American Economic Review 2022, 112(12): 3911–3940
https://doi.org/10.1257/aer.20201856

Optimality of Matched-Pair Designs in Randomized Controlled Trials[†]

By Yuehao Bai*

In randomized controlled trials, treatment is often assigned by stratified randomization. I show that among all stratified randomization schemes that treat all units with probability one half, a certain matched-pair design achieves the maximum statistical precision for estimating the average treatment effect. In an important special case, the optimal design pairs units according to the baseline outcome. In a simulation study based on datasets from ten randomized controlled trials, this design lowers the standard error for the estimator of the average treatment effect by 10 percent on average, and by up to 34 percent, relative to the original designs. (*JEL* C13, C21)

This paper studies the optimality of matched-pair designs in randomized controlled trials (RCTs). Matched-pair designs are examples of stratified randomization, in which the researcher partitions a set of units into strata (groups) based on their observed covariates and assigns a fraction of units in each stratum to treatment. A matched-pair design is a stratified randomization scheme with two units in each stratum.

Stratified randomization is prevalent in economics. Among the 5,000 RCTs in the AEA RCT Registry, more than 800 are stratified. The schemes in these papers, however, differ vastly in terms of the covariates used to stratify and how fine the strata are. Among these 800 RCTs, around 50 use matched-pair designs. Moreover, 56 percent of the researchers interviewed in Bruhn and McKenzie (2009) have used matched-pair designs at some point in their research. Yet, despite the frequency with which applied researchers make decisions about how to stratify, there are few general econometric results on whether matched-pair designs lead to better precision of estimators of treatment effects than other stratified randomization schemes and the best way to pair units.

*University of Michigan (email: yuehaob@umich.edu). Isaiah Andrews was the coeditor for this article. I am deeply grateful for the encouragement and guidance from my advisors Azeem Shaikh, Stephane Bonhomme, Alex Torgovitsky, and Leonardo Bursztyn. I thank the coeditor and the referees for numerous comments that helped improve the paper greatly. I thank Marinho Bertanha, John Bound, Charlie Brown, Aibo Gong, Florian Gunsilius, Sara Heller, Wooyong Lee, Jonathan Roth, Joshua Shea, Max Tabord-Meehan, Dean Yang, Basit Zafar, and Qinyue Zhou for extensive feedback on the paper. I thank Rex Hsieh and Jiehan Xu for outstanding research assistance. I also thank Bobbie Goettler for excellent copyediting.

[†]Go to https://doi.org/10.1257/aer.20201856 to visit the article page for additional materials and author disclosure statements.

I derive the exact form of the stratified randomization scheme that has the maximum statistical precision for estimating the average treatment effect (ATE). The optimal scheme is a matched-pair design. In an important special case, the optimal design is to order the units according to the baseline values of the primary outcome variable of interest and then pair the adjacent units. When I simulate this simple design using data from ten recent papers in the *American Economic Journal: Applied Economics*, I find it lowers the standard error of the difference-in-means estimator by 10 percent on average, and by up to 34 percent, relative to the designs actually used in those studies. I also find some more complicated stratifications with strata of four units according to multiple covariates could further lower both the mean-squared error (MSE) and the standard error. Based on these findings, I make practical recommendations across a wide range of empirical settings.

In Section II, I study settings where the treated fractions are identically $1/2$ across strata. In such settings, a common estimator for the ATE is the difference in the means of the treated and control groups. The properties of the difference-in-means estimator, however, vary substantially with how the researcher stratifies. To begin, consider the thought experiment where we know the distributions of the potential outcomes. Let $Y(1)$ denote the potential outcome if a unit is treated and let $Y(0)$ denote the potential outcome if it is not treated. Let X denote the observed, baseline covariates. I define an index function $E\left[Y(1) + Y(0) \,|\, X\right]$, the expected sum of the potential outcomes given the covariates. My first result shows the MSE of the difference-in-means estimator is minimized by a matched-pair design, where units are ordered according to this index function and paired adjacently. My optimality result holds at any sample size and without any distributional assumption beyond the existence of moments. In particular, my result does not rely on restrictions on treatment effects heterogeneity.

I describe a special case where the optimal stratification is feasible even without knowing the index function. Suppose X contains a single covariate. Further suppose both $Y(1)$ and $Y(0)$ are higher in expectation when X is higher, so that $E\left[Y(1) + Y(0) \,|\, X\right]$ is increasing in X. In this case, pairing units according to X is optimal. An important example in empirical practice is when X is the baseline value of the primary outcome variable of interest. For instance, in Angrist and Lavy (2009), the primary outcome variable of interest is a test score and the treatment is an educational program, so we expect a higher baseline test score (X) implies a higher endline test score ($Y(1)$ and $Y(0)$) in expectation.

If researchers are unsure about the monotonicity condition, or if multiple covariates are available, then the optimal stratification is generally unknown because the index function is generally unknown. As such, Section III studies several feasible procedures. With multiple covariates, I study pairing units to minimize the (Mahalanobis) distances of the covariates. In settings with auxiliary data, such as data from pilot studies, I propose several matched-pair designs in which the index function is approximated by a proxy based on the auxiliary data.

In Section IV, to compare the performance of these practical procedures, I study the asymptotic properties of the difference-in-means estimator. I show that relative to not stratifying, pairing according to any function of the covariates can only reduce the limiting variance of the difference-in-means estimator. Moreover, the limiting

variance is lower if the stratifying variables explain a larger proportion of the variation in $Y(1) + Y(0)$.

In Section V, I conduct a simulation study using data from a systematically selected set of ten RCTs from recent issues of the *American Economic Journal: Applied Economics*. Relative to the original stratifications used in those ten papers, if the researchers had just paired the units according to their baseline outcomes, then the MSE of the difference-in-means estimator would be 24 percent smaller on average and 56 percent smaller in some cases. The standard error of the difference-in-means estimator would be 10 percent smaller on average and 34 percent smaller in some cases.

Among all methods in the simulation, pairing units to minimize the sum of the squared Mahalanobis distances of the covariates usually leads to the smallest MSEs. When the number of covariates is large, however, the standard error could be even larger than the standard error when pairing according to the baseline outcome alone. Intuitively, this is because the quality of the variance estimator is lower when the curse of dimensionality is more severe. An alternative that balances the MSE and the standard error is to match units into sets of four, instead of pairs, to minimize the sum of the squared Mahalanobis distance of the covariates. Such a method has both smaller MSEs and standard errors than pairing according to the baseline outcome alone while being computationally more intensive.

I conclude with recommendations for empirical practice in Section VI. I recommend different stratifications based on the availability of auxiliary datasets and whether one main outcome of interest clearly dominates the others. All of my recommended procedures are defined by pairing units or matching units into sets of four according to all or a subset of the available covariates.

Related Literature.—This paper is most closely related to Barrios (2013) and Tabord-Meehan (forthcoming). Barrios (2013) studies minimizing the variance of the difference-in-means estimator. He is the first to show pairing units according to my index function is optimal among all matched-pair designs, albeit under the assumption of homogeneous treatment effects. My optimality result holds among all stratified randomization schemes and with heterogeneous treatment effects. Tabord-Meehan (forthcoming) studies optimality within a class of stratification trees. Because the number of strata is fixed in his asymptotic framework, he can optimize over the treated fraction in each stratum. In a matched-pair design, the number of strata is half of the sample size and hence not fixed as the sample size increases, so matched-pair designs are precluded in his framework. In online Appendix C.2, I elaborate on the comparison between the two papers and further note that combining our procedures is straightforward.

The following papers also study matched-pair designs: Greevy et al. (2004) study pairing units to minimize the sum of the squared Mahalanobis distances of the covariates. Imai (2008) studies matched-pair designs, focusing on the sample ATE. The inference methods in this paper build on and extend those in Bai, Romano, and Shaikh (2021). In addition, inference under matched-pair designs has also been studied in Abadie and Imbens (2008), who assume a different sampling framework; Fogarty (2018a, b), who provides conservative estimators for the limiting variance; and de Chaisemartin and Ramirez-Cuellar (2021) in a finite-population setting.

I. Setup and Notation

Let Y_i denote the observed outcome of interest for the ith unit, let D_i denote the treatment status for the ith unit, and let X_i denote the observed, baseline covariates for the ith unit. Further denote by $Y_i(1)$ the potential outcome of the ith unit if treated and by $Y_i(0)$ if not treated. As usual, the observed outcome is related to the potential outcomes and treatment status by the relationship

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i).$$

For ease of exposition, I assume the sample size is even and denote it by $2n$. I assume $\big((Y_i(1), Y_i(0), X_i) : 1 \leq i \leq 2n\big)$ is an i.i.d. sequence of random vectors. Note the potential outcomes and the covariates are drawn from a population and hence are random instead of fixed. For any random vector indexed by i, A_i, define $A^{(n)} = (A_1, \ldots, A_{2n})'$. The main parameter of interest is the ATE:

$$\theta = E\big[Y_i(1) - Y_i(0)\big].$$

In stratified randomization, I first partition the set of units into strata. Formally, I define a stratification $\lambda = \{\lambda_s : 1 \leq s \leq S\}$ as a partition of $\{1, \ldots, 2n\}$:

(i) $\lambda_s \cap \lambda_{s'} = \varnothing$ for all s and s' such that $1 \leq s \neq s' \leq S$;

(ii) $\bigcup_{1 \leq s \leq S} \lambda_s = \{1, \ldots, 2n\}$.

Let Λ_n denote the set of all stratifications of $2n$ units. Define $n_s = |\lambda_s|$ and τ_s as the treated fraction in stratum λ_s. A matched-pair design is simply a stratified randomization scheme with $S = n$ and $n_s = 2$ for $1 \leq s \leq S$. I define $\Lambda_n^{\text{pair}} \subseteq \Lambda_n$ as the set of all matched-pair designs for $2n$ units.

I make the following assumption on the treatment assignment scheme:

ASSUMPTION 1: *Given the covariates $X^{(n)}$, treatment status is determined as follows: independently for $1 \leq s \leq S$, uniformly at random choose $n_s\tau_s$ units in λ_s, and assign $D_i = 1$ to them and $D_i = 0$ to the other units in λ_s. Furthermore, $\tau_s = \frac{1}{2}$ for $1 \leq s \leq S$.*

Assumption 1 implies

$$\big(Y^{(n)}(0), Y^{(n)}(1)\big) \perp D^{(n)} \mid X^{(n)}. \tag{1}$$

In other words, treatment status and potential outcomes are conditionally independent given the covariates. Assumption 1 also implies n_s has to be even because a unit cannot be cut in half. Note the distribution of the vector of treatment status $D^{(n)}$ depends on λ. Most results below can be extended to settings where $\tau_s, 1 \leq s \leq S$ are identical but not $1/2$, or where they are additionally allowed to vary across subpopulations. See Remark 2 for details.

For all treatment assignment schemes in the main text, I estimate the ATE by the difference in the means of the treated and control groups. Formally, for $d \in \{0,1\}$, define

$$\hat{\mu}_n(d) = \frac{1}{n} \sum_{1 \le i \le 2n : D_i = d} Y_i.$$

The difference-in-means estimator is defined as

$$\hat{\theta}_n = \hat{\mu}_n(1) - \hat{\mu}_n(0).$$

The difference-in-means estimator is widely used because it is simple and transparent. Under Assumption 1, it coincides with the ordinary least squares (OLS) estimator for the coefficient in the linear regression of the outcome on treatment status and strata fixed effects and the OLS estimator from the fully saturated version of that regression, both of which are also widely used in analyses of RCTs. See, for example, Duflo, Glennerster, and Kremer (2007); Glennerster and Takavarasha (2013); and Crépon et al. (2015).

II. Optimal Stratification

This section studies the optimal stratification. To preview the results, define the index function

$$g(x) = E\big[Y_i(1) + Y_i(0) \,|\, X_i = x\big]. \tag{2}$$

I show the optimal stratification is given by ordering the units according to $g_i = g(X_i)$ and then pairing the adjacent units. In the special case where X_i is a scalar and $E\big[Y_i(1) \,|\, X_i = x\big]$ and $E\big[Y_i(0) \,|\, X_i = x\big]$ are both weakly increasing (or both weakly decreasing) in x, the optimal stratification is given by ordering the units according to X_i and then pairing the adjacent units.

The analysis in this section is conditional on $X^{(n)}$. In this section only, instead of the population ATE, I focus on the ATE conditional on $X^{(n)}$:

$$\theta_n = \frac{1}{2n} \sum_{1 \le i \le 2n} E\big[Y_i(1) - Y_i(0) | X_i\big].$$

Focusing on θ_n simplifies the discussion. Moreover, conditional on a fixed sample with covariates $X^{(n)}$, I can only hope to be unbiased for θ_n instead of θ. The conclusions of the theorems in this section are the same regardless of whether the parameter of interest is θ_n or θ.

My objective function is the MSE of $\hat{\theta}_n$ for θ_n conditional on $X^{(n)}$ under a stratification $\lambda \in \Lambda_n$:

$$\mathrm{MSE}\big(\lambda \,|\, X^{(n)}\big) = E_\lambda\Big[\big(\hat{\theta}_n - \theta_n\big)^2 \,|\, X^{(n)}\Big].$$

Here, the notation E_λ indicates the distribution of the vector of treatment status $D^{(n)}$ depends on the stratification. I consider minimizing the conditional MSE over the set of all stratifications:

$$\min_{\lambda \in \Lambda_n} \mathrm{MSE}\big(\lambda \,|\, X^{(n)}\big). \tag{3}$$

In what follows, I derive the optimal stratification as the solution to (3). I emphasize that by a simple bias-variance decomposition, one can show (3) is equivalent to the problem where θ_n is replaced by θ, so focusing on θ_n in this section is genuinely without loss of generality.

Solving (3) involves two intermediate results, each carrying additional insights into the problem. To describe the first intermediate result, I define the ex ante bias of $\hat{\theta}_n$ for θ_n conditional on $X^{(n)}$ as

$$\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}\big(\hat{\theta}_n \,|\, X^{(n)}\big) \;=\; E_\lambda\big[\hat{\theta}_n \,|\, X^{(n)}\big] - \theta_n,$$

and the ex post bias of $\hat{\theta}_n$ for θ_n conditional on $X^{(n)}$ and $D^{(n)}$ as

$$\mathrm{Bias}^{\mathrm{post}}_{n}\big(\hat{\theta}_n \,|\, X^{(n)}, D^{(n)}\big) \;=\; E\big[\hat{\theta}_n \,|\, X^{(n)}, D^{(n)}\big] - \theta_n.$$

Here, ex ante bias refers to the bias conditional only on the covariates, before treatment status is realized; ex post bias refers to the bias conditional on both the covariates and treatment status, after treatment status is realized. Note in the definition of the ex post bias, the λ subscript does not appear because $D^{(n)}$ is already given. Note from the definition of the difference-in-means estimator that

$$\hat{\theta}_n \;=\; \frac{1}{n} \sum_{1 \le i \le 2n} \big(Y_i(1)\, D_i - Y_i(0)(1 - D_i)\big). \tag{4}$$

By Assumption 1, the marginal treatment probability of each unit satisfies $E_\lambda\big[D_i | X^{(n)}\big] = \frac{1}{2}$, and together with the conditional independence assumption in (1), they imply

$$E_\lambda\big[\hat{\theta}_n \,|\, X^{(n)}\big] \;=\; \theta_n.$$

Therefore, the ex ante bias is identically zero across $\lambda \in \Lambda_n$, which is not surprising because the ex ante bias should be zero if we run an experiment. By the law of iterated expectations,

$$E_\lambda\Big[\mathrm{Bias}^{\mathrm{post}}_{n}\big(\hat{\theta}_n \,|\, X^{(n)}, D^{(n)}\big) \,|\, X^{(n)}\Big] \;=\; \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}\big(\hat{\theta}_n \,|\, X^{(n)}\big) \;=\; 0,$$

so the mean of the ex post bias over the distribution of treatment status equals the ex ante bias, which is zero.

The first intermediate result is a decomposition of the conditional MSE in (3). Because $E_\lambda\left[\hat{\theta}_n - \theta_n \,|\, X^{(n)}\right] = 0$, by the law of total variance,

$$\begin{aligned}(5)\qquad \mathrm{MSE}\left(\lambda \,|\, X^{(n)}\right) &= \mathrm{var}\left[\hat{\theta}_n - \theta_n \,|\, X^{(n)}\right] \\ &= E_\lambda\left[\mathrm{var}\left[\hat{\theta}_n - \theta_n \,|\, X^{(n)}, D^{(n)}\right] \,|\, X^{(n)}\right] \\ &\quad + \mathrm{var}_\lambda\left[E\left[\hat{\theta}_n \,|\, X^{(n)}, D^{(n)}\right] - \theta_n \,|\, X^{(n)}\right].\end{aligned}$$

For any $\lambda \in \Lambda_n$, the first term on the right-hand side of (5) equals

$$\begin{aligned}&E_\lambda\left[\frac{1}{n^2}\sum_{1\le i\le 2n}\left(\mathrm{var}\left[Y_i(1) \,|\, X_i\right] D_i + \mathrm{var}\left[Y_i(0) \,|\, X_i\right]\left(1 - D_i\right)\right) \,|\, X^{(n)}\right] \\ &\quad = \frac{1}{2n^2}\sum_{1\le i\le 2n}\left(\mathrm{var}\left[Y_i(1) \,|\, X_i\right] + \mathrm{var}\left[Y_i(0) \,|\, X_i\right]\right),\end{aligned}$$

which is identical across all $\lambda \in \Lambda_n$. Note I used the conditional independence assumption in (1), the facts that θ_n is a constant given $X^{(n)}$, that $D_i\left(1 - D_i\right) = 0$ for $1 \le i \le 2n$, and that $E_\lambda\left[D_i \,|\, X^{(n)}\right] = \frac{1}{2}$. Hence, (3) is further equivalent to minimizing the second term on the right-hand side of (5), which is the variance of the ex post bias:

$$\mathrm{var}_\lambda\left[\mathrm{Bias}_n^{\mathrm{post}}\left(\hat{\theta}_n \,|\, X^{(n)}, D^{(n)}\right) \,|\, X^{(n)}\right].$$

The discussion so far leads to my first intermediate result:

LEMMA 1: *Suppose the treatment assignment scheme satisfies Assumption 1. Then,* (3) *is equivalent to*

$$\min_{\lambda\in\Lambda_n} \mathrm{var}_\lambda\left[\mathrm{Bias}_n^{\mathrm{post}}\left(\hat{\theta}_n \,|\, X^{(n)}, D^{(n)}\right) \,|\, X^{(n)}\right].$$

Next, I describe the second intermediate result in solving (3). The result states any stratification is a convex combination of matched-pair designs. Formally, for $\lambda, \lambda' \in \Lambda_n^{\mathrm{pair}}$ and $\delta \in [0,1]$, define $\delta\lambda \oplus \left(1 - \delta\right)\lambda'$ as the randomization between λ and λ' such that λ is implemented with probability δ. Define the convex hull formed by all convex combinations of any finite number of matched-pair designs as

$$\begin{aligned}\mathrm{co}\left(\Lambda_n^{\mathrm{pair}}\right) = \Bigg\{ &\bigoplus_{1\le j\le J} \delta_j \lambda^j : \lambda^j \in \Lambda_n^{\mathrm{pair}}, \\ &\delta_j \ge 0 \text{ for } 1 \le j \le J, \sum_{1\le j\le J} \delta_j = 1, 1 \le J < \infty\Bigg\}.\end{aligned}$$

In other words, a member of the convex hull is the "mixing" of J matched-pair designs, where J is finite.

For example, suppose $2n = 4$. Then, four stratifications are possible:

$$\lambda^0 = \big\{\{1,2,3,4\}\big\},$$

$$\lambda^1 = \big\{\{1,2\},\{3,4\}\big\},$$

$$\lambda^2 = \big\{\{1,3\},\{2,4\}\big\},$$

$$\lambda^3 = \big\{\{1,4\},\{2,3\}\big\}.$$

Stratification λ^0 puts four units in the same stratum, λ^1 pairs 1 and 2 together and 3 and 4 together, and λ^2 and λ^3 are defined similarly. Note that implementing each of the three matched-pair designs with probability $1/3$ is equivalent to implementing λ^0, in the sense that the distributions of (D_1, D_2, D_3, D_4) are the same under the two implementations. Indeed, under λ^1, (D_1, D_2, D_3, D_4) takes the following four values each with probability $1/4$: $(1,0,1,0)$, $(1,0,0,1)$, $(0,1,1,0)$, $(0,1,0,1)$. Similarly, under λ^2, it takes the following four values each with probability $1/4$: $(1,0,0,1)$, $(1,1,0,0)$, $(0,0,1,1)$, $(0,1,1,0)$. Under λ^3, it takes the following four values each with probability $1/4$: $(1,0,1,0)$, $(1,1,0,0)$, $(0,1,0,1)$, $(0,0,1,1)$. Accordingly, under $\frac{1}{3}\lambda^1 \oplus \frac{1}{3}\lambda^2 \oplus \frac{1}{3}\lambda^3$, it takes the following six values each with probability $1/6$: $(1,1,0,0)$, $(1,0,1,0)$, $(1,0,0,1)$, $(0,1,1,0)$, $(0,1,0,1)$, $(0,0,1,1)$. This distribution is the same as that of (D_1, D_2, D_3, D_4) under λ^0, where two out of four units are treated uniformly at random. As a result, $\lambda^0 \in \text{co}\big(\big\{\lambda^1, \lambda^2, \lambda^3\big\}\big)$, meaning λ^0 can be written as a convex combination of the three matched-pair designs.

I show in online Appendix A that the result above holds in general and summarize it into the following lemma:

LEMMA 2: *If the treatment assignment scheme satisfies Assumption 1, then* $\Lambda_n \subseteq \text{co}\big(\Lambda_n^{\text{pair}}\big)$. *In other words, any stratification is a convex combination of matched-pair designs.*

Combining Lemmas 1 and 2 to minimize the MSE as in (3) is now straightforward. To state the result, I need an equivalent notation for matched-pair designs. Recall that a permutation of $\{1, \ldots, 2n\}$ is a function that maps $\{1, \ldots, 2n\}$ onto itself. Let Π_n denote the group of all permutations of $\{1, \ldots, 2n\}$. A matched-pair design is a stratified randomization scheme with

$$\lambda = \big\{\big\{\pi(2s-1), \pi(2s)\big\} : 1 \leq s \leq n\big\},$$

where $\pi \in \Pi_n$. Recall the definition of the index function g in (2) and order the units by defining $\pi^g \in \Pi_n$ that satisfies $g_{\pi^g(1)} \leq \ldots \leq g_{\pi^g(2n)}$. Define the stratification

$$\lambda^g\big(X^{(n)}\big) = \big\{\big\{\pi^g(2s-1), \pi^g(2s)\big\} : 1 \leq s \leq n\big\}. \tag{6}$$

The stratification in (6) is given by ordering the units according to g_i and then pairing the adjacent units. I now show it minimizes the MSE as in (3).

For each $\lambda \in \Lambda_n$, define $V(\lambda)$ as the objective in Lemma 1. Recall $g^{(n)} = (g_1, \ldots, g_n)'$. Then,

$$\begin{aligned} V(\lambda) &= \operatorname{var}_\lambda\Big[E\Big[\hat{\theta}_n \,|\, X^{(n)}, D^{(n)}\Big] - \theta_n \,|\, X^{(n)}\Big] \\ &= \frac{1}{n^2}\operatorname{var}_\lambda\Big[\sum_{1\le i\le 2n}\Big(D_i E\big[Y_i(1)\,|\,X_i\big] - (1-D_i)E\big[Y_i(0)\,|\,X_i\big]\Big)\,|\,X^{(n)}\Big] \\ &= \frac{1}{n^2}\operatorname{var}_\lambda\Big[\sum_{1\le i\le 2n} D_i\Big(E\big[Y_i(0)\,|\,X_i\big] + E\big[Y_i(1)\,|\,X_i\big]\Big)\,|\,X^{(n)}\Big] \\ &= \frac{1}{n^2}\big(g^{(n)}\big)'\operatorname{var}_\lambda\big[D^{(n)}\,|\,X^{(n)}\big]g^{(n)}, \end{aligned}$$

where the first equality follows from the definition of the ex post bias, the second equality follows from (4) and the fact that θ_n is a constant given $X^{(n)}$, and the last two equalities follow by inspection. Recall the variance of D_i is $\frac{1}{4}$. Also recall that for a matched-pair design, the covariance between treatment status of the two units in a pair is $-\frac{1}{4}$, and that of units across pairs is 0. Therefore, for any $\lambda = \big\{\{\pi(1),\pi(2)\}, \ldots, \{\pi(2n-1),\pi(2n)\}\big\} \in \Lambda_n^{\text{pair}}$,

$$V(\lambda) = \frac{1}{4n^2}\sum_{1\le s\le n}\big(g_{\pi(2s-1)} - g_{\pi(2s)}\big)^2.$$

Therefore, $V(\lambda)$ is proportional to the sum of squared distances of g within each pair. By Lemma B.1 in the online Appendix, which is a simple consequence of the Hardy-Littlewood-Pólya rearrangement inequality, $V\big(\lambda^g\big(X^{(n)}\big)\big) \le V(\lambda)$ for any $\lambda \in \Lambda_n^{\text{pair}}$. Therefore, $\lambda^g\big(X^{(n)}\big)$ minimizes the MSE among Λ_n^{pair}, the set of all matched-pair designs.

To conclude $\lambda^g\big(X^{(n)}\big)$ is optimal among the set of all stratifications Λ_n, note each stratification is a mixing of matched-pair designs, and no "mixed strategy" has a better payoff than the optimal "pure strategy." Formally, by Lemma 2, any $\lambda \in \Lambda_n$ can be written as

$$\lambda = \bigoplus_{1\le j\le J}\delta_j\lambda^j,$$

where $\lambda^j \in \Lambda_n^{\text{pair}}$, $\delta_j \ge 0$ for $1 \le j \le J$, and $\sum_{1\le j\le J}\delta_j = 1$. As a result,

$$\begin{aligned} \operatorname{MSE}\big(\lambda\,|\,X^{(n)}\big) &= \sum_{1\le j\le J}\delta_j\operatorname{MSE}\big(\lambda^j\,|\,X^{(n)}\big) \\ &\ge \min_{1\le j\le J}\operatorname{MSE}\big(\lambda^j\,|\,X^{(n)}\big) \ge \operatorname{MSE}\Big(\lambda^g\big(X^{(n)}\big)\,|\,X^{(n)}\Big), \end{aligned}$$

where the equality follows from the definition of the MSE, the first inequality follows because any weighted average of a set of numbers is weakly larger than the minimum across them, and the last inequality follows because $\lambda^g\big(X^{(n)}\big)$ minimizes $\operatorname{MSE}\big(\lambda\,|\,X^{(n)}\big)$ across Λ_n^{pair}. Therefore, I have established my main theorem on the optimal stratification:

THEOREM 1: *Suppose the treatment assignment scheme satisfies Assumption 1. Then, the matched-pair design defined in (6) minimizes the MSE as in (3). In other words, the optimal stratification is given by ordering the units according to g_i and then pairing the adjacent units.*

REMARK 1: *Note the optimal stratification does not depend on knowledge of the conditional variances of $Y_i(1)$ and $Y_i(0)$ given X_i.*

REMARK 2: *Theorem C.1 in the online Appendix examines settings where the treated fractions are identical across strata but not $\frac{1}{2}$. Formally, suppose $\tau_s = \tau = \frac{l}{k}$ for $1 \leq s \leq S$, where $l, k \in \mathbf{N}$, $0 < l < k$, and l and k are mutually prime. Define*

$$g^{\tau}(X_i) = \frac{E\left[Y_i(1) \mid X_i\right]}{\tau} + \frac{E\left[Y_i(0) \mid X_i\right]}{1-\tau}, \tag{7}$$

where g^{τ} adjusts for the treatment probability by inverse probability weighting. The optimal stratification is defined by the following algorithm:

(*i*) *Order the units according to $g^{\tau}(X_i)$.*

(*ii*) *Put the first k units in the first stratum, the second k units in the second stratum, and so on.*

(*iii*) *Uniformly at random assign l of the k units in each stratum to treatment.*

In this case, the optimal design is not paired, but stratified randomization with the appropriate group size remains optimal. For examples in this spirit of small strata, see Bold et al. (2018) and Brown and Andrabi (2020).

REMARK 3: *Rerandomization, studied by Morgan and Rubin (2012, 2015), is an alternative to stratified randomization. Rerandomization takes random draws of treatment status until it falls in an admissible set. The admissible set is usually defined as the collection of treatment assignments under which the distance between the treated and control units is below a threshold. The notion of distance can be, for instance, the (Mahalanobis) distance in the covariates or the distance in g. In matched-pair designs, units are matched to minimize the distance between treated and control units. As such, each possible realization of the vector of treatment status under a matched-pair design not only belongs to the admissible set but also attains the smallest distance within the admissible set. For example, suppose each distinct value of the covariate appears twice in the sample. Then, a matched-pair design is equivalent to rerandomization with the distance threshold set to zero.*

Note from (2) that the index function g_i is a scalar regardless of the dimension of X_i. Moreover, the optimal stratification depends not on the values but merely on the ordering of g_i. For instance, if X_i is univariate and $g(x)$ is monotonic in x, then the optimal stratification in (6) is given by ordering the units by X_i and then pairing the adjacent units. This scenario arises in many settings, especially if X_i is the baseline

value of the primary outcome variable of interest, which is collected in the baseline survey before treatment is assigned. For instance, Angrist and Lavy (2009) study the effect of an educational program on test scores. In their paper, X_i is the baseline test score, so we expect both $E\left[Y_i(1) \mid X_i\right]$ and $E\left[Y_i(0) \mid X_i\right]$ are weakly increasing in X_i. I record this result as a theorem. Let $\pi^X \in \Pi_n$ be such that $X_{\pi^X(1)} \leq \ldots \leq X_{\pi^X(2n)}$.

THEOREM 2: *Suppose X_i is univariate, the treatment assignment scheme satisfies Assumption 1, and $g(x)$ in (2) is monotonic in x. Then,*

$$\lambda^g\left(X^{(n)}\right) = \left\{\left\{\pi^X(2s-1), \pi^X(2s)\right\} : 1 \leq s \leq n\right\}.$$

In other words, the optimal stratification is given by ordering the units according to their covariate values and then pairing the adjacent units.

III. Feasible Procedures

The optimal stratification in Theorem 1 depends on the index function g, which is generally unknown, so the optimal stratification is also generally unknown. Therefore, researchers often need to approximate the index function with some proxies, possibly with the help of auxiliary data. This section studies a wide range of feasible stratification methods. Some procedures are based on data from pilot experiments, which are smaller-scale copies of the main experiment run on the same population. Depending on the availability of a pilot experiment and its sample size, different procedures are available. I switch the parameter of interest back to the population ATE θ, recalling that all results in the previous section hold for both θ_n and θ.

A. *Settings without Pilot Data*

According to Theorem 2, if X_i is univariate and the index function $g(x)$ is monotonic in x, then the optimal stratification is given by pairing units according to X_i. A prominent example is where X_i is the baseline value of the primary outcome variable of interest, and $E\left[Y_i(1) \mid X_i = x\right]$ and $E\left[Y_i(0) \mid X_i = x\right]$ are both weakly increasing or both weakly decreasing in x.

Even if the monotonicity condition fails, units can still be paired according to their baseline outcomes. Theorem 3 and Remark 4 below study the limiting variance of the difference-in-means estimator. They reveal that if we need to choose a single covariate to pair on, the smallest limiting variance is attained by pairing units according to a covariate that explains the largest proportion of the variation in the potential outcomes. Bruhn and McKenzie (2009) note the baseline outcome is often such a covariate. Simulation evidence in Section V further shows pairing units according to the baseline outcome performs better than the status-quo methods in terms of both the MSE and the standard error of the difference-in-means estimator.

Regardless of whether the baseline outcome is available, if X_i is multivariate, then researchers can also pair units to minimize the sum of the squared Mahalanobis distances of the covariates:

$$d(x_1, x_2) = (x_1 - x_2)' \hat{\Sigma}_n^{-1} (x_1 - x_2). \tag{8}$$

Here, $\hat{\Sigma}_n$ is the sample variance matrix of X. Equation (8) is simply the squared Euclidean distance if $\hat{\Sigma}_n$ is the identity matrix, and $\hat{\Sigma}_n^{-1}$ serves as a scale normalization because different covariates may be measured in different units or have different standard deviations. Note

$$d(x_1, x_2) \;=\; \| \hat{\Sigma}_n^{-1/2}(x_1 - x_2) \|^2,$$

where $\hat{\Sigma}_n^{-1/2}$ is the square root of $\hat{\Sigma}_n$. So the Mahalanobis distance between x_1 and x_2 equals the Euclidean distance between $\hat{\Sigma}_n^{-1/2} x_1$ and $\hat{\Sigma}_n^{-1/2} x_2$.

When the baseline outcome is unavailable but a large amount of auxiliary data are available, I can calculate a sample counterpart of (6). In general, the auxiliary data need to come from pilot experiments, but in one special case, even observational data suffice. If the conditional ATEs are homogeneous, meaning

$$E\big[Y_i(1) - Y_i(0) \,|\, X_i\big] \;=\; E\big[Y_i(1) - Y_i(0)\big] \text{ with probability one,} \tag{9}$$

then the ordering of g_i is the same as that of $E\big[Y_i(0) \,|\, X_i\big]$. Suppose we have an observational dataset where the distribution of $\big(Y_i(0), X_i\big)$ is the same as that in the main experiment. As an example, suppose in an RCT to study the effect of educational program on test scores, the researcher has administrative data on the test scores of the previous cohort, and the distributions of $\big(Y_i(0), X_i\big)$ are the same across the two cohorts. Then, they can estimate $E\big[Y_i(0) \,|\, X_i = x\big]$ by a nonparametric regression using the data for the previous cohort and pair the units in the current cohort according to the predicted values in the regression. A key requirement is that the estimator for $E\big[Y_i(0) \,|\, X_i = x\big]$ is consistent in the sense of Assumption 4 and Theorem 5 below. Then, as the sample sizes of the auxiliary data and the main experiment both increase, the limiting variance of $\hat{\theta}_n$ when units are paired according to the predicted values of $E\big[Y_i(0) \,|\, X_i\big]$ is the same as that under the optimal stratification in (6).

B. *Settings with Large Pilots*

Next, I consider settings with data from a pilot experiment. Let m denote the sample size of the pilot experiment. I assume the pilot units are drawn from the same population as the main experiment.

I start by investigating settings where the sample size of the pilot experiment is large. Formally, in the asymptotic framework, I allow both m and n to go to infinity. I pair units according to a suitable estimator $\tilde{g}_m$ of the index function g, where $\tilde{g}_m$ comes from a nonparametric regression using the pilot data. Again, a key requirement is that $\tilde{g}_m$ is consistent for g in the sense of Assumption 4 and Theorem 5 below. Then, as the sample sizes of the auxiliary data and the main experiment both increase, the limiting variance of $\hat{\theta}_n$ when units are paired according to $\tilde{g}_m$ is the same as that under the optimal stratification in (6).

If the pilot data are imperfect in the sense that they does not come from the same population as the main experiment, or if the estimation method for constructing $\tilde{g}_m$ is not flexible enough, then $\tilde{g}_m$ may not converge to g but instead to another function h. In that case, the limiting variance of $\hat{\theta}_n$ is different from that under the optimal

stratification in (6) and depends on h, but it is still smaller than that under no stratification. See Theorem 5 and Remark 5 for details.

C. *Settings with Small Pilots*

In practice, even if pilot data are available, their sample size is often small. In those settings, pairing units according to $\tilde{g}_m$ generally does not ensure efficiency, unlike in settings with large pilots. We may be concerned that $\tilde{g}_m$ is a poor approximation of the index function g, and as a result, if units are paired according to $\tilde{g}_m$, then both the conditional MSE and the limiting variance of $\hat{\theta}_n$ are large.

Researchers could of course ignore the information in the small pilot and implement the procedures in Section IIIA. If they would like to incorporate information from the pilot experiment, they can consider the following procedure. For $d \in \{0,1\}$, let $\tilde{\beta}_m(d)$ denote the OLS estimators of the linear regression coefficients among the treated or untreated units in the pilot experiment and let $\tilde{\Omega}_m(d)$ denote the variance estimators in OLS assuming homoskedasticity (see online Appendix C.5 for details). Further define

$$\tilde{\beta}_m = \tilde{\beta}_m(1) + \tilde{\beta}_m(0)$$

$$\tilde{\Omega}_m = \tilde{\Omega}_m(1) + \tilde{\Omega}_m(0).$$

I pair the units to minimize the sum of the following distances of the covariates:

$$d^{\text{pen}}(x_1, x_2) = \left(x_1'\tilde{\beta}_m - x_2'\tilde{\beta}_m\right)^2 + (x_1 - x_2)'\tilde{\Omega}_m(x_1 - x_2). \tag{10}$$

To shed some light on the behavior of such a minimization problem, I consider two extreme cases. If $\tilde{\Omega}_m = 0$, which means $\tilde{\beta}_m$ is very precise, then the solution is given by pairing units according to $\tilde{g}_m = X_i'\tilde{\beta}_m$. If $\tilde{\Omega}_m$ is large, which means $\tilde{\beta}_m$ is very imprecise, then the second term on the right-hand side of (10) dominates the first term, so the solution is close to a paired matching weighted by $\tilde{\Omega}_m$. Therefore, the solution can be viewed as penalizing the pairing according to $\tilde{g}_m(x) = x'\tilde{\beta}_m$, with the penalization determined by the variance estimator $\tilde{\Omega}_m$. I refer to the solution as the penalized matched-pair design. In online Appendix C.5, I show it is optimal in a Bayesian framework.

D. *Other Practical Considerations*

Each matched-pair design discussed in this section has a counterpart where units are matched into sets of four instead of pairs. Specifically, I first pair the units and then pair the pairs using the midpoints of all pairs, as in Section 4 of Bai, Romano, and Shaikh (2021). Such a design is also discussed by Athey and Imbens (2017). It often increases the MSE relative to its paired version but often improves inference for the ATE, especially with multiple covariates. In particular, simulation evidence in Section V shows that, with multiple covariates, the test with matched sets of

four usually has the correct size but the test with matched pairs often severely underrejects. I refer interested readers to Section IVC for a detailed discussion.

A frequent concern in experiments is attrition, meaning units in the baseline survey may drop out in the follow-up survey, so their covariates are available but outcomes are not. I emphasize that even when a unit attrites, the entire pair may not need to be dropped. If attrition happens, then I redefine the difference-in-means estimator using only nonattritors. If attrition is independent of treatment status conditional on the covariates, then this estimator is consistent for the ATE for nonattritors. I refer interested readers to online Appendix C.3 for details. The case with differential attrition, as in the setting of Lee (2009), is an interesting topic for future work.

Another related question that frequently arises in the design of experiments is that some studies are implemented in multiple waves. Although a full-length discussion of such settings is beyond the scope of the paper, a possible solution is to implement the procedures discussed in this section repeatedly. For instance, in the first wave, researchers could pair the units according to their baseline outcomes. In the second wave, they could use the data from the first wave as pilot data, and implement the pilot-based procedures discussed earlier in this section. They can repeatedly implement the pilot-based procedures in the following waves. In online Appendix C.6, I discuss how to pool the data from multiple waves for estimation and inference.

IV. Asymptotic Results and Inference

The optimality result in Section II pinpoints the optimal stratification but is silent on how the feasible procedures in Section III compare with each other. To make such a comparison, this section studies the asymptotic properties of the difference-in-means estimator. I also provide inference methods for the ATE under different stratifications. The main difficulty in deriving the theoretical results is that under matched-pair designs, treatment status across units is heavily dependent; in fact, treatment status of the two units in a pair is perfectly correlated. I extend the results in Bai, Romano, and Shaikh (2021) by allowing units to be paired according to functions of the covariates instead of the covariates themselves, and furthermore allowing the function to be random and dependent on auxiliary data. To begin, I make the following mild moment restriction on the distributions of potential outcomes.

ASSUMPTION 2: $E\left[Y_i^2(d)\right] < \infty$ *for* $d \in \{0,1\}$.

A. *Pairing on Nonrandom Functions*

I provide general results when units are paired according to a measurable function h that maps from the support of X_i into $\mathbf{R}$. The results can be easily specialized to the procedures in Section III. Let $\pi^h \in \Pi_n$ be such that $h_{\pi^h(1)} \leq \ldots \leq h_{\pi^h(2n)}$ and define the stratification that pairs units according to h as

$$\lambda^h\left(X^{(n)}\right) = \left\{\left\{\pi^h(2s-1), \pi^h(2s)\right\} : 1 \leq s \leq n\right\}.$$

To describe the requirements on h, define $\mathbf{H}$ to be the set of all measurable functions mapping from the support of X_i into $\mathbf{R}$ such that the following three conditions hold:

(i) $0 < E\left[\text{var}\left[Y_i(d) \mid h(X_i)\right]\right]$ for $d \in \{0,1\}$.

(ii) $E\left[Y_i^r(d) \mid h(X_i) = z\right]$ is Lipschitz in z for $r = 1,2$ and $d = 0,1$.

(iii) $E\left[h^2(X_i)\right] < \infty$.

Condition (i) is a mild restriction to rule out degenerate situations and to permit the application of suitable laws of large numbers and central limit theorems, and (iii) is another mild moment restriction to ensure the pairs are "close" in the limit. Some restrictive primitive conditions for (ii) are provided in online Appendix B.2. I assume h lies in the set $\mathbf{H}$.

ASSUMPTION 3: $h \in \mathbf{H}$.

The next theorem establishes the limiting distribution of $\hat{\theta}_n$ when units are paired according to h, where h satisfies Assumption 3.

THEOREM 3: *Suppose the treatment assignment scheme satisfies Assumption 1, the distribution of the data satisfies Assumption 2, and h satisfies Assumption 3. Then, when units are paired according to h, as* $n \to \infty$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, \varsigma_h^2\right),$$

where

$$\varsigma_h^2 = \text{var}\left[Y_i(1)\right] + \text{var}\left[Y_i(0)\right] - \frac{1}{2}E\left[\left(E\left[Y_i(1) + Y_i(0) \mid h(X_i)\right] - E\left[Y_i(1) + Y_i(0)\right]\right)^2\right]. \tag{11}$$

REMARK 4: *In online Appendix C.1, I show the minimum of* ς_h^2 *over* $h \in \mathbf{H}$ *occurs when* $h = g$. *The law of iterated expectations implies*

$$\varsigma_h^2 = \text{var}\left[Y_i(1)\right] + \text{var}\left[Y_i(0)\right] - \frac{1}{2}\text{var}\left[g(X_i)\right] + \frac{1}{2}E\left[\text{var}\left[g(X_i) \mid h(X_i)\right]\right],$$

so the increase in the limiting variance when pairing according to h instead of g is proportional to $E\left[\text{var}\left[g(X_i) \mid h(X_i)\right]\right]$, *the average conditional variance of* $g(X_i)$ *given* $h(X_i)$. *Therefore, among all functions* $h \in \mathbf{H}$, *choosing an h that minimizes* $E\left[\text{var}\left[g(X_i) \mid h(X_i)\right]\right]$ *is optimal. Intuitively, the optimal h explains the largest proportion of the variation in* $Y(1)$ *and* $Y(0)$.

REMARK 5: *Theorem 3 immediately leads to three insights on the comparison of different treatment assignment schemes:*

(*i*) *Stratifications with a small number of large strata can be characterized by a function h mapping from the support of X_i into $\{1, \ldots, S\}$, such that unit i is in stratum s if and only if $h(X_i) = s$. Bugni, Canay, and Shaikh (2018) show the limiting variance of $\hat{\theta}_n$ under such a stratification equals ς_h^2. Therefore, $\varsigma_h^2 > \varsigma_g^2$ unless $g(X_i) = E[g(X_i) \mid h(X_i)]$ with probability one, which means $g(X_i)$ is constant within each stratum.*

(*ii*) *The stratification $\{\{1, \ldots, 2n\}\}$ with all units in one stratum can be written as $\lambda^{h_c}(X^{(n)})$, where h_c is a constant function. For any h that satisfies Assumption 3, $\varsigma_{h_c}^2 > \varsigma_h^2$ unless $E[g(X_i) \mid h(X_i)]$ is constant with probability one. As a result, in terms of the limiting variance of $\hat{\theta}_n$, any stratification is weakly better than not stratifying at all.*

(*iii*) *It follows from straightforward calculation that for any $h \in \mathbf{H}$, ς_h^2 is weakly less than and typically strictly less than the limiting variance of $\hat{\theta}_n$ when treatment status is determined by i.i.d. coin flips.*

Theorem C.3 in the online Appendix studies a procedure that "breaks up" a stratification with a small number of large strata. I further allow the treated fractions to vary across strata. I show the limiting variance of $\hat{\theta}_n$ is weakly smaller if I implement small-strata designs similar to the ones described in Remark 2 separately within each stratum.

Next, I consider inference for the ATE when units are paired according to $h \in \mathbf{H}$. For any prespecified $\theta_0 \in \mathbf{R}$, I am interested in testing

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0 \tag{12}$$

at level $\alpha \in (0, 1)$. To do so, it suffices to provide a consistent estimator for the limiting variance ς_h^2 in (11). To describe such an estimator, for $d \in \{0, 1\}$, define the variance estimator among units with $D = d$ as

$$\hat{\sigma}_n^2(d) = \frac{1}{n} \sum_{1 \leq i \leq 2n : D_i = d} \big(Y_i - \hat{\mu}_n(d)\big)^2.$$

In addition, define

$$\hat{\rho}_n = \frac{2}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} \big(Y_{\pi^h(4j-3)} + Y_{\pi^h(4j-2)}\big)\big(Y_{\pi^h(4j-1)} + Y_{\pi^h(4j)}\big) \tag{13}$$

and

$$\hat{\varsigma}_{h,n}^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\hat{\rho}_n + \frac{1}{2}\big(\hat{\mu}_n(1) + \hat{\mu}_n(0)\big)^2. \tag{14}$$

The calculation in online Appendix C.4 shows $\hat{\varsigma}_{h,n}^2$ is nonnegative. The correction term $\hat{\rho}_n$ is constructed by averaging the product of the sum of the outcomes of adjacent *pairs of pairs*, as in Bai, Romano, and Shaikh (2021).

The following theorem shows the variance estimator in (14) is consistent for the limiting variance in (11).

THEOREM 4: *Suppose the treatment assignment scheme satisfies Assumption 1, the distribution of the data satisfies Assumption 2, and h satisfies Assumption 3. Then, when units are paired according to h, as* $n \to \infty$, $\hat{\varsigma}^2_{h,n}$ *defined in* (14) *satisfies*

$$\hat{\varsigma}^2_{h,n} \xrightarrow{P} \varsigma^2_h.$$

REMARK 6: *The correction term* $\hat{\rho}_n$ *in* (13) *is crucial for the consistency of* $\hat{\varsigma}^2_{h,n}$ *in* (14). *In commonly used tests including the two-sample t-test* (*Riach and Rich 2002; Gelman and Hill 2007; Duflo, Glennerster, and Kremer 2007*) *and the "matched pairs" t-test* (*Moses 2006; Hsu and Lachenbruch 2007; Armitage, Berry, and Matthews 2008; Imbens and Rubin 2015; Athey and Imbens 2017*), *the test statistics are studentized by variance estimators whose limits in probability are weakly greater than* ς^2_h, *so these tests are asymptotically conservative in the sense that the limiting size is no greater than and typically strictly less than the nominal level. For instance, a 5 percent level test could have a size of 1 percent. In fact, the limiting size of the "matched pairs" t -test is strictly less than the nominal level unless* (9) *holds. I refer interested readers to Bai, Romano, and Shaikkh* (*2021*) *for details.*

REMARK 7: *Let* $\tilde{h}_m$ *be a function of the pilot data such that* $\tilde{h}_m \in \mathbf{H}$ *with probability one. Then, the proof of Theorems 3 and 4 implies the conclusions therein hold for* $h = \tilde{h}_m$ *conditional on the pilot data with probability one. Because probabilities are bounded between zero and one and hence are uniformly integrable, the same conclusions hold unconditionally too. In particular, the variance estimator in* (14) *is valid even when the sample size of the pilot experiment is small and fixed.*

B. *Pairing on Random Functions*

The discussion in the last subsection applies to settings where units are paired according to a fixed function $h \in \mathbf{H}$ or a random function $\tilde{h}_m$ such that $\tilde{h}_m \in \mathbf{H}$ with probability one. Such settings are most relevant when the pilot sample size m is small. Next, I consider settings where $\tilde{h}_m$ converges to a fixed function $h \in \mathbf{H}$ in a suitable sense as $m \to \infty$. Let Q_X denote the marginal distribution of X_i.

ASSUMPTION 4: *Let* $\tilde{h}_m$ *be a random function depending on the auxiliary data that maps from the support of* X_i *into* $\mathbf{R}$, *and satisfies*

$$\int |\tilde{h}_m(x) - h(x)|^2 Q_X(dx) \xrightarrow{P} 0$$

as $m \to \infty$.

Assumption 4 is commonly referred to as the L^2 consistency of the $\tilde{h}_m$ for h. When the dimension of X_i is fixed and suitable smoothness conditions hold, L^2 consistency is satisfied by series and sieves estimators (Newey 1997; Chen 2007) and kernel estimators (Li and Racine 2007). In some high-dimensional settings, when the dimension of X_i increases with n at suitable rates, it is satisfied by the least absolute shrinkage and selection operator (LASSO) estimator (Bühlmann and van de Geer 2011; Belloni, Chernozhukov, and Hansen 2014), regression trees and

random forests (Györfi et al. 2002; Wager and Walther 2015), neural nets (White 1990; Farrell, Liang, and Misra 2018), and support vector machines (Steinwart and Christmann 2008). The results therein are either exactly as stated in Assumption 4 or one of the following:

(i) $\sup_x |\tilde{h}_m(x) - h(x)| \xrightarrow{P} 0$ as $m \to \infty$.

(ii) $E\left[|\tilde{h}_m(x) - h(x)|^2\right] \to 0$ as $m \to \infty$.

It is straightforward to see (i) implies Assumption 4. Furthermore, (ii) implies Assumption 4 by Markov's inequality.

The next theorem shows that if $\tilde{h}_m$ is L^2 consistent for h, then as the sample sizes of both the pilot and main experiments increase, the limiting variance of $\hat{\theta}_n$ when units are paired according to $\tilde{h}_m$ is the same as that when units are paired according to h.

THEOREM 5: *Suppose the treatment assignment scheme satisfies Assumption 1, the distribution of the data satisfies Assumption 2, h satisfies Assumption 3, and $\tilde{h}_m$ satisfies Assumption 4. Then, when units are paired according to $\tilde{h}_m$, as $m, n \to \infty$,*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, \varsigma_h^2\right),$$

and

$$\hat{\varsigma}^2_{\tilde{h}_m, n} \xrightarrow{P} \varsigma_h^2.$$

REMARK 8: *Note the assumptions in Theorems 3 and 4 and those in 5 are nonnested and differ in whether the sample size of the pilot experiment stays fixed or goes to infinity in the asymptotic framework. Theorems 3 and 4 do not require $\tilde{h}_m$ to be consistent for any fixed function and allow m to be fixed asymptotically, but require $\tilde{h}_m \in \mathbf{H}$ with probability one. On the other hand, Theorem 5 does not require $\tilde{h}_m \in \mathbf{H}$ but requires $m \to \infty$ and $\tilde{h}_m$ to be L^2 consistent for h.*

C. *Pairing on Multiple Covariates*

We briefly comment on inference when units are paired according to multiple covariates. For the settings in (8) and (10), the variance estimators are slightly more complicated than that in (14) because the distances in (8) and (10) cannot be written as distances between two scalars, but the correction term is similar in spirit to (13). I defer the discussion to online Appendix C.5. In addition, note combining data from both the pilot and main experiments for estimation and inference is possible. I defer the discussion to online Appendix C.6.

When units are paired using multiple covariates, the simulation evidence in Section V shows that when the sample size is not large enough relative to the number of covariates, the size of the test is often strictly smaller than the nominal level. The reason is that the asymptotic results rely on the assumption that units are "close," in

the sense that a suitable normalization of the sum of distances between the covariates within each pair is close to zero. When the sample size is not large enough relative to the number of covariates, the procedures here suffer from the curse of dimensionality, so the units paired together are not close enough in terms of their covariates, and hence, the asymptotic results do not approximate the finite-sample distribution of $\hat{\theta}_n$ very well. The problem is mitigated by matching units into sets of four instead of pairs. Specifically, I first pair the units and then pair the pairs using the midpoints of all pairs, as in Section 4 of Bai, Romano, and Shaikh (2021). In online Appendix C.7, I propose a valid test for (12) when units are matched into sets of four. Simulation evidence in Section V shows the size of my proposed test is close to the nominal level in finite sample.

V. Simulation

In this section, I examine the performance of the practical procedures in Section III and the inference methods in Section IV via a simulation study calibrated to a systematically selected set of ten RCTs from recent issues of the *American Economic Journal: Applied Economics*. I focus on settings with small or no pilots, because they are the most common settings in practice. I searched the 11 issues from October 2018 to April 2021 and collected 28 papers running RCTs. I exclude 11 papers for which the treatment is assigned at the cluster level instead of the unit level. I further exclude four papers with a network/spillover structure. I also exclude one paper for which the sample size is too small (less than 20). Finally, I exclude two papers for which the data are confidential. I end up with ten papers, which are listed in Table 1. For each paper, I list whether the baseline outcome is available, the original randomization method, and the number of additional covariates besides the baseline outcome in the main regression specification of the paper. The full details of the data are available in online Appendix D.

For each paper, I denote the sample size by $2n$. I use the original sample except for Barrera-Osorio, Linden, and Saavedra (2019), where the original data contain 15,759 observations and 16 covariates, so one replication in the simulation takes almost six hours. For Barrera-Osorio, Linden, and Saavedra (2019) only, I take half of the observations as the population to reduce the computational time for one replication to about an hour, which is about the same as that using the next largest dataset. I begin by imputing the unobserved potential outcomes. For the ith unit, I denote the original data by $(Y_i^*, D_i^*, X_{1i}^*, X_{2i}^*)$, where Y_i^* denotes its observed outcome, D_i^* denotes its treatment status, X_{1i}^* denotes its baseline outcome if available, and X_{2i}^* denotes the other covariates in the main regression specification of the paper. Let $Y_i^*(1), Y_i^*(0)$ denote the potential outcomes for the ith unit. For the ith unit, $Y_i^*(D_i)$ is observed, and I construct $Y_i^*(1 - D_i)$ according to the following models:

Model 1: $Y_i^*(1) = Y_i^*(0)$.

Model 2: $Y_i^*(1 - D_i) = Y_{j(i)}^*$, where the $j(i)$th unit is the closest unit to the ith unit in terms of the Mahalanobis distance of (X_1^*, X_2^*) among units with $D_j \neq D_i$.

Table 1—Papers Selected for the Simulation Study

Paper	Baseline	Original stratification	Number of covariates
1. Herskowitz (2021)	×	none	4
2. Lee et al. (2021)	✓	rerandomization	4
3. Abel et al. (2020)	✓	gender	7
4. Gerber et al. (2020)	✓	state	11
5. Deserranno et al. (2019)	✓	none	1
6. Barrera-Osorio et al. (2019)	×	baseline grade, gender	16
7. Himmler et al. (2019)	×	GPA (4 strata)	8
8. Abel et al. (2019)	✓	none	10
9. de Mel et al. (2019)	✓	region, sector	7
10. Lafortune et al. (2018)	✓	none	8

Notes: For each paper, I list whether the baseline outcome is available, the original stratification method, and the number of covariates besides the baseline outcome in the main regression specification of the paper. Lee et al. (2021) assign treatment status by rerandomization, and the other papers use stratified randomization (possibly with only one stratum).

Model 3: $Y_i^*(1-D_i) = Y_{j(i)}^*$, where the $j(i)$th unit is the closest unit to the ith unit in terms of the baseline outcome X_1^*, if available, among units with $D_j \neq D_i$.

In model 1, treatment effects are homogeneous, so (9) holds. In models 2 and 3, treatment effects are heterogeneous. Note the baseline outcome predicts the potential outcomes better in model 3 than in model 2.

For each replication, I simulate new data $\left(\left(Y_i(1), Y_i(0), X_{i1}, X_{2i}\right) : 1 \leq i \leq 2n\right)$ by drawing $2n$ units from the empirical distribution of $\left(\left(Y_i^*(1), Y_i^*(0), X_{1i}^*, X_{2i}^*\right) : 1 \leq i \leq 2n\right)$, so each unit in the original data is drawn with equal probability, with replacement.

For each paper, I implement several stratifications. Note the covariates may have been selected ex post by authors on the basis of predictive power, while ideally I would like to include only the covariates specified in the preanalysis plans. Unfortunately, only one paper includes such information in the AEA RCT Registry, and the preregistered covariates are the same as those in the regression analysis. I stratify on the baseline outcome whenever it is available. To stratify based on data from pilot experiments, I reached out to the authors of all ten papers to request pilot data. Nine out of ten replied, eight of whom said they did not run a pilot, and the last said they ran a pilot, but the data were lost. Therefore, I simulate pilot data by drawing with replacement from the empirical distribution at a sample size of $\lfloor 0.2 \cdot (2n) \rfloor$. To study the setting with a small and fixed pilot, I fix the pilot data throughout all replications. I also consider several stratifications with matched sets of four, as in Athey and Imbens (2017). Specifically, strata are constructed by first pairing the units and then pairing the pairs according to their midpoints, as in Section 4 of Bai, Romano, and Shaikh (2021). The complete list of stratifications are as follows:

(a) MP X: matched pairs to minimize the sum of the squared Mahalanobis distances in (8) of all covariates X.

(b) MS X: matched sets of four to minimize the sum of the squared Mahalanobis distances of X.

(c) MP base: matched pairs according to the baseline outcome, if available.

(d) MS base: matched sets of four according to the baseline outcome, if available.

(e) MP X2: matched pairs to minimize the sum of the squared Mahalanobis distances of X_2, namely, all covariates in the main regression specification except the baseline outcome.

(f) MP pilot: matched pairs according to $\tilde{g}_m$ from the pilot, where $\tilde{g}_m$ is given by the OLS.

(g) MP pen: the penalized matched pairs given by minimizing the sum of the distances in (10) of all covariates.

(h) Origin: stratification used in the original paper, if not one of (i) through (vii).

(i) None: no stratification, meaning all units are in one stratum and exactly half are treated.

(i') None-reg: no stratification with the estimator given by the OLS estimator of the coefficient on D in the linear regression of Y on a constant, D, and X.

The original stratifications are listed in Table 1. I do not consider rerandomization in Lee et al. (2021), because the exact implementation is unclear from the original paper, and inference under rerandomization is complicated. See also Remark 3 for a comparison between rerandomization and matched-pair designs. Online Appendix D contains the results for several additional stratifications. Although it is interesting to investigate the performance of regression adjustment with the stratifications in Origin, inference with regression adjustment under stratified randomization is still an open question.

I consider the following inference methods:

(i) Matched pairs: (a) (adj) the adjusted t-test with the variance estimator in (14); (b) (MPt) the test with the variance estimator in Theorem 10.1 of Imbens and Rubin (2015), which is equivalent to the "matched pairs" t-test in Bai, Romano, and Shaikh (2021).

(ii) Matched sets of four: (adj4) the adjusted t-test with the variance estimator in (S.50) in online Appendix D;

(iii) Original: the test in equation (23) of Bugni, Canay, and Shaikkh (2018), which is asymptotically exact under stratified randomization.

(iv) No stratification: without regression adjustment, the two-sample t-test with the variance estimator given by $\hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0)$; with regression adjustment, White's heteroskedasticity-robust standard error.

For matched sets of four, Athey and Imbens (2017) propose a test in a sampling framework different from ours. I show in online Appendix C.7 that because of the differences in sampling frameworks, the test in Athey and Imbens (2017) does not control size in my setting unless the conditional ATEs are homogeneous. Therefore, I defer these simulation results to online Appendix D.

For each paper, each model, and each stratification, across 1,000 replications, I calculate three metrics of performance: (i) the MSE of estimating θ using $\hat{\theta}_n$, reflecting the precision of the estimator; (ii) the average rejection probability of testing (12) for $\theta_0 = \theta$, reflecting the size of the test; and (iii) the average standard error, which directly determines the length of the confidence interval for the ATE.

The main results of the simulation study are summarized in Table 2. I only report the summary statistics across all papers and models and defer the raw numbers to online Appendix D. In particular, for each stratification, I report the average and $[\min, \max]$ across all papers and models of

(i) the ratio between the MSE under the particular stratification and the MSE under no stratification,

(ii) the size of the test, and

(iii) the ratio between the average standard error under the particular stratification and the average standard error under no stratification.

Rows are labeled according to the stratifications.

A. *Statistical Precision*

In this subsection, I discuss major takeaways about statistical precision (specifically, MSE) from Table 2. I focus on five questions that are particularly relevant to empirical practice.

First, how much statistical precision are researchers leaving on the table with their current stratification methods? To answer this question, I compare the MSEs under the stratifications used in the original paper (origin) and the MSEs when pairing according to the baseline outcome (MP base). Relative to the original stratifications used in those ten papers, if the researchers had just paired the units according to their baseline outcomes, the MSE would be 24 percent smaller on average and 56 percent smaller in some cases. In fact, in many models, the MSE under original stratification is almost the same as the MSE when not stratifying. As a result, if researchers had paired the units according to their baseline outcomes, they could have reached the same statistical precision with a much smaller sample size.

Second, how much statistical precision would researchers leave on the table by pairing units according to the baseline outcome rather than using more complicated feasible methods? Note MP X usually has the smallest MSEs across all methods. On average, the MSE under MP base is about 39 percent larger than that under MP X, and 10 percent larger than that under MS X. As a result, researchers indeed sacrifice some statistical precision by pairing units according to the baseline outcome along instead of pairing or matching into sets of four according to all covariates. Note,

TABLE 2—SUMMARY STATISTICS FOR MSES, SIZE, AND STANDARD ERRORS FOR EACH STRATIFICATION ACROSS ALL PAPERS AND MODELS

	Stratification	MSE (ratio versus none)	Size (percent)		SE (ratio versus none)	
			adj/adj4	MPt	adj/adj4	MPt
(a)	MP X	0.549	2.267	3.533	0.870	0.810
		[0.304, 0.830]	[0.300, 4.800]	[1.700, 6.400]	[0.720, 0.968]	[0.530, 1.092]
(b)	MS X	0.695	5.148	–	0.828	–
		[0.464, 0.927]	[3.600, 6.600]	–	[0.663, 0.946]	–
(c)	MP base	0.762	4.771	4.781	0.886	0.885
		[0.404, 1.030]	[3.200, 5.800]	[2.800, 6.200]	[0.629, 0.989]	[0.633, 1.003]
(d)	MS base	0.792	5.257	–	0.882	–
		[0.404, 0.982]	[4.400, 6.300]	–	[0.629, 0.991]	–
(e)	MP X2	0.685	3.126	3.852	0.919	0.874
		[0.362, 0.923]	[0.500, 6.900]	[1.300, 6.300]	[0.840, 0.979]	[0.608, 1.091]
(f)	MP pilot	0.666	3.578	4.107	0.879	0.855
		[0.387, 0.873]	[2.100, 5.500]	[2.700, 5.600]	[0.653, 0.969]	[0.605, 1.063]
(g)	MP pen	0.542	2.296	3.330	0.862	0.806
		[0.280, 0.826]	[0.400, 4.800]	[1.300, 5.600]	[0.625, 0.965]	[0.506, 1.093]
(h)	Origin	1.007	5.292	–	0.980	–
		[0.918, 1.114]	[3.700, 7.300]	–	[0.950, 0.999]	–
(i)	None (benchmark)	1.000	5.089	–	1.000	–
		[1.000, 1.000]	[3.600, 6.900]	–	[1.000, 1.000]	–
(i')	None-reg	0.948	4.900	–	0.980	–
		[0.775, 1.012]	[3.200, 6.700]	–	[0.880, 1.034]	–

Notes: For each stratification, I report summary statistics across all papers and models of (i) the ratio between the MSE under the particular stratification and the MSE under no stratification, (ii) the size of testing (12) for $\theta_0 = \theta$ at the 5 percent level, in percentage, and (iii) the ratio between the average standard error under the particular stratification and the average standard error under no stratification. The tests used in this table are as follows: for matched-pair designs, the adjusted t-test with the variance estimator in (14) (adj) and the test in Imbens and Rubin (2015) (MPt); for matched sets of four, the adjusted t-test with the variance estimator in (S.50) in online Appendix C.7 (adj4); for the original stratifications, the test in equation (23) of Bugni, Canay, and Shaikh (2018); for no stratification, the two-sample t-test; for the regression-adjusted estimator, the t-test with White's heteroskedasticity-robust standard error. For each metric, I show the average and $[\min, \max]$ across all papers and models. Rows are labeled according to the stratifications. Columns are labeled according to the metrics. For size and standard errors, the second column corresponds to MPt for matched-pair designs and the first column corresponds to the other tests. The definitions of the stratifications can be found in the main text.

however, that MP base picks up about half of the difference between the MSEs under the best feasible method (MP X) and the status-quo methods (Origin).

Third, what is the value of collecting just the baseline outcome rather than other covariates? To answer this question, I compare the MSEs under MP base and MP X2. Note the MSE under MP base is on average only 11 percent larger than and sometimes almost the same as that under MP X2. As a result, the statistical precision of pairing according to the baseline outcome alone is comparable to the statistical precision of pairing according to all other covariates. Note the number of other covariates is close to or larger than ten in most cases, so the per-covariate return for collecting all of them is limited relative to collecting the baseline outcome alone.

Fourth, are matched sets of four better than matched pairs in terms of the MSE? To answer this question, I compare the MSEs of the MS methods and the MP methods. The MSE under MS base is 4 percent larger than that under MP base, and the MSE under MS X is 27 percent larger than that under MP X. Therefore, the statistical precision is higher with matched pairs. The difference is pronounced when I match

according to multiple covariates but tiny when I match only according to the baseline outcome.

Fifth, does the best pairing method based on a small pilot dominate pairing according to X? In other words, what is the value of having a small pilot? First, note the MSE under MP pen is 19 percent smaller than that under MP pilot, so the penalized matched-pair design indeed has better precision than the naïve plug-in procedure. Meanwhile, the MSE under MP pen is almost the same as that under MP X, and even in the most favorable case, it is only about 8 percent smaller. Therefore, the return for a small pilot in terms of the MSE is negligible.

I also study the performance of regression-adjusted estimators in stratifications with one stratum (none-reg). With one stratum, the regression-adjusted estimator usually has slightly smaller MSEs than the difference-in-means estimator. In almost all cases, however, the MSE is larger than those under all methods with matched pairs or matched sets of four, regardless of whether all or only a subset of the covariates in the regression adjustment are used in the matching. Online Appendix D contains the results for the regression-adjusted estimator in Lin (2013), which additionally includes the interactions of treatment status and covariates. The results are qualitatively similar to those for none-reg. Therefore, most of the gains in precision from matching according to the covariates cannot be retrieved by controlling for the same covariates via ex post regression adjustment.

B. *Inference Methods*

Next, I discuss the properties of the inference methods. I start with the size of the tests. For matched pairs, note both the adjusted t-test and the "matched pairs" t-test control size well across all papers and models. When I pair units according to the baseline outcome (MP base), the size of the adjusted t-test is almost always close to 5 percent. The size of the "matched pairs" t-test is also close to 5 percent. The underrejection phenomenon for the "matched pairs" t-test in Remark 6 is very mild, reflecting the treatment effects heterogeneity is not very large. Several relatively noticeable cases include, for model 2 of paper 5, the size of the "matched pairs" t-test is 3.9 percent but the size of the adjusted t-test is 5.4 percent; for model 3 of paper 5, the size of the "matched pairs" t-test is 2.8 percent but the size of the adjusted t-test is 3.2 percent.

When I pair units according to multiple covariates (MP X, MP X2, and MP pen), the "matched pairs" t-test is still conservative except in model 1, for the same reason as mentioned in Remark 6. At the same time, the adjusted t-test also becomes conservative—its size is often smaller than 5 percent. The reason is that the asymptotic results in this paper rely on the assumption that units are "close," in the sense that a suitable normalization of the sum of the distances between the covariates within each pair is close to zero. When pairing according to multiple covariates, however, the procedures suffer from the curse of dimensionality, so the units paired together are not close enough in terms of their covariates. Therefore, the asymptotic results do not approximate the finite-sample distribution of $\hat{\theta}_{\eta}$ very well, and my variance estimator does not approximate the actual variance of $\hat{\theta}_{n}$ very well.

When matching according to multiple covariates, the conservativeness of the tests is somewhat alleviated by matching units into sets of four instead of pairs. The size of the test under MS X is close to 5 percent even when the size of the test

under MP X is much smaller than 5 percent. On the other hand, such a difference is virtually nonexistent when I match only according to the baseline outcome—the size of the test under MP base and that under MS base are both close to 5 percent. Our current asymptotic framework cannot explain the difference in size because the variance estimators for both MP and MS methods are consistent for the limiting variance. The exact reason for the difference is an interesting topic for future work.

I now turn to the standard errors. The findings are mostly similar to those for the MSEs, though with some important exceptions. The standard error under MP base is 10 percent smaller on average and 34 percent smaller in some cases than that under origin. At the same time, although the MSE under MP X is smaller than that under MS X, the standard error of MP X is larger than that under MS X. In fact, the standard error under MP X is about the same as that under MP base, and the standard error under MP X2 is often larger than that under MP base. Therefore, although pairing according to multiple covariates is desirable for the MSE, it is often not the best choice for inference, because the standard error is too large and the size of the test could be strictly smaller than the nominal level. By matching units into sets of four instead of pairs according to the same set of covariates, researchers could lower the standard error and bring the size close to the nominal level. Note, however, that the MSE will increase, as discussed in Section VA.

I emphasize the validity of my tests relies on the assumptions on the sampling framework. In this paper, I assume units are drawn from a superpopulation, and the potential outcomes and the covariates are random. De Chaisemartin and Ramirez-Cuellar (2021), on the other hand, study a finite-population setting in which the potential outcomes and the covariates are fixed. Such a setting is particularly relevant if we have a convenience sample instead of a random sample drawn from a large population. De Chaisemartin and Ramirez-Cuellar (2021) show in these settings that if the number of pairs is small, then the tests in my paper and Bai, Romano, and Shaikh (2021) may not control size, and the "matched pairs" *t*-test in Imbens and Rubin (2015) could become preferable.

C. *Multiple Outcomes*

Finally, I consider settings with multiple outcomes. I take the example of Abel et al. (2019), where a primary outcome, a secondary outcome, and the baseline outcomes of both are available. The paper studies job-searching behaviors. The primary outcome is the search hours and the secondary outcome is the number of applications sent. I study the estimation of the ATE of the secondary outcome. The missing potential outcomes are imputed as in Model 1, assuming the treatment effect is zero for everyone, and Model 3, using the nearest neighbor in terms of the baseline value of the secondary outcome. I consider the following stratifications.

> MP 2: matched pairs according to the baseline value of the secondary outcome.
>
> MS 2: matched sets of four according to the baseline value of the secondary outcome.

MP 1: matched pairs according to the baseline value of the primary outcome.

MS 1: matched sets of four according to the baseline value of the primary outcome.

MP 1+2: matched pairs to minimize the sum of the squared Mahalanobis distances in (8) of the baseline values of both outcomes.

MS 1+2: matched sets of four to minimize the sum of the squared Mahalanobis distances of the baseline values of both outcomes.

None: no stratification, meaning all units are in one stratum and exactly half are treated.

In light of the results in Section VA, I only consider the adjusted *t*-tests. For each stratification, I calculate the MSE, the size of testing (12) with $\theta_0 = \theta$, and the average standard error. The results are displayed in Table 3. Rows are labeled according to the stratifications. As expected, because the secondary outcome is of interest, for both models, stratifying on the baseline value of the secondary outcome produces smaller MSEs than stratifying on the baseline value of the primary outcome. For both models, the MSEs when stratifying on the baseline value of the primary outcome are close to the MSEs with one stratum, reflecting that the baseline value of the primary outcome is a poor predictor of the secondary outcome. The smallest MSE is attained by pairing according to the baseline values of both outcomes, and the second smallest MSE is attained by matching units into sets of four according to both baseline outcomes. In all models, the size of the test is close to the nominal level. The test under MP 1+2 slightly underrejects for the same reason as in Section VB, although the problem is mild because I only match on two variables. The standard errors are ranked in the same way as the MSEs except for that of MP 1+2. In both models, MS 1+2 produces the smallest standard errors across all methods.

VI. Discussion and Recommendations for Empirical Practice

Based on the theoretical results, in settings with large pilots, I recommend researchers to pair units according to the estimated index function from nonparametric regressions. If the conditional ATEs are homogeneous and researchers have access to a large observational dataset from the same population as that of the main experiment, then I recommend pairing according to predicted values from nonparametric regressions of the outcome on the covariates in the observational dataset. In what follows, I focus on settings with small or no pilots because these are the most common settings in practice.

A simple approach that researchers can take, assuming there is only one primary outcome of interest and its baseline value is available, is to pair units according to the baseline outcome. Indeed, if the baseline outcome is the only available covariate and the index function is monotonic in it, then my theoretical results show pairing units according to the baseline outcome is optimal at any sample size. The simulation results in Section V also show pairing according to the baseline outcome

TABLE 3—MSEs, SIZE, AND STANDARD ERRORS FOR ESTIMATING THE ATE OF THE SECONDARY OUTCOME IN ABEL ET AL. (2019)

	Model 1 $\theta = 0$			Model 3 $\theta = 0.4449$		
	MSE (ratio versus none)	Size (percent)	SE (ratio versus none)	MSE (ratio versus none)	Size (percent)	SE (ratio versus none)
MP 2	0.760	5.9	0.835	0.645	4.7	0.799
MS 2	0.756	6.0	0.835	0.689	5.8	0.799
MP 1	0.988	4.3	0.980	1.010	4.4	0.986
MS 1	1.117	6.6	0.980	1.070	5.7	0.987
MP 1+2	0.558	4.2	0.769	0.568	4.0	0.783
MS 1+2	0.615	4.9	0.760	0.598	4.3	0.777
None	1.000	4.5	1.000	1.000	4.7	1.000

Notes: For each stratification, I report (i) the MSE, (ii) the size of testing (12) for $\theta_0 = \theta$ at the 5 percent level, in percentage, and (iii) the average standard error. The parameter of interest is the ATE of the secondary outcome. The tests used in this table are as follows: for matched-pair designs, the adjusted t-test with the variance estimator in (14) (adj); for matched sets of four, the adjusted t-test with the variance estimator in (S.50) of the online Appendix (adj4). Rows are labeled according to stratifications. Columns are labeled according to the models and metrics. I display the value of θ for each model. The definitions of the stratifications can be found in the main text.

improves upon the status quo methods in terms of both the MSE and the standard error of the difference-in-means estimator. Further, this approach has the advantage of simplicity.

When multiple covariates are available, an attractive alternative is to match units into pairs or sets of four according to the baseline outcome and other covariates. Unless the number of covariates is very small, I recommend matched sets of four over pairs because the standard error is usually smaller with matched sets of four. In my simulation study, when I use the baseline outcome together with all the covariates that the authors control for in their regressions, matching units into sets of four leads to smaller MSEs and standard errors than pairing on the baseline outcome alone. Note that the good performance of this design could be due to the fact that the authors selected the covariates with the best predictive power ex post, something that is not feasible at the time of randomization. Nevertheless, forming sets of four according to the baseline outcome and other covariates is an attractive alternative to pairing according to the baseline outcome alone.

In my simulation study, when multiple outcomes are of interest, pairing on one of them may not improve the MSE of the other outcomes. In those settings, researchers could consider matching units into sets of four to minimize the sum of the squared Mahalanobis distances of the baseline values of all outcomes of interest and perhaps some additional covariates.

A further question is whether pilot experiments are worth running for the sole purpose of improving the precision for estimating the ATE. My simulation results only show minor gains in statistical precision when using pilot-based stratifications instead of matching directly on the covariates. Therefore, although pilot experiments are essential for other aspects of the design of the main experiment, they are not as helpful in improving the precision of the estimator for the ATE.

Another natural question is whether one can retrieve the gains in precision from matched pairs or sets of four units by controlling for the same set of covariates

through ex post regression adjustment. Our simulation results show the answer is negative; although with one stratum, regression adjustment slightly lowers the MSE and the standard error relative to the unadjusted difference-in-means estimator. I further note the difference-in-means is unbiased for the ATE in finite sample under all stratifications considered in this paper, while the regression-adjusted estimators are only consistent for the ATE asymptotically. A very interesting direction for future work is to combine regression adjustment with stratifications defined by matched pairs or matched sets of four units.

For inference, researchers can use the test with the variance estimator in (14). They can also use the test in Theorem 10.1 of Imbens and Rubin (2015), which is valid albeit sometimes conservative. Finally, I emphasize that my framework assumes units are drawn from a superpopulation and the potential outcomes and the covariates are random. If we have a convenience sample instead of a random sample drawn from a large population, and the sample size is small, then de Chaisemartin and Ramirez-Cuellar (2021) show the tests in this paper and Bai, Romano, and Shaikh (2021) may not control size, and the test in Imbens and Rubin (2015) could become preferable.

REFERENCES

Abadie, Alberto, and Guido W. Imbens. 2008. "Estimation of the Conditional Variance in Paired Experiments." *Annales d'E´conomie et de Statistique* 91–92: 175–87.

Abel, Martin, Rulof Burger, Eliana Carranza, and Patrizio Piraino. 2019. "Bridging the Intention-Behavior Gap? The Effect of Plan-Making Prompts on Job Search and Employment." *American Economic Journal: Applied Economics* 11 (2): 284–301.

Abel, Martin, Rulof Burger, Eliana Carranza, and Patrizio Piraino. 2019. "Replication data for: Bridging the Intention-Behavior Gap? The Effect of Plan-Making Prompts on Job Search and Employment." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Abel, Martin, Rulof Burger, and Patrizio Piraino. 2020. "The Value of Reference Letters: Experimental Evidence from South Africa." *American Economic Journal: Applied Economics* 12 (3): 40–71.

Abel, Martin, Rulof Burger, and Patrizio Piraino. 2020. "Data for: The Value of Reference Letters: Experimental Evidence from South Africa." *American Economic Journal: Applied Economics*, 12 (3): 40–71.

Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review* 99 (4): 1384–414.

Armitage, Peter, Geoffrey Berry, and John Nigel Scott Matthews. 2008. *Statistical Methods in Medical Research*. Hoboken, NJ: John Wiley and Sons.

Athey, Susan and Guido W. Imbens. 2017. "The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments*, Vol. 1, edited by Abhijit Banerjee and Esther Duflo, 73–140. Amsterdam: Elsevier.

Bai, Yuehao. 2022. "Replication Data for: Optimality of Matched-Pair Designs in Randomized Controlled Trials." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E171321V1.

Bai, Yuehao, Joseph P. Romano, and Azeem M. Shaikh. 2021. "Inference in Experiments with Matched Pairs." *Journal of the American Statistical Association* 1–12.

Barrera-Osorio, Felipe, Leigh L. Linden, and Juan E. Saavedra. 2019. "Medium- and Long-Term Educational Consequences of Alternative Conditional Cash Transfer Designs: Experimental Evidence from Colombia." *American Economic Journal: Applied Economics* 11 (3): 54–91.

Barrera-Osorio, Felipe, Leigh L. Linden, and Juan E. Saavedra. 2019. "Replication data for: Medium- and Long-Term Educational Consequences of Alternative Conditional Cash Transfer Designs: Experimental Evidence from Colombia." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Barrios, Thomas. 2013. "Optimal Stratification in Randomized Experiments." Unpublished.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *Review of Economic Studies* 81 (2): 608–50.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2018. "Experimental Evidence on Scaling Up Education Reforms in Kenya." Journal of Public Economics 168: 1–20.

Brown, Christina, and Tahir Andrabi. 2020. "Inducing Positive Sorting through Performance Pay: Experimental Evidence from Pakistani Schools." Unpublished.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–32.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh. 2018. "Inference under Covariate-Adaptive Randomization." *Journal of the American Statistical Association* 113 (524): 1784–96.

Bühlmann, Peter, and Sara van de Geer. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin, Heidelberg: Springer-Verlag.

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-nonparametric Models." In *Handbook of Econometrics*, Vol. 6, edited by James J. Heckman and Edward E. Leamer, 5549–5632. Amsterdam: Elsevier.

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté. 2015. "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco." *American Economic Journal: Applied Economics* 7 (1): 123–50.

de Chaisemartin, Clement, and Jamie Ramirez-Cuellar. 2021. "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" Unpublished.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2019. "Labor Drops: Experimental Evidence on the Return to Additional Labor in Microenterprises." *American Economic Journal: Applied Economics* 11 (1): 202–35.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2019. "Replication data for: Labor Drops: Experimental Evidence on the Return to Additional Labor in Microenterprises." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Deserranno, Erika, Miri Stryjan, and Munshi Sulaiman. 2019. "Leader Selection and Service Delivery in Community Groups: Experimental Evidence from Uganda." *American Economic Journal: Applied Economics* 11 (4): 240–67.

Deserranno, Erika, Miri Stryjan, and Munshi Sulaiman. 2019. "Replication data for: Leader Selection and Service Delivery in Community Groups: Experimental Evidence from Uganda." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, Vol. 4, edited by T. Paul Schultz and John A. Strauss, 3895–3962. Amsterdam: Elsevier.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra. 2018. "Deep Neural Networks for Estimation and Inference." Unpublished.

Fogarty, Colin B. 2018a. "On Mitigating the Analytical Limitations of Finely Stratified Experiments." *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) 80 (5): 1035–56.

Fogarty, Colin B. 2018b. "Regression-Assisted Inference for the Average Treatment Effect in Paired Experiments." *Biometrika* 105 (4): 994–1000.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Gerber, Alan, Mitchell Hoffman, John Morgan, and Collin Raymond. 2020. "One in a Million: Field Experiments on Perceived Closeness of the Election and Voter Turnout." *American Economic Journal: Applied Economics* 12 (3): 287–325.

Gerber, Alan, Mitchell Hoffman, John Morgan, and Collin Raymond. 2020. "Data for: One in a Million: Field Experiments on Perceived Closeness of the Election and Voter Turnout." *American Economic Journal: Applied Economics* 12 (3): 287–325.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton: Princeton University Press.

Greevy, Robert, Bo Lu, Jeffery H. Silber, and Paul Rosenbaum. 2004. "Optimal Multivariate Matching before Randomization." Biostatistics 5 (2): 263–75.

Györfi, László, Michael Kohler, Adam Krzyzak, and Harro Walk. 2002. *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer.

Herskowitz, Sylvan. 2021. "Gambling, Saving, and Lumpy Liquidity Needs." *American Economic Journal: Applied Economics* 13 (1): 72–104.

Herskowitz, Sylvan, and International Food Policy Research Institute. 2020. "Replication Data for: Gambling, Saving, and Lumpy Liquidity Needs." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Himmler, Oliver, Robert Jäckle, and Philipp Weinschenk. 2019. "Soft Commitments, Reminders, and Academic Performance." *American Economic Journal: Applied Economics* 11 (2): 114–42.

Himmler, Oliver, Robert Jäckle, and Philipp Weinschenk. 2019. "Data for: Soft Commitments, Reminders, and Academic Performance." *American Economic Journal: Applied Economics* 11(2): 114–142.

Hsu, Henry, and Peter A. Lachenbruch. 2007. "Paired *t* Test." In *Encyclopedia of Biostatistics*, edited by Peter Armitage and Theodore Colton.. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/0470011815.b2a15112.

Imai, Kosuke. 2008. "Variance Identification and Efficiency Analysis in Randomized Experiments under the Matched-Pair Design." *Statistics in Medicine* 27 (24): 4857–73.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Lafortune, Jeanne, Julio Riutort, and Jose Tessada. 2018. "Role Models or Individual Consulting: The Impact of Personalizing Micro-Entrepreneurship Training." American Economic Journal: Applied Economics 10 (4): 222–45.

Lafortune, Jeanne, Julio Riutort, and Jose Tessada. 2019. "Replication data for: Role Models or Individual Consulting: The Impact of Personalizing Micro-entrepreneurship Training." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–1102.

Lee, Jean N., Jonathan Morduch, Saravana Ravindran, Abu Shonchoy, and Hassan Zaman. 2021. "Poverty and Migration in the Digital Age: Experimental Evidence on Mobile Banking in Bangladesh." *American Economic Journal: Applied Economics* 13 (1): 38–71.

Lee, Jean N., Jonathan Morduch, Saravana Ravindran, Abu Shonchoy, and Hassan Zaman. 2021. "Data and Code for: Poverty and Migration in the Digital Age: Experimen- tal Evidence on Mobile Banking in Bangladesh." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Li, Qi, and Jeffrey S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *Annals of Applied Statistics* 7 (1): 295–318.

Morgan, Kari Lock, and Donald B. Rubin. 2012. "Rerandomization to Improve Covariate Balance in Experiments." *Annals of Statistics* 40 (2): 1263–82.

Morgan, Kari Lock, and Donald B. Rubin. 2015. "Rerandomization to Balance Tiers of Covariates." *Journal of the American Statistical Association* 110 (512): 1412–21.

Moses, Lincoln E. 2006. "Matched Pairs *t*-Tests." In *Encyclopedia of Statistical Sciences*. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781118445112.stat01660.

Newey, Whitney K. 1997. "Convergence Rates and Asymptotic Normality for Series Estimators." *Journal of Econometrics* 79 (1): 147–68.

Riach, Peter A., and Judith Rich. 2002. "Field Experiments of Discrimination in the Market Place*." *Economic Journal* 112 (483): F480–518.

Steinwart, Ingo, and Andreas Christmann. 2008. *Support Vector Machines*. New York: Springer.

Tabord-Meehan, Max. Forthcoming. "Stratification Trees for Adaptive Randomization in Randomized Controlled Trials." *Review of Economic Studies.*

Wager, Stefan, and Guenther Walther. 2015. "Adaptive Concentration of Regression Trees, with Application to Random Forests." Unpublished.

White, Halbert. 1990. "Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings." *Neural Networks* 3 (5): 535–49.