p}{\overline{n}} \left(\frac{\tau_{1p}}{2} + \frac{\tau_{2p}}{2} \right)$$

where $\tau_{gp} = \frac{1}{n_{gp}} \sum_{i=1}^{n_{gp}} [y_{igp}(1) - y_{igp}(0)]$ denotes the average treatment effect in unit g of pair p.[17] τ^* is a weighted average of the pair-specific average treatment effects $(\tau_{1p} + \tau_{2p})/2$. Those pair-specific average treatment effects give equal weight to the average treatment effect in each unit, rather than weighting them according to their number of observations like τ_p does.Imai et al. (2009) show that

$$\mathbb{V}(\widetilde{\tau}) = \frac{1}{4P^2} \sum_p \frac{n_p^2}{\overline{n}^2} \left(\Delta_p(1) + \Delta_p(0) \right)^2,$$

where $\Delta_p(1) \equiv \overline{y}_{1p}(1) - \overline{y}_{2p}(1)$ and $\Delta_p(0) \equiv \overline{y}_{1p}(0) - \overline{y}_{2p}(0)$. They propose various estimators of that variance, and show that they are upward biased. Instead, we rely on the fact $\widetilde{\tau}$ can be obtained by weighted least squares to propose an estimator whose properties have not been studied in the randomization-inference framework we consider: the PCVE attached to $\widetilde{\tau}$.

First, the following lemma extends Lemma C.1 to the PCVE in a weighted OLS regression.[18]

Lemma G.1 (Pair-clustered variance estimator for $\widetilde{\tau}$)**.** $\widehat{\mathbb{V}}_{pair}(\widetilde{\tau}) = \frac{1}{P^2} \sum_p \frac{n_p^2}{\overline{n}^2} \left[\widehat{\tau}_p - \widetilde{\tau} \right]^2$.

Proof. See Web Appendix H.

Then, we study the asymptotic distribution of the t-statistic attached to $\widetilde{\tau}$ and $\widehat{\mathbb{V}}_{pair}(\widetilde{\tau})$. To do so, we make the following assumption.

Assumption 5.

1. For all g and p, $1 \leq n_{gp} \leq N$ for some fixed $N < +\infty$.
2. As $P \to +\infty$, $\frac{1}{P} \sum_p \left(\frac{n_p}{\overline{n}} \right)^2$, $\frac{1}{P} \sum_p \left(\frac{n_p}{\overline{n}} \right)^2 \mathbb{E}[\widehat{\tau}_p]$, and $\frac{1}{P} \sum_p \left(\frac{n_p}{\overline{n}} \right)^2 (\mathbb{E}[\widehat{\tau}_p])^2$ converge to strictly positive constants, and $\tau^* = \frac{1}{P} \sum_p \frac{n_p}{\overline{n}} \left(\frac{\tau_{1p}}{2} + \frac{\tau_{2p}}{2} \right)$ converges to a constant τ^∞.

[17] With a slight abuse of notation, τ_{1r} and τ_{2r} refer to the ATE in pairs 1 and 2 of POP r, while τ_{1p} and τ_{2p} refer to the ATE in units 1 and 2 of pair p.

[18] We follow the definition of clustered variance estimators for weighted least squares in Equation (15) of Cameron and Miller (2015).

3. As $P \to +\infty$, $\sum_{p=1}^{P} \mathbb{E}\left[|\frac{n_p}{\overline{n}}|^{2+\epsilon}|\widehat{\tau}_p - \mathbb{E}[\widehat{\tau}_p]|^{2+\epsilon}\right] / \widetilde{S}_P^{2+\epsilon} \to 0$ for some $\epsilon > 0$, where $\widetilde{S}_P^2 \equiv P^2\mathbb{V}(\widetilde{\tau})$.

Point 1 of Assumption 5 requires that the number of observations in every unit is greater than 1 and lower than some fixed N. Combined with Point 2 of Assumption 3, Point 2 of Assumption 5 ensures that $P\widehat{\mathbb{V}}_{pair}(\widetilde{\tau})$ converges towards a strictly positive limit. Point 3 guarantees that we can apply the Lyapunov central limit theorem to $\left(\frac{n_p}{\overline{n}}\widehat{\tau}_p\right)_{p=1}^{+\infty}$. Let $\sigma_{wls}^2 = \lim_{P\to+\infty} \frac{P\mathbb{V}(\widetilde{\tau})}{P\mathbb{V}(\widetilde{\tau}) + \frac{1}{P}\sum_p \left(\frac{n_p}{\overline{n}}\right)^2 (\mathbb{E}(\widehat{\tau}_p) - \tau^{\infty})^2}$.

Theorem G.2. *If Assumptions 1 and 5, and Points 1 and 2 of Assumption 3 hold,* $(\widetilde{\tau} - \tau^*)/\sqrt{\widehat{\mathbb{V}}_{pair}(\widetilde{\tau})} \xrightarrow{d} \mathcal{N}(0, \sigma_{wls}^2)$. $\sigma_{wls}^2 \leq 1$, *and if* $\tau_{gp} = \tau$ *for every* g *and* p, *or if* $n_{1p} = n_{2p}$ *and* $\tau_p = \tau$ *for every* p, *then* $\sigma_{wls}^2 = 1$.

Proof. See Web Appendix H.

This theorem shows that when the number of pairs grows, the t-statistic of the weighted least squares estimator using the PCVE 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 units, or when the treatment effect is homogenous across pairs and in every pair the two units have the same number of observations.

Theorem G.2 shows that when units have different numbers of observations, the PCVE attached to $\widetilde{\tau}$ is upward biased asymptotically. We now show that the same holds for $\widehat{\tau}_{fe}$, the pair fixed effects estimator, provided one applies some kind of degrees-of-freedom correction to its PCVE. As shown in Point 3 of Lemma C.1, the PCVE of $\widehat{\tau}_{fe}$ is $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P} \omega_p^2 \left(\widehat{\tau}_p - \widehat{\tau}_{fe}\right)^2$. Let $\widetilde{\omega}_p = \omega_p \left(1 - 2\omega_p\right)^{-1/2}$.

Lemma G.3 (The adjusted PCVE for $\widehat{\tau}_{fe}$ is upward biased)**.** *Under Assumption 1, and if* $\omega_p \leq 1/2$ *for all* p, $\mathbb{E}\left[\sum_p \widetilde{\omega}_p^2 (\widehat{\tau}_p - \widehat{\tau}_{fe})^2\right] = \mathbb{V}(\widehat{\tau}_{fe})\left(1 + \sum_p \widetilde{\omega}_p^2\right) + \sum_p \widetilde{\omega}_p^2 [\mathbb{E}(\widehat{\tau}_p - \widehat{\tau}_{fe})]^2$.

Proof. See Web Appendix H. □

Lemma G.3 shows that the adjusted PCVE, where the ω_p are replaced by $\widetilde{\omega}_p$, is upward biased for the variance of $\widehat{\tau}_{fe}$. The adjustment in $\widetilde{\omega}_p$ is similar to a degrees-of-freedom adjustment. In fact, under Assumption 2, the adjusted PCVE is equal to $\frac{P}{P-2}\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$. The requirement that $\omega_p \leq 1/2$ for all p is mild. For instance, if $n_{1p} = n_{2p}$ for all p, this only requires that every pair has fewer observations than all other pairs combined. If there is an integer L such that $n_p \leq L$ for every p, one can show that $\liminf_{P\to+\infty} \mathbb{E}\left[P\left(\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) - \mathbb{V}(\widehat{\tau}_{fe})\right)\right] \geq 0$: the unadjusted PCVE is also upward biased asymptotically. When the number of observations varies across units, $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ does not

coincide with the estimator of the variance of $\widehat{\tau}_{fe}$ considered in Imai et al. (2009). It seems that Lemma G.3 above is the first result to justify the use of the PCVE attached to $\widehat{\tau}_{fe}$, in paired RCTs where the number of observations varies across units.

G.2 Ratio of the UCVE and PCVE with pair fixed effects

In this subsection, we derive the ratio of the UCVE and PCVE with pair fixed effects when units have different numbers of observations.

Lemma G.4 (Ratio of the UCVE and PCVE with pair fixed effects when units have different numbers of observations)**.** $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \sum_p \zeta_p \left(\left(\frac{n_{1p}}{n_p} \right)^2 + \left(\frac{n_{2p}}{n_p} \right)^2 \right)$*, where, for all* p $\zeta_p \geq 0$ *and* $\sum_p \zeta_p = 1$*. Therefore,* $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) \in \left[\frac{1}{2}, 1\right]$.

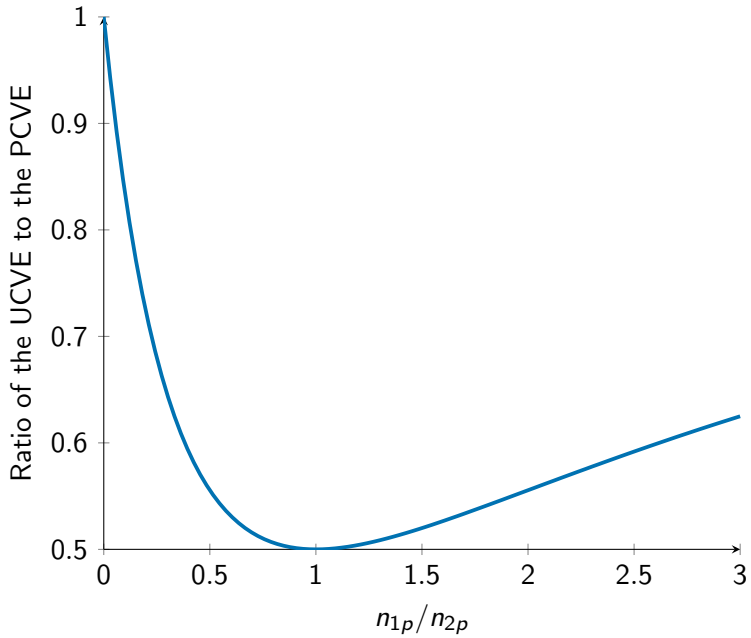
Proof. The formula for $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ follows from Points 3 and 4 of Lemma C.1, with

$$\zeta_p = \frac{\omega_p^2 \left(\widehat{\tau}_p - \widehat{\tau}_{fe}\right)^2}{\sum_{p=1}^P \omega_p^2 \left(\widehat{\tau}_p - \widehat{\tau}_{fe}\right)^2}.$$

$n_{1p}^2 + n_{2p}^2 \leq (n_{1p} + n_{2p})^2$, so $\left(\frac{n_{1p}}{n_p} \right)^2 + \left(\frac{n_{2p}}{n_p} \right)^2 \leq 1$. $(n_{1p} - n_{2p})^2 = n_{1p}^2 - 2n_{1p}n_{2p} + n_{2p}^2 \geq 0$, so $2n_{1p}^2 + 2n_{2p}^2 \geq (n_{1p} + n_{2p})^2$, and $\left(\frac{n_{1p}}{n_p} \right)^2 + \left(\frac{n_{2p}}{n_p} \right)^2 \geq \frac{1}{2}$. Therefore, $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) \in \left[\frac{1}{2}, 1\right]$. □

Lemma G.4 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 the squared shares that each unit accounts for in the pair. The sum of these squared shares is included between a half and one, so this ratio is included between a half and one. Figure 2 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 close to 1/2 when n_{1p}/n_{2p} is included between 0.5 and 2, meaning that the first unit has between half and twice as many observations as the second one. For instance, if in every pair, one unit has twice as many observations as the other, then the ratio of the two variances is equal to 5/9. Based on Figure 2, one can also derive an upper bound for $\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})/\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$, when n_{1p}/n_{2p} varies across pairs. For instance, if in every pair, one unit has at most twice as many observations as the other, as should often be the case in practice, then the ratio of the two variances is at most equal to 5/9. Overall, Lemma G.4 shows that Point 2 of Lemma 3.1 still approximately holds when units in each pair have different numbers of observations, unless they have an extremely unbalanced number of observations.

Figure 2: Ratio of Unit-Clustered and Pair-Clustered Variance Estimators with Pair Fixed Effects



Note: UCVE and PCVE stand for unit- and pair clustered variance estimators, respectively. n_{1p} and n_{2p} are the number of observations in units 1 and 2 of pair p, respectively.

H Proofs of the results in the Web Appendix

H.1 Proof of Theorem B.1

The proof relies on Lemma H.1 and on the two equations below.

Using a similar reasoning as that used to show Equation (55) in the proof of Lemma H.1, one can show that

$$\mathbb{E}\left[\left|\widehat{Y}_p(d)\right|^{2+\epsilon}\right] \leq M_1 < +\infty. \tag{16}$$

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

By Lemma H.1, Assumption 2, and Point 2 of Assumption 3,

$$\widehat{\tau} = \frac{1}{P}\sum_p \widehat{\tau}_p \xrightarrow{\mathbb{P}} \lim_{P\to+\infty} \frac{1}{P}\sum_p \mathbb{E}[\widehat{\tau}_p] = \lim_{P\to+\infty} \frac{1}{P}\sum_p \tau_p = \lim_{P\to+\infty} \tau. \tag{17}$$

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 (55). 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 the Lyapunov 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,

$$(\widehat{\tau} - \tau)/\sqrt{\mathbb{V}(\widehat{\tau})} \xrightarrow{d} \mathcal{N}(0,1). \tag{18}$$

Then,

$$\begin{aligned} P\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) - P\mathbb{V}(\widehat{\tau}) &= \sum_{p=1}^P \frac{\widehat{\tau}_p^2}{P} - \widehat{\tau}^2 - \sum_{p=1}^P \frac{\mathbb{V}(\widehat{\tau}_p)}{P} \\ &= \sum_{p=1}^P \frac{\widehat{\tau}_p^2}{P} - \widehat{\tau}^2 - \sum_{p=1}^P \frac{\mathbb{E}[\widehat{\tau}_p^2] - \mathbb{E}[\widehat{\tau}_p]^2}{P} \\ &= \sum_{p=1}^P \frac{\widehat{\tau}_p^2 - \mathbb{E}[\widehat{\tau}_p^2]}{P} - \widehat{\tau}^2 + \sum_{p=1}^P \frac{\tau_p^2}{P} \qquad (19) \\ &\xrightarrow{\mathbb{P}} \lim_{P \to +\infty} \frac{1}{P} \sum_{p=1}^P (\tau_p - \tau)^2. \qquad (20) \end{aligned}$$

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 (19). $\sum_{p=1}^P \frac{\widehat{\tau}_p^2 - \mathbb{E}[\widehat{\tau}_p^2]}{P} \xrightarrow{\mathbb{P}} 0$ by Lemma H.1. Then, $\widehat{\tau}^2 \xrightarrow{\mathbb{P}} \lim_{P \to +\infty} \tau^2$ by Equation (17) and the continuous mapping theorem (CMT). Equation (20) follows from these facts, and from Point 2 of Assumption 3.

Given Equation (20), Point 2 of Assumption 3, the Slutsky Lemma and the CMT, as $P \to +\infty$,

$$\frac{\widehat{\tau} - \tau}{\sqrt{\widehat{\mathbb{V}}_{pair}(\widehat{\tau})}} = \frac{\widehat{\tau} - \tau}{\sqrt{\mathbb{V}(\widehat{\tau})}} \sqrt{\frac{P\mathbb{V}(\widehat{\tau})}{P\widehat{\mathbb{V}}_{pair}(\widehat{\tau})}} \xrightarrow{d} \mathcal{N}(0, \sigma_{pair}^2). \tag{21}$$

Finally, by Lemma 3.1, $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = \widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$, and by Assumption 2, $\widehat{\tau} = \widehat{\tau}_{fe}$.

QED.

Point 2

By Lemma 3, $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = 2\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe})$, so given Point 1 of this theorem, the result follows.

QED.

Point 3

$$
\begin{aligned}
&P\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) - P\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) \\
&= \frac{2}{P}\sum_p \widehat{Y}_p(1)\widehat{Y}_p(0) - 2\frac{1}{P}\sum_p \widehat{Y}_p(1)\frac{1}{P}\sum_p \widehat{Y}_p(0) \\
&\xrightarrow{\mathbb{P}} 2\lim_{P\to+\infty}\left\{\frac{1}{P}\sum_p \mathbb{E}[\widehat{Y}_p(1)\widehat{Y}_p(0)] - \mathbb{E}[\widehat{Y}(1)]\,\mathbb{E}[\widehat{Y}(0)]\right\} \\
&= 2\lim_{P\to+\infty}\frac{1}{P}\sum_p\left\{\left(\overline{y}_p(0)-\overline{y}(0)\right)\left(\overline{y}_p(1)-\overline{y}(1)\right) - \frac{1}{2}\sum_g(\overline{y}_{gp}(0)-\overline{y}_p(0))(\overline{y}_{gp}(1)-\overline{y}_p(1))\right\}. \quad (22)
\end{aligned}
$$

The first equality follows from Equation (11). 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 (16) and the Cauchy-Schwarz inequality, from the fact that $\mathbb{E}\left[\left|\widehat{Y}_p(d)\right|^{1+\epsilon/2}\right]$ 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 3. The result follows from Equations (22), (20), and (18), and a reasoning similar to that used to prove Equation (21). **QED**.

H.2 Proof of Lemma C.1

Point 1

First, we introduce the formulas for the PCVE and UCVE in a general linear regression. Let ϵ_{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))

$$
\widehat{\mathbb{V}}_{pair}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{p=1}^{P}\left(\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}\epsilon_{igp}\boldsymbol{X}_{igp}\right)\left(\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}\epsilon_{igp}\boldsymbol{X}_{igp}\right)'\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}. \quad (23)
$$

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

$$
\widehat{\mathbb{V}}_{unit}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{p=1}^{P}\sum_{g=1}^{2}\left(\sum_{i=1}^{n_{gp}}\epsilon_{igp}\boldsymbol{X}_{igp}\right)\left(\sum_{i=1}^{n_{gp}}\epsilon_{igp}\boldsymbol{X}_{igp}\right)'\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}. \quad (24)
$$

Subtract from Equation (1) the average outcome in the population $\overline{Y} \equiv \frac{1}{n}\sum_p\sum_g\sum_i Y_{igp} =$

$\widehat{\alpha} + \widehat{\tau}\overline{W} + \bar{\epsilon}$, where $\overline{W} \equiv \frac{1}{n}\sum_p \sum_g \sum_i W_{gp}$, and $\bar{\epsilon} \equiv \frac{1}{n}\sum_p \sum_g \sum_i \epsilon_{igp} = 0$ by construction. Then,

$$Y_{igp} - \overline{Y} = \widehat{\tau}(W_{gp} - \overline{W}) + \epsilon_{igp}. \tag{25}$$

Apply Equation (23) to the residuals and covariates of the regression defined by Equation (25).[19] Then,

$$\widehat{\mathbb{V}}_{pair}(\widehat{\tau}) = \frac{\sum_p \left[\sum_g (W_{gp} - \overline{W}) \sum_i \epsilon_{igp}\right]^2}{\left[\sum_p \sum_g \sum_i (W_{gp} - \overline{W})^2\right]^2}. \tag{26}$$

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

$$\begin{aligned}\sum_p \left[\sum_g (W_{gp} - \overline{W}) \sum_i \epsilon_{igp}\right]^2 &= \sum_p \left[(1-\overline{W})SET_p - \overline{W}SEU_p\right]^2 \\ &= \sum_p \left[\frac{C}{n}SET_p - \frac{T}{n}SEU_p\right]^2. \end{aligned} \tag{27}$$

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

$$\begin{aligned}\left[\sum_p \sum_g \sum_i (W_{gp} - \overline{W})^2\right]^2 &= \left[\sum_p \sum_g (W_{gp} - \overline{W})^2 n_{gp}\right]^2 \\ &= \left[(1-\overline{W})^2 \sum_p T_p + \overline{W}^2 \sum_p C_p\right]^2 \\ &= \left[\frac{C^2}{n^2}T + \frac{T^2}{n^2}C\right]^2 \\ &= \left[\frac{CT}{n}\right]^2. \end{aligned} \tag{28}$$

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.

[19] The clustered variance estimators of $\widehat{\tau}$ in the demeaned regression in Equation (25) and in the regression with an intercept in Equation (1) are equal (Cameron and Miller, 2015).

Then, combining Equations (26), (27) and (28),

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

QED.

Point 2

Apply Equation (24) to the residuals and covariates of the regression defined by Equation (25). Then,

$$\widehat{\mathbb{V}}_{unit}(\widehat{\tau}) = \frac{\sum_p \sum_g \left[(W_{gp} - \overline{W}) \sum_i \epsilon_{igp}\right]^2}{\left[\sum_p \sum_g \sum_i (W_{gp} - \overline{W})^2\right]^2}. \tag{29}$$

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

$$\begin{aligned}\sum_p \sum_g \left[(W_{gp} - \overline{W}) \sum_i \epsilon_{igp}\right]^2 &= \sum_p \sum_g (W_{gp} - \overline{W})^2 \left(\sum_i \epsilon_{igp}\right)^2 \\ &= \sum_p \left[(1 - \overline{W})^2 SET_p^2 + \overline{W}^2 SEU_p^2\right] \\ &= \sum_p \left[\frac{C^2}{n^2} SET_p^2 + \frac{T^2}{n^2} SEU_p^2\right]. \end{aligned} \tag{30}$$

The second equality follows from the definition of SET_p and SEU_p. The third equality follows from the definition of T and C. Then, combining Equations (28), (29) and (30),

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

QED.

Point 3

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. Averaging Equation (2) across units in pair p,

$$\overline{Y}_p = \widehat{\tau}_{fe}\overline{W}_p + \widehat{\gamma}_p + \overline{u}_p, \tag{31}$$

where $\overline{Y}_p = \frac{1}{n_p}\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} Y_{igp}$, $\overline{W}_p = \frac{1}{n_p}\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} W_{gp} = \frac{1}{n_p}\sum_{g=1}^{2} W_{gp}n_{gp} = \frac{T_p}{n_p}$, and $\overline{u}_p = \frac{1}{n_p}\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} u_{igp}$. Substracting Equation (31) from Equation (2),

$$Y_{igp} - \overline{Y}_p = \widehat{\tau}_{fe}(W_{gp} - \overline{W}_p) + u_{igp} - \overline{u}_p. \tag{32}$$

$\{u_{ijp'}\}$ is orthogonal to the pair-p fixed effect indicator $\{\delta_{igp}\}$, so

$$\begin{aligned}\sum_{p'=1}^{P}\sum_{g=1}^{2}\sum_{i=1}^{n_{jp'}} u_{ijp'}\delta_{igp} = 0 \\ \Leftrightarrow \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} u_{igp} = 0,\end{aligned} \tag{33}$$

where the equivalence holds because $\delta_{igp} = 1$ if and only if observation i belongs to pair p. This implies that for all p $\overline{u}_p = 0$. Equation (32) then becomes a regression with one covariate and the same residuals as in Equation (2):

$$Y_{igp} - \overline{Y}_p = \widehat{\tau}_{fe}(W_{gp} - \overline{W}_p) + u_{igp}. \tag{34}$$

Now, it follows from Equations (23) and (34) that[20]

$$\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \frac{\left[\sum_{p=1}^{P}\left(\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} u_{igp}(W_{gp} - \overline{W}_p)\right)^2\right]}{\left(\sum_{p=1}^{P}\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}(W_{gp} - \overline{W}_p)^2\right)^2}. \tag{35}$$

[20]The clustered variance estimators of $\widehat{\tau}_{fe}$ in the regression residualized from the pair fixed effects in Equation (34) and in the regression with pair fixed effects in Equation (2) are equal (Cameron and Miller, 2015).

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

$$\begin{aligned}\left[\sum_p \sum_g \sum_i (W_{gp} - \overline{W}_p)^2\right]^2 &= \left[\sum_p \sum_g (W_{gp} - \overline{W}_p)^2 n_{gp}\right]^2 \\ &= \left[\sum_p [T_p(1-\overline{W}_p)^2 + C_p \overline{W}_p^2]\right]^2 \\ &= \left[\sum_p \left(T_p \frac{C_p^2}{n_p^2} + C_p \frac{T_p^2}{n_p^2}\right)\right]^2 \\ &= \left[\sum_p \frac{T_p C_p}{n_p}\right]^2 \\ &= \left[\sum_p (n_{1p}^{-1} + n_{2p}^{-1})^{-1}\right]^2 . \end{aligned} \tag{36}$$

The numerator of $\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe})$ is equal to

$$\begin{aligned}\sum_{p=1}^{P} \left(\sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} u_{igp}(W_{gp} - \overline{W}_p)\right)^2 &= \sum_{p=1}^{P} \left(\sum_{g=1}^{2} (W_{gp} - \overline{W}_p) \sum_{i=1}^{n_{gp}} u_{igp}\right)^2 \\ &= \sum_{p=1}^{P} \left(-\overline{W}_p (SET_{p,fe} + SEU_{p,fe}) + SET_{p,fe}\right)^2 \\ &= \sum_{p=1}^{P} (SET_{p,fe})^2 , \end{aligned} \tag{37}$$

where $SET_{p,fe} + SEU_{p,fe} = \sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} u_{igp} = 0$ from Equation (33). Finally,

$$\begin{aligned}
SET_{p,fe} &= \sum_{g,i} W_{gp}[Y_{igp} - \widehat{\gamma}_p - \widehat{\tau}_{fe} W_{gp}] \\
&= \sum_{g,i} W_{gp} Y_{igp} - (\widehat{\gamma}_p + \widehat{\tau}_{fe}) \sum_{g,i} W_{gp} \\
&= \sum_{g,i} W_{gp} Y_{igp} - (\overline{Y}_p - \widehat{\tau}_{fe} \overline{W}_p + \widehat{\tau}_{fe}) \overline{W}_p n_p \\
&= \sum_{g,i} W_{gp} Y_{igp} - \overline{W}_p \sum_{g,i} Y_{igp} - (1 - \overline{W}_p) \widehat{\tau}_{fe} \overline{W}_p n_p \\
&= \sum_{g,i} W_{gp} Y_{igp} - \overline{W}_p \left[\sum_{g,i} W_{gp} Y_{igp} + \sum_{g,i} (1 - W_{gp}) Y_{igp} \right] - (1 - \overline{W}_p) \overline{W}_p n_p \widehat{\tau}_{fe} \\
&= (1 - \overline{W}_p) \sum_{g,i} W_{gp} Y_{igp} - \overline{W}_p \sum_{g,i} (1 - W_{gp}) Y_{igp} - (1 - \overline{W}_p) \overline{W}_p n_p \widehat{\tau}_{fe} \\
&= (1 - \overline{W}_p) \overline{W}_p n_p \left(\frac{\sum_{g,i} W_{gp} Y_{igp}}{\overline{W}_p n_p} - \frac{\sum_{g,i} (1 - W_{gp}) Y_{igp}}{(1 - \overline{W}_p) n_p} - \widehat{\tau}_{fe} \right) \\
&= \frac{n_{1p} n_{2p}}{n_p^2} n_p \left(\frac{\sum_{g,i} W_{gp} Y_{igp}}{\sum_{g,i} W_{gp}} - \frac{\sum_{g,i} (1 - W_{gp}) Y_{igp}}{\sum_{g,i} (1 - W_{gp})} - \widehat{\tau}_{fe} \right) \\
&= \frac{n_{1p} n_{2p}}{n_{1p} + n_{2p}} (\widehat{\tau}_p - \widehat{\tau}_{fe}).
\end{aligned} \tag{38}$$

The first equality follows from the definition of $SET_{p,fe}$. The second equality follows from the definition of u_{igp} in Equation (2). The third equality follows from the definition of $\overline{W}_p$ and Equations (31) and (33). The ninth equality follows from the definition of $\widehat{\tau}_p$.

Therefore, combining Equations (35), (36), (37) and (38),

$$\widehat{\mathbb{V}}_{pair}(\widehat{\tau}_{fe}) = \sum_{p=1}^{P} \omega_p^2 (\widehat{\tau}_p - \widehat{\tau})^2$$

QED.

Point 4

Applying the definition of the UCVE from Equation (24) to the regression in Equation (34),

$$\widehat{\mathbb{V}}_{unit}(\widehat{\tau}_{fe}) = \frac{\left[\sum_{p=1}^{P} \sum_{g=1}^{2} \left(\sum_{i=1}^{n_{gp}} u_{igp} (W_{gp} - \overline{W}_p) \right)^2 \right]}{\left(\sum_{p=1}^{P} \sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} (W_{gp} - \overline{W}_p)^2 \right)^2}. \tag{39}$$

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

$$\begin{aligned}
\sum_{p=1}^{P}\sum_{g=1}^{2}\left(\sum_{i=1}^{n_{gp}} u_{igp}(W_{gp}-\overline{W}_p)\right)^2 &= \sum_{p=1}^{P}\sum_{g=1}^{2}(W_{gp}-\overline{W}_p)^2\left(\sum_{i=1}^{n_{gp}} u_{igp}\right)^2 \\
&= \sum_{p=1}^{P}\left((1-\overline{W}_p)^2 SET_{p,fe}^2 + \overline{W}_p^2 SEU_{p,fe}^2\right) \\
&= \sum_{p=1}^{P} SET_{p,fe}^2\left(\frac{C_p^2}{n_p^2}+\frac{T_p^2}{n_p^2}\right) \\
&= \sum_{p=1}^{P}\frac{C_p^2T_p^2}{n_p^2} SET_{p,fe}^2\left(\frac{1}{T_p^2}+\frac{1}{C_p^2}\right) \\
&= \sum_{p=1}^{P}(n_{1p}^{-1}+n_{2p}^{-1})^{-2} SET_{p,fe}^2\left(\frac{1}{n_{1p}^2}+\frac{1}{n_{2p}^2}\right). \qquad (40)
\end{aligned}$$

The second equality follows from the definitions of $SET_{p,fe}$ and $SEU_{p,fe}$. The third equality follows from Equation (33), 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. Finally, combining Equations (36), (38), (39) and (40),

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

QED.

H.3 Proof of Lemma D.1

Point 1

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

Taking expected value,

$$\begin{aligned}
\mathbb{E}[\widehat{\mathbb{V}}_{pop}(\widehat{\tau})] &= \frac{1}{P^2}\sum_{r=1}^{R}\mathbb{E}(\widehat{\tau}_{1r}^2 + \widehat{\tau}_{2r}^2 - 2\widehat{\tau}_{1r}\widehat{\tau}_{2r}), \\
&= \frac{1}{P^2}\sum_{r=1}^{R}(\mathbb{V}(\widehat{\tau}_{1r}) + \mathbb{V}(\widehat{\tau}_{2r}) + \tau_{1r}^2 + \tau_{2r}^2 - 2\tau_{1r}\tau_{2r}), \\
&= \frac{1}{P^2}\sum_{p=1}^{P}\mathbb{V}(\widehat{\tau}_p) + \frac{1}{P^2}\sum_{r=1}^{R}(\tau_{1r} - \tau_{2r})^2, \\
&= \mathbb{V}(\widehat{\tau}) + \frac{1}{P^2}\sum_{r=1}^{R}(\tau_{1r} - \tau_{2r})^2. \qquad (41)
\end{aligned}$$

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). **QED**.

Point 2

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

QED.

Point 3

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

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

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

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 2 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 2 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.

H.4 Proof of Theorem D.2

Point 1

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

The second equality follows from the properties of the variance. As $P \to +\infty$, by Lemma H.1, $\sum_{p=1}^{P}\frac{\widehat{\tau}_p^2 - \mathbb{E}[\widehat{\tau}_p^2]}{P} \xrightarrow{\mathbb{P}} 0$. 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 (55) 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 (18) and (42) and a reasoning similar to that used to prove Equation (21).

QED.

Point 2

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

The first equality follows from Point 2 of Lemma 3.1. The convergence arrow follows from Equations (20) and (42). The result follows from the previous display, Equation (18), and a reasoning similar to that used to prove Equation (21).

QED.

Point 3

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

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

Point 4 is straightforward so we do not prove it.

QED.

H.5 Proof of Lemma G.1

Let e_{igp} be the residual from the weighted least squares regression. One has

$$Y_{igp} = \widetilde{\alpha} + \widetilde{\tau}W_{gp} + e_{igp}.$$

Let $\widetilde{Y} = \frac{1}{n}\sum_{i,g,p} V_{gp} Y_{igp}$. The previous display implies that

$$\begin{aligned}\widetilde{Y} &= \widetilde{\alpha}\sum_{i,g,p}\frac{V_{gp}}{n} + \widetilde{\tau}\frac{1}{n}\sum_{i,g,p} V_{gp}W_{gp} + \frac{1}{n}\sum_{i,g,p} V_{gp}e_{igp} \\ &= 2\widetilde{\alpha} + \widetilde{\tau},\end{aligned}$$

where the second equality follows from $\frac{1}{n}\sum_{i,g,p} V_{gp}e_{igp} = 0$, by the first-order condition attached to $\widetilde{\alpha}$ in the weighted OLS minimization problem. Then, combining the two preceding displays implies that

$$Y_{igp} - \frac{1}{2}\widetilde{Y} = \widetilde{\tau}\left(W_{gp} - \frac{1}{2}\right) + e_{igp}. \tag{43}$$

The next step is to compute the clustered variance estimators for the weighted least squares estimator. To do so, we apply Equation (15) in Cameron and Miller (2015) to the residuals and covariates of the regression defined by Equation (43). This equation implies that

$$\widehat{\mathbb{V}}_{pair}(\widetilde{\tau}) = \frac{\sum_p \left[\sum_g V_{gp}(W_{gp} - \frac{1}{2})\sum_i e_{igp}\right]^2}{\left[\sum_p \sum_g \sum_i V_{gp}(W_{gp} - \frac{1}{2})^2\right]^2}. \tag{44}$$

Let $\widehat{Y}_{igp} = \widetilde{\alpha} + W_{gp}\widetilde{\tau}$, $\widehat{Y}(0) = \widetilde{\alpha}$, and $\widehat{Y}(1) = \widetilde{\alpha} + \widetilde{\tau}$. Note that

$$\begin{aligned}\sum_{i,g} W_{gp}\frac{e_{igp}}{n_{gp}} &= \sum_{i,g} W_{gp}(Y_{igp} - \widehat{Y}_{igp})/n_{gp} \\ &= \sum_g W_{gp}\overline{y}_{gp}(1) - \widehat{Y}(1)\sum_g W_{gp} \\ &= \widehat{Y}_p(1) - \sum_{p'}\frac{n_{p'}}{n}\widehat{Y}_{p'}(1)\end{aligned} \tag{45}$$

The second equality follows from $W_{gp}Y_{igp} = W_{gp}Y_{igp}(1)$, the definition of $\overline{y}_{gp}(1)$ and $W_{gp}\widehat{Y}_{igp} = W_{gp}\widehat{Y}(1)$. The third equality follows from the definition of $\widehat{Y}_p(1)$, Point 2 of Assumption 1, and the definition of $\widehat{Y}(1)$.

Likewise,

$$\sum_{i,g}(1 - W_{gp})\frac{e_{igp}}{n_{gp}} = \widehat{Y}_p(0) - \sum_{p'}\frac{n_{p'}}{n}\widehat{Y}_{p'}(0) \tag{46}$$

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

$$\begin{aligned}
\sum_p \left[\sum_g V_{gp}\left(W_{gp}-\frac{1}{2}\right)\sum_i e_{igp}\right]^2 &= \sum_p \left[\sum_g n_p\left(W_{gp}-\frac{1}{2}\right)(W_{gp}+1-W_{gp})\sum_i \frac{e_{igp}}{n_{gp}}\right]^2 \\
&= \sum_p n_p^2\left[\left(1-\frac{1}{2}\right)\sum_{i,g} W_{gp}\frac{e_{igp}}{n_{gp}} - \frac{1}{2}\sum_{i,g}(1-W_{gp})\frac{e_{igp}}{n_{gp}}\right]^2 \\
&= \sum_p \frac{n_p^2}{4}\left[\widehat{Y}_p(1) - \sum_{p'}\frac{n_{p'}}{n}\widehat{Y}_{p'}(1) - \widehat{Y}_p(0) + \sum_{p'}\frac{n_{p'}}{n}\widehat{Y}_{p'}(0)\right]^2 \\
&= \sum_p \frac{n_p^2}{4}\left[\widehat{\tau}_p - \widetilde{\tau}\right]^2. \qquad (47)
\end{aligned}$$

The second equality follows from the fact that $W_{gp}-\frac{1}{2} = 1-\frac{1}{2}$ for the treated units and $W_{gp}-\frac{1}{2} = -\frac{1}{2}$ for the untreated units. The third equality follows from Equations (45) and (46).

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

$$\begin{aligned}
\left[\sum_p\sum_g\sum_i V_{gp}\left(W_{gp}-\frac{1}{2}\right)^2\right]^2 &= \left[2n\frac{1}{4}\right]^2 \\
&= \frac{n^2}{4}. \qquad (48)
\end{aligned}$$

Then, combining Equations (44), (47) and (48),

$$\widehat{\mathbb{V}}_{pair}(\widetilde{\tau}) = \sum_p \frac{n_p^2}{n^2}\left[\widehat{\tau}_p - \widetilde{\tau}\right]^2 = \frac{1}{P^2}\sum_p \frac{n_p^2}{\overline{n}^2}\left[\widehat{\tau}_p - \widetilde{\tau}\right]^2. \qquad (49)$$

QED.

H.6 Proof of Theorem G.2

It follows from Lemma H.1 that

$$\widetilde{\tau} - \tau^* = \frac{1}{P}\sum_p \frac{n_p}{\overline{n}}(\widehat{\tau}_p - \mathbb{E}[\widehat{\tau}_p]) \xrightarrow{\mathbb{P}} 0, \qquad (50)$$

and

$$\frac{1}{P}\sum_p \left(\frac{n_p}{\overline{n}}\right)^2\left[\widehat{\tau}_p^2 - \mathbb{E}(\widehat{\tau}_p^2)\right] \xrightarrow{\mathbb{P}} 0. \qquad (51)$$

By a similar argument to the one used in the proof of Lemma H.1, one can also show that

$$\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2\left[\widehat{\tau}_p-\mathbb{E}(\widehat{\tau}_p)\right]\xrightarrow{\mathbb{P}}0. \tag{52}$$

We now use Point 3 of Assumption 5 to derive the asymptotic distribution of $(\widetilde{\tau}-\tau^*)/(\widetilde{S}_P/P)$. As $\sum_{p=1}^{P}\mathbb{E}\left[|\frac{n_p}{\overline{n}}|^{2+\epsilon}|\widehat{\tau}_p-\mathbb{E}\left[\widehat{\tau}_p\right]|^{2+\epsilon}/\widetilde{S}_P^{2+\epsilon}\right]\to 0$ for some $\epsilon>0$ (by Point 3 of Assumption 5), then, by the Lyapunov central limit theorem, $(\widetilde{\tau}-\tau^*)/(\widetilde{S}_P/P)=\sum_p\frac{n_p}{\overline{n}}(\widehat{\tau}_p-\mathbb{E}[\widehat{\tau}_p])/\widetilde{S}_P\xrightarrow{d}\mathcal{N}(0,1)$ as $P\to+\infty$, as $\widetilde{S}_P^2=P^2\mathbb{V}(\widetilde{\tau})=\sum_{p=1}^{P}\mathbb{V}\left(\frac{n_p}{\overline{n}}\widehat{\tau}_p\right)$.
Therefore,

$$(\widetilde{\tau}-\tau^*)/\sqrt{\mathbb{V}(\widetilde{\tau})}\xrightarrow{d}\mathcal{N}(0,1). \tag{53}$$

Then,

$$\begin{aligned}
&P\widehat{\mathbb{V}}_{pair}(\widetilde{\tau})-P\mathbb{V}(\widetilde{\tau})\\
&=\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2(\widehat{\tau}_p-\widetilde{\tau})^2-\frac{1}{P}\sum_{p=1}^{P}\left(\frac{n_p}{\overline{n}}\right)^2\mathbb{V}(\widehat{\tau}_p)\\
&=\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2(\widehat{\tau}_p-\widetilde{\tau})^2-\frac{1}{P}\sum_{p=1}^{P}\left(\frac{n_p}{\overline{n}}\right)^2[\mathbb{E}(\widehat{\tau}_p^2)-\mathbb{E}[\widehat{\tau}_p]^2]\\
&=\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2(\widehat{\tau}_p^2-2\widetilde{\tau}\widehat{\tau}_p+\widetilde{\tau}^2)-\frac{1}{P}\sum_{p=1}^{P}\left(\frac{n_p}{\overline{n}}\right)^2[\mathbb{E}(\widehat{\tau}_p^2)-\mathbb{E}[\widehat{\tau}_p]^2]\\
&=\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2(\widehat{\tau}_p^2-\mathbb{E}[\widehat{\tau}_p^2])-2\widetilde{\tau}\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2\widehat{\tau}_p+\widetilde{\tau}^2\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2+\frac{1}{P}\sum_{p=1}^{P}\left(\frac{n_p}{\overline{n}}\right)^2\mathbb{E}[\widehat{\tau}_p]^2\\
&\xrightarrow{\mathbb{P}}-2\tau^{\infty}\lim_{P\to+\infty}\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2\mathbb{E}[\widehat{\tau}_p]+(\tau^{\infty})^2\lim_{P\to+\infty}\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2+\lim_{P\to+\infty}\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2\mathbb{E}[\widehat{\tau}_p]^2\\
&=\lim_{P\to+\infty}\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2\left[\mathbb{E}[\widehat{\tau}_p]^2-2\tau^{\infty}\mathbb{E}[\widehat{\tau}_p]+(\tau^{\infty})^2\right]\\
&=\lim_{P\to+\infty}\frac{1}{P}\sum_{p=1}^{P}\left(\frac{n_p}{\overline{n}}\right)^2[\mathbb{E}[\widehat{\tau}_p]-\tau^{\infty}]^2.
\end{aligned} \tag{54}$$

The first equality follows from Equation (49) and the fact that the $(\widehat{\tau}_p)_{p=1}^{P}$ are independent across p by Point 3 of Assumption 1. The second equality follows from the definition of variance. The convergence in probability follows from Equations (50) and (51), (52), and Point 2 of Assumption 5.

Then,

$$\frac{\widetilde{\tau} - \tau^*}{\sqrt{\widehat{\mathbb{V}}_{pair}(\widetilde{\tau})}} = \frac{\widetilde{\tau} - \mathbb{E}[\widetilde{\tau}]}{\sqrt{\mathbb{V}(\widetilde{\tau})}} \sqrt{\frac{P\mathbb{V}(\widetilde{\tau})}{P\widehat{\mathbb{V}}_{pair}(\widetilde{\tau})}}$$
$$\xrightarrow{d} \mathcal{N}(0, \sigma^2_{wls}).$$

The convergence in distribution follows from Equation (54), Equation (53), Lemma H.2, the Slutsky Lemma, and the CMT.

QED.

H.7 Proof of Lemma G.3

$$\begin{aligned}
\mathbb{E}\left[\sum_p \widetilde{\omega}_p^2(\widehat{\tau}_p - \widehat{\tau}_{fe})^2\right] &= \sum_p \widetilde{\omega}_p^2 \,\mathbb{E}[(\widehat{\tau}_p - \widehat{\tau}_{fe})^2] \\
&= \sum_p \widetilde{\omega}_p^2 \left[\mathbb{V}(\widehat{\tau}_p - \widehat{\tau}_{fe}) + [\mathbb{E}(\widehat{\tau}_p - \widehat{\tau}_{fe})]^2\right] \\
&= \sum_p \widetilde{\omega}_p^2 \left[\mathbb{V}(\widehat{\tau}_p) + \mathbb{V}(\widehat{\tau}_{fe}) - 2\mathrm{Cov}(\widehat{\tau}_p, \widehat{\tau}_{fe}) + [\mathbb{E}(\widehat{\tau}_p - \widehat{\tau}_{fe})]^2\right] \\
&= \sum_p \widetilde{\omega}_p^2 \left[\mathbb{V}(\widehat{\tau}_p) + \mathbb{V}(\widehat{\tau}_{fe}) - 2\omega_p \mathbb{V}(\widehat{\tau}_p) + [\mathbb{E}(\widehat{\tau}_p - \widehat{\tau}_{fe})]^2\right] \\
&= \sum_p \widetilde{\omega}_p^2 \left[1 - 2\omega_p\right] \mathbb{V}(\widehat{\tau}_p) + \mathbb{V}(\widehat{\tau}_{fe}) \sum_p \widetilde{\omega}_p^2 + \sum_p \widetilde{\omega}_p^2 [\mathbb{E}(\widehat{\tau}_p - \widehat{\tau}_{fe})]^2 \\
&= \sum_p \omega_p^2 \mathbb{V}(\widehat{\tau}_p) + \mathbb{V}(\widehat{\tau}_{fe}) \sum_p \widetilde{\omega}_p^2 + \sum_p \widetilde{\omega}_p^2 [\mathbb{E}(\widehat{\tau}_p - \widehat{\tau}_{fe})]^2 \\
&= \mathbb{V}(\widehat{\tau}_{fe}) \left(1 + \textstyle\sum_p \widetilde{\omega}_p^2\right) + \displaystyle\sum_p \widetilde{\omega}_p^2 [\mathbb{E}(\widehat{\tau}_p - \widehat{\tau}_{fe})]^2
\end{aligned}$$

The first equality follows from the linearity of the expectation and the fact that the weights ω_p are not stochastic. The fourth equality follows from Point 3 of Assumption 1. The sixth equality follows from the definition of $\widetilde{\omega}_p$. The seventh equality follows from the definition of the variance, the definition of $\widehat{\tau}_{fe}$ and Point 3 of Assumption 1.

QED.

H.8 Auxiliary Lemmas to prove Theorems B.1, D.2, and G.2

Lemma H.1. *Let $q \geq 1$, under Points 2 and 3 of Assumption 1, and Assumption 2 or Point 1 of Assumption 3,*

$$\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^q \left[\widehat{\tau}_p^q - \mathbb{E}(\widehat{\tau}_p^q)\right] \xrightarrow{\mathbb{P}} 0$$

Proof. Assumption 2 implies Point 1 of Assumption 3, so it is sufficient to show that the result holds under Points 2 and 3 of Assumption 1, and Point 1 of Assumption 3.

Note that by Point 3 of Assumption 1, $((\frac{n_p}{\overline{n}}\widehat{\tau}_p)^q - \mathbb{E}[(\frac{n_p}{\overline{n}}\widehat{\tau}_p)^q])_{p=1}^P$, $q \geq 1$, is a sequence of independent random variables with mean zero.

Note that, for all p,

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{n_p}{\overline{n}}\widehat{\tau}_p\right|^{q+\epsilon}\right]^{1/(q+\epsilon)} &= \frac{n_p}{\overline{n}}\,\mathbb{E}\left[\left|\widehat{Y}_p(1) - \widehat{Y}_p(0)\right|^{q+\epsilon}\right]^{1/(q+\epsilon)} \\
&\leq N\left(\left(\mathbb{E}\left[\left|\widehat{Y}_p(1)\right|^{q+\epsilon}\right]\right)^{1/(q+\epsilon)} + \left(\mathbb{E}\left[\left|\widehat{Y}_p(0)\right|^{q+\epsilon}\right]\right)^{1/(q+\epsilon)}\right) \\
&= N\left(\left(\mathbb{E}\left[\left|\sum_g W_{gp}\overline{y}_{gp}(1)\right|^{q+\epsilon}\right]\right)^{1/(q+\epsilon)} + \left(\mathbb{E}\left[\left|\sum_g (1-W_{gp})\overline{y}_{gp}(0)\right|^{q+\epsilon}\right]\right)^{1/(q+\epsilon)}\right) \\
&\leq N\left(\sum_g\left(\mathbb{E}\left[\left|W_{gp}\overline{y}_{gp}(1)\right|^{q+\epsilon}\right]\right)^{1/(q+\epsilon)} + \sum_g\left(\mathbb{E}\left[\left|(1-W_{gp})\overline{y}_{gp}(0)\right|^{q+\epsilon}\right]\right)^{1/(q+\epsilon)}\right) \\
&= N\left(\sum_g\left(\mathbb{E}[W_{gp}]\left|\overline{y}_{gp}(1)\right|^{q+\epsilon}\right)^{1/(q+\epsilon)} + \sum_g\left(\mathbb{E}[1-W_{gp}]\left|\overline{y}_{gp}(0)\right|^{q+\epsilon}\right)^{1/(q+\epsilon)}\right) \\
&= N\left(\sum_g\left(\frac{1}{2}\left|\overline{y}_{gp}(1)\right|^{q+\epsilon}\right)^{1/(q+\epsilon)} + \sum_g\left(\frac{1}{2}\left|\overline{y}_{gp}(0)\right|^{q+\epsilon}\right)^{1/(q+\epsilon)}\right) \\
&< N\frac{4}{2^{1/(q+\epsilon)}}M < +\infty. \qquad (55)
\end{aligned}$$

The first equality follows from the definition of $\widehat{\tau}_p$. The first inequality follows from Minkowski's inequality, and from Point 1 of Assumption 5. 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 the LLN in Lemma 1 in Liu (1988), the previous facts and the fact that almost sure

convergence implies convergence in probability, one concludes that

$$\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^q\left[\widehat{\tau}_p^q-\mathbb{E}(\widehat{\tau}_p^q)\right]\xrightarrow{\mathbb{P}}0. \tag{56}$$

QED.

Lemma H.2. *[Strictly positive limit for $P\mathbb{V}(\widetilde{\tau})$] Under Point 2 of Assumption 3 and Point 1 of Assumption 5,* $\lim_{P\to+\infty}P\mathbb{V}(\widetilde{\tau})>0$.

Proof. Note that

$$\begin{aligned}\lim_{P\to+\infty}P\mathbb{V}(\widetilde{\tau}) &= \lim_{P\to+\infty}\frac{1}{P}\sum_{p}\left(\frac{n_p}{\overline{n}}\right)^2\mathbb{V}(\widehat{\tau}_p)\\ &\geq \frac{1}{N^2}\lim_{P\to+\infty}\frac{1}{P}\sum_{p}\mathbb{V}(\widehat{\tau}_p)\\ &= \frac{1}{N^2}\lim_{P\to+\infty}P\mathbb{V}(\widehat{\tau})\\ &> 0.\end{aligned}$$

The first equality follows from the definition of $\widetilde{\tau}$ and Point 3 of Assumption 1. The first inequality follows from the fact that $0<\frac{1}{N}\leq\frac{n_p}{\overline{n}}\leq N$ (which follows from Point 1 of Assumption 5). The second equality follows from the definition of $\mathbb{V}(\widehat{\tau})$. The second inequality follows from Point 2 of Assumption 3.

QED.