The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation*

Kosuke Imai† Gary King‡ Clayton Nall§

First Draft: July 17, 2007
This Draft: January 18, 2008

Abstract

A basic feature of many field experiments is that investigators are only able to randomize clusters of individuals — such as households, communities, firms, medical practices, schools, or classrooms — even when the individual is the unit of interest. To recoup some of the resulting efficiency loss, many studies pair similar clusters and randomize treatment within pairs. Other studies (including almost all published political science field experiments) avoid pairing, in part because some prominent methodological articles claim to have identified serious problems with this "matched-pair cluster-randomized" design. We prove that all such claims about problems with this design are unfounded. We then show that the estimator for matched-pair designs favored in the literature is appropriate only in situations where matching is not needed. To address this problem without modeling assumptions, we generalize Neyman's (1923) approach and propose a simple new estimator with much improved statistical properties. We also introduce methods to cope with individual-level noncompliance, which most existing approaches assume away. We show that from the perspective of, among other things, bias, efficiency, power, or robustness, and in large samples or small, pairing should be used in cluster-randomized experiments whenever feasible; failing to do so

*Our proposed methods can be implemented using an R package, *experiment* (Imai, 2007), which is available for download at the Comprehensive R Archive Network (http://cran.r-project.org). We thank Kevin Arceneaux, Jake Bowers, Paula Diehr, Ben Hansen, Jennifer Hill, and Dylan Small for helpful comments, and Neil Klar and Allan Donner for detailed suggestions and gracious extended conversations. For research support, our thanks go to the National Institute of Public Health of Mexico, the Mexican Ministry of Health, the National Science Foundation (SES-0550873), the Princeton University Committee on Research in the Humanities and Social Sciences, and the Institute for Quantitative Social Science at Harvard.

†Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6610, Fax: 973–556–1929, kimai@princeton.edu, http://imai.princeton.edu

‡David Florence Professor of Government, Harvard University (Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge MA 02138; http://GKing.Harvard.edu, King@Harvard.edu, (617) 495-2027).

§Ph.D. Candidate, Department of Government, Harvard University (Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge MA 02138); nall@fas.harvard.edu.

is equivalent to discarding a considerable fraction of one's data. We develop these techniques in the context of a randomized evaluation we are conducting of the Mexican Universal Health Insurance Program.

Keywords: causal inference, community intervention trials, field experiments, group-randomized trials, place-randomized trials, health policy, matched-pair design, noncompliance, power.

1 Introduction

For political, ethical, or administrative reasons, researchers conducting field experiments are often unable to randomize treatment assignment to individuals and so instead randomize treatments to clusters of individuals (Murray, 1998; Donner and Klar, 2000; Raudenbush *et al.*, 2007). For example, 19 (68%) of the 28 field experiments we found published in major political science journals since 2000 randomized households, precincts, city-blocks, or villages even though individual voters were the inferential target (e.g., Arceneaux, 2005); in public health and medicine, where "the number of trials reporting a cluster design has risen exponentially since 1997" (Campbell, 2004), randomization occurs at the level of health clinics, physicians, or other administrative and geographical units even though individuals are the units of interest (e.g., Sommer *et al.*, 1986; Varnell *et al.*, 2004); and numerous education researchers randomize schools, classrooms, or teachers instead of students (e.g., Angrist and Lavy, 2002).

Since statistical efficiency drops when randomizing clusters of individuals instead of individuals themselves (Cornfield, 1978), many researchers attempt to recoup a portion of this lost efficiency by pairing clusters, based on the similarity of available background characteristics, before randomly assigning one cluster within each pair to receive the treatment assignment (e.g., Ball and Bogatz, 1972; Gail *et al.*, 1992; Hill *et al.*, 1999). Since matching prior to random treatment assignment can greatly improve the efficiency of causal effect estimation (Box *et al.*, 1978; Greevy *et al.*, 2004), and matching in pairs can be substantially more efficient than matching in larger blocks, *matched-pair, cluster-randomization* would appear to be a very attractive design for field experiments (Author cite redacted). The design is especially useful for public policy experiments since, when used properly, it can be robust to some interventions by politicians and others that have ruined many policy evaluations, such as when office-holders unexpectedly arrange program benefits for their constituents in some control group clusters (King *et al.*, 2007).

Unfortunately, despite its apparent benefits and common usage, this experimental design has an uncertain status within parts of the methodological literature. For example, Klar and Donner (1997) claim that certain "analytic limitations" make paired randomization, or at least the existing methods available to analyze data from this design, inappropriate. These claimed lim-

itations, which include "the restriction of prediction models to cluster-level baseline risk factors (for example, cluster size), the inability to test for homogeneity of ... [causal effects across clusters], and difficulties in estimating the intracluster correlation coefficient, a measure of similarity among cluster members" (Klar and Donner 1997, p.1754; see also Donner and Klar 2004), are also made by many other researchers and even various clinical trial standards organizations (e.g., Feng *et al.*, 2001; Medical Research Council, 2002). Another reason to worry about the matched-pair cluster randomized design is that, to our knowledge, there exists no published formal evaluation of the statistical properties of the estimator recommended in the methodological literature. In fact, the literature does not even offer a precise definition of the causal effect to be estimated under matched-pair and other cluster-randomized designs (except implicitly under model-based approaches, although in the literature the models to be used are not always made explicit). Finally, in a widely-cited article, Martin *et al.* (1993) claim that in very small sample sizes, pairing can reduce statistical power.

In this paper, we prove that each of the claims regarding analytical limitations of the matched-pair design is incorrect. We also show that the power calculations leading Martin *et al.* (1993) to recommend against matched pairs in small samples is dependent on an assumption of equal cluster sizes that vitiates the advantage of pair matching; we show in real data that the assumption is wrong and without it pair matching typically improves both efficiency and power a great deal even in extremely small sample sizes (such as with only three matched pairs). In fact, because we show that the efficiency gain of matched pair designs depends on the correlation of cluster means weighted by cluster size, the advantage can be much larger than the unweighted correlations examined in the literature seems to indicate, even when cluster sizes are independent of the outcome.

Furthermore, we prove that the standard causal effect estimator for matched-pair designs in the literature depends on unrealistic assumptions, such as homogeneity across clusters, that apply only when matching is not needed to begin with. To ameliorate this situation, we define the causal effect quantities of interest, offer new simple nonparametric design-based estimators and standard errors, and prove that they have more desirable statistical properties. We then extend our estimator to situations with individual-level noncompliance, which all but a few prior analyses

have ignored and yet is a basic feature of most cluster-randomized experiments. We also show that under the matched-pair cluster-randomized design, researchers only need to assume no interference from one matched-pair to the next, as compared to unit-level randomization designs that have to satisfy the much more dubious assumption of no influence between an individual and his or her family members, neighbors, friends, and coworkers. With the results and new estimators offered here, the ambiguity in the literature vanishes: pair matching should be used whenever feasible.

2 Evaluation of the Mexican Universal Health Insurance Program

As a running example, we introduce a randomized evaluation we are conducting of *Seguro Popular de Salud* (SPS, the Mexican universal health insurance program). The program's "aim is to provide social protection in health to the 50 million uninsured Mexicans" (Frenk *et al.*, 2003, p.1667), constituting about half the population, through one of the largest health policy reforms in any country in the last twenty years (King *et al.*, 2007). The government intends to spend an additional one percent of GDP on health compared to 2002 when the program is fully rolled out.

To take advantage of SPS services, individuals must formally affiliate with the program. Mexican states apply to the federal government for permission and funds to encourage certain numbers of people in chosen areas to affiliate. The federal government approves these requests only when local health clinics are brought up to federal standards, including sufficient hours of operation, drugs, medical equipment, medical personnel, etc. When an area is approved for affiliation, individuals who affiliate are supposed to receive preventative and regular medical care, as well as pharmaceuticals and medical procedures, all without cost.

A key goal of the program, and the main goal of the evaluation at this early stage, is to reduce out-of-pocket expenditures for health, and especially catastrophic health expenditures that may account for a large fraction of a family's disposable income.

Because it will take several years for local health clinics and hospitals to be built and brought up to federal standards, and because the money available to spend on SPS is limited at any one time, the program needed to be rolled out in stages. As such, a randomized evaluation was

possible. A set of 12,284 geographic "health clusters" that tile the country were defined. Each health cluster includes the local health clinic or hospital and the catchment area around it, such that the travel time to the clinic, using transportation methods available to locals, is less than a day. One hundred of these clusters were selected in negotiation with the Mexican government for which randomization and evaluation were acceptable. Clusters were then paired based on variables measuring population, socio-demographics, poverty, education, and health infrastructure. Finally, one "treatment" cluster was randomly chosen within each of the fifty pairs to receive the benefits of SPS and encouragement for people to affiliate. (For expository reasons, we assume perfect compliance with affiliation encouragement until Section 6.) The remaining cluster will receive SPS at some future date but is the "control" cluster for our experiment. For details of the design, see King *et al.* (2007).

Our data include aggregate characteristics of the health clusters, collected from the census and other government sources, as well as two surveys of about 32,000 individuals randomly selected within each of the 100 clusters. One baseline survey was conducted at the time of randomization and another, which we use in this paper, was conducted 10 months later. Follow-up surveys are planned for later on, when the health effects of SPS will be more likely, as well.

3 Matched-Pair, Cluster-Randomized Experiments

We now introduce matched-pair cluster randomized experiments, including the theories of inference commonly applied (Section 3.1), the formal definitions, notation, and assumptions used in (Section 3.2), and the quantities of interest typically sought (Section 3.3).

3.1 Theories of Inference

Two theories of statistical inference have been applied to the analysis of matched-pair cluster-randomized data: model-based and permutation-based analyses. We describe these here, followed by a description of the design-based inference from which our work is derived.

The first existing theory of inference applied to matched-pair designs is model-based, using generalized mixed-effects, generalized estimating equations, or multi-level models (Feng *et al.*,

2001). Most of these approaches work only if the modeling assumptions are correct and many rely on asymptotic approximations. Model-based and model-assisted approaches have proven to be powerful in other areas, but are not in the spirit of experimental work that goes to great lengths and expense to avoid these types of assumptions in the first place.

The second theory of inference commonly applied to this type of design is Fisher (1935)'s permutation-based approach which seeks to construct nonparametric hypothesis tests with inference based only on the random assignment of treatment. Permutation inference requires no models or large-sample approximations, but focuses on sample causal effects and typically depends on the unrealistic assumption of constant treatment effects across clusters when inferring the magnitude of causal effects. (See also Gail *et al.* (1996) and Braun and Feng (2001) who combine permutation inference with parametric models to estimate average treatment effects, and Small *et al.* (In-press) who entertain alternative assumptions for quantile effects.)

The methods developed in this paper use Neyman (1923)'s approach to statistical inference, which is well known but has not before been attempted for matched-pair cluster randomized experiments. Like Fisher, Neyman's approach is also design (or "randomization") based and nonparametric, but it goes further to avoid constant treatment effect and homoskedasticity assumptions and can provide valid inferences about both sample and population average treatment effects. The estimators we derive under this approach also turn out to be simple to understand and considerably easier to compute (requiring only weighted means and no numerical optimization, simulation, or iterations) compared with the above approaches.

3.2 Formal Design Definition, Notation, and Assumptions

Consider a matched-pair, cluster-randomized experiment where a total of $2m$ observed clusters (or groups) are paired, based on a certain known function of the observed cluster characteristics, prior to the randomization of a binary treatment. We assume the jth cluster in the kth pair contains observed n_{jk} units, where $j = 1, 2$, and thus the total number of observed units is equal to $n = \sum_{k=1}^{m}(n_{1k} + n_{2k})$. The two clusters within each matched pair are randomly ordered.

Under the matched-pair design, simple randomization of an indicator variable, Z_k for $k = 1, 2, \ldots, m$, is conducted independently across the resulting m pairs. For a matched-pair with

$Z_k = 1$, the first cluster within pair k is treated (in our case, assigned encouragement to affiliate with SPS), and the second cluster is assigned control. In contrast, for a matched-pair with $Z_k = 0$, the first cluster of the pair is the control whereas the second is treated. Thus, if we use T_{jk} to represent the treatment indicator variable for the jth cluster in the kth matched-pair, then $T_{1k} = Z_k$ and $T_{2k} = 1 - Z_k$. For the moment, we consider an intention-to-treat (ITT) analysis, or equivalently assume that all units in the same cluster receive the same treatment so that everyone encouraged to affiliate does, and everyone who is not encouraged does not. We extend this methodology to unit-level noncompliance in Section 6.

We denote $Y_{ijk}(T_{jk})$ as the potential outcomes under the treatment ($T_{jk} = 1$) and control ($T_{jk} = 0$) conditions for the ith unit in the jth cluster of the kth matched-pair. The observed outcome variable can then be written as $Y_{ijk} = T_{jk}Y_{ijk}(1) + (1 - T_{jk})Y_{ijk}(0)$. Finally, the order of clusters within each pair is assumed to be randomized so that the population distribution of a pair $(Y_{i1k}(1), Y_{i1k}(0))$ is identical to $(Y_{i2k}(1), Y_{i2k}(0))$ (though this equality may not hold in sample).

A defining feature of cluster-randomized experiments is that the potential outcomes for the ith unit in the jth cluster of the kth matched-pair are a function of the cluster-level randomized treatment variable, T_{jk}, rather than its unit-level treatment counterpart. Similarly, the unit-level causal effect, $Y_{ijk}(1) - Y_{ijk}(0)$, represents the difference between two unit-level potential outcomes that are the functions of the cluster-level treatment variable. Thus, in cluster-randomized trials, the usual assumption of no interference (Cox, 1958; Rubin, 1990) needs to be made *only* at the cluster level. Moreover, in the case of the matched-pair design, assuming no interference only between pairs of clusters is sufficient. This advantage of matched-pair designs can be substantial if contagion or social influence is present at the individual level, where for example individuals may affect the behavior of neighbors or friends, but such interference does not exist across clusters. In this way, matched-pair designs greatly reduce the assumptions necessary compared to unit-level randomization designs. We formalize this assumption as follows.

Assumption 1 (No Interference between Matched-Pairs) *Let $Y_{ijk}(\mathbf{T})$ be the potential outcomes for the ith unit in the jth cluster of the kth matched-pair where $\mathbf{T}$ is a $(m \times 2)$ matrix whose (j, k) element is T_{jk}. Then, if $T_{jk} = T'_{jk}$, then $Y_{ijk}(\mathbf{T}) = Y_{ijk}(\mathbf{T}')$.*

The assumption allows us to write $Y_{ijk}(T_{jk})$ rather than $Y_{ijk}(\mathbf{T})$. Since $T_{1k} = Z_k$ and $T_{2k} =$

$1 - Z_k$, it is clear that $Y_{ijk}(T_{jk})$ only depends on Z_k. Given that in many social experiments the assumption of no interference between units is highly unrealistic (Sobel, 2006), cluster-randomized trials offer an attractive alternative in the field experiments where social interactions among units are expected to occur. Indeed, in our Mexican experiment, this assumption is reasonable because most of the clusters in our experiment were not geographically contiguous to each other, and travel between these areas is often not easy. However, especially in small villages, unit-level no interference assumptions would have been implausible.

Finally, we formalize the cluster-level treatment assignment as follows.

Assumption 2 (Cluster Randomization under Matched-Pair Design) *The potential outcomes are independent of the randomization indicator variable:* $(Y_{ijk}(1), Y_{ijk}(0)) \perp\!\!\!\perp Z_k$, *for all* $i, j,$ *and* k. *Also,* Z_k *is independent across matched-pairs, and* $\Pr(Z_k) = 1/2$ *for all* k.

The assumption also implies $(Y_{ijk}(1), Y_{ijk}(0)) \perp\!\!\!\perp T_{jk}$ since T_{jk} is a function of Z_k.

3.3 Quantities of Interest

We now offer the precise definitions of the average causal effects of interest in matched-pair cluster-randomized experiments which, to our knowledge, have not been formally defined. At least two types of each of four distinct quantities may be of interest in these experiments. We begin with the four quantities, which define the target population, and then discuss the two types, which clarify the role of interference or contagion between treatment and control clusters in each pair. Section 6 introduces additional quantities of interest when individual-level noncompliance exists. (All the quantities below are based on causal effects defined as grouped individual-level phenomena; we discuss causal quantities defined at the cluster level in Section 4.4.)

3.3.1 Target Population Quantities

Table 1 offers an overview of the four target population causal effects. All four quantities represent the causal treatment effect (the potential outcome under treatment minus the potential outcome under control) averaged over different sets of units.

The first quantity is the *sample average treatment effect* (SATE or ψ_S) which we define as an

Quantities		Clusters	Units within Clusters	Inferential Target
ψ_S	SATE	Observed	Observed	Observed sample
ψ_C	CATE	Observed	Sampled	Population within observed clusters
ψ_U	UATE	Sampled	Observed	Observable units within the population of clusters
ψ_P	PATE	Sampled	Sampled	Population

Table 1: Quantities of Interest: For each causal effect, this table lists whether clusters and units within clusters are treated as observed and fixed or instead as a sample from a larger population. The resulting inferential target is also given.

average over the set of all units in the observed sample (which we denote as $\mathcal{S}$):

$$\psi_S \equiv \mathbb{E}_{\mathcal{S}}(Y(1) - Y(0)) = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{2}\sum_{i=1}^{n_{jk}}(Y_{ijk}(1) - Y_{ijk}(0)), \tag{1}$$

where the sums go over matched-pairs, the two clusters within each pair, and the units within each cluster, respectively.

The second quantity treats the observed clusters as fixed (and not necessarily representative of some other population) and the units within clusters as randomly sampled from the (finite) population of units within each cluster. This gives the *cluster average treatment effect* (CATE or ψ_C):

$$\psi_C \equiv \mathbb{E}_{\mathcal{C}}(Y(1) - (0)) = \frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{2}\sum_{i=1}^{N_{jk}}(Y_{ijk}(1) - Y_{ijk}(0)), \tag{2}$$

where the expectation is taken over the the set $\mathcal{C}$ which contains all observed units within the sample clusters, N_{jk} is the known (and finite) population cluster size, and $N \equiv \sum_{k=1}^{m}(N_{1k} + N_{2k})$. Throughout, we assume simple random sampling within each cluster for the sake of simplicity, but other random sampling procedures can easily be accommodated by applying appropriate unit-level weights. Thus, the only difference between SATE and CATE is whether each unit within clusters is treated as fixed or as randomly drawn based on some known sampling mechanism.

A third quantity treats the clusters as randomly sampled from a larger population of clusters, but the units within the sampled clusters are treated as fixed and not necessarily randomly sampled. The inferential target is the set $\mathcal{U}$, which includes all units in the population of clusters that would be observed if its cluster were in the observed sample. This is what we call the *unit average*

treatment effect (UATE or ψ_U) and is defined as:

$$\psi_U \equiv \mathbb{E}_{\mathcal{U}}(Y(1) - Y(0)). \tag{3}$$

The final quantity of interest is the *population average treatment effect* (PATE or ψ_P), which is defined as:

$$\psi_P \equiv \mathbb{E}_{\mathcal{P}}(Y(1) - Y(0)), \tag{4}$$

where the expectation is taken over the entire population $\mathcal{P}$ — that is, the population of units within the population of clusters.

For simplicity throughout, we assume an infinite population of clusters, but this is easily extended to finite populations at some cost in additional notation. We also assume that all quantities of interest are at the individual level; studies seeking causal effects of treatments on variables measured at the cluster level can use methods designed for unit-level randomization experiments, and so we do not discuss those further here.

3.3.2 Interference

We now discuss experimental interference (1) among individuals in the same cluster, (2) between clusters in different pairs, and most centrally for this section (3) interference between the treatment and control clusters within the same pair.

First, when interference among individuals within a cluster exists, the potential outcomes of one person (or unit) within a cluster may be different, depending on whether or not others nearby were (also) assigned to receive treatment. This can happen in a health policy study through disease contagion, social pressure, or family members or friends helping each other financially. This type of interference is expected to some degree and not assumed away in cluster-randomized trials where all individuals within a cluster receive the same treatment assignment. Thus, all four quantities in cluster-randomized experiments allow for interference among individuals within a cluster. The degree of interference among individuals within a cluster is treated as a consequence of treatment and thus part of the outcome under study. Understanding the effect of some treatment on individuals independent of and isolated from other individuals is best left to studies where individual randomization is possible and interference among individuals can be eliminated.

Second, Assumption 1 clarifies that matched-pair cluster randomized experiment assumes no interference between clusters in different pairs. We continue to maintain this assumption as Sobel (2006) demonstrates that without it even the definition of a causal effect is not straightforward.

A third concern in matched-pair cluster-randomized studies is interference between treatment and control clusters within the same pair. For example, if one cluster is assigned to SPS and receives hospitals, doctors, pharmaceuticals, and financial protection from catastrophic expenditures, it can be that a neighboring cluster assigned control in the same pair will contain some individuals who are envious and perhaps even depressed as a consequence. This type of interference between clusters within a pair can be thought of in two ways. In the first, which we call *no-interference*, we define our chosen causal effect (SATE, CATE, UATE, or PATE) so that the treatment assignment in one cluster has no effect on the potential outcomes of units in the control cluster. In the second, which we call *with-interference*, we define our chosen causal effect to have whatever level of interference happens to exist in the data in our particular experiment.

We do not complicate our notation by defining symbols and acronyms to accommodate the interference distinction as we do with our target population quantities. The issue is that estimating the no-interference version of SATE, CATE, UATE, or PATE when there exists interference in the data turns out to be difficult, and feasible only with assumption-laden estimators. In contrast, the with-interference version is easy to estimate since it accepts whatever level of non-interference one's data happens to dish out.

Of course, having a quantity that is easy to estimate is not a satisfactory substitute for having an estimate of the quantity that is genuinely of interest. The best way around this dilemma is to use these facts from the outset in designing the experiment. For example, in determining the matches for a matched-pair study, we can select clusters to pair that are not contiguous. We can also select pairs that are not contiguous to other pairs. Following rules like this reduces the level of interference among clusters and makes the difference between the no-interference and with-interference quantities considerably smaller. We can then use the simpler estimators not requiring corrections for interference and get estimates closer to the desired quantity.

Any study seeking to make causal inferences should identify the precise causal effect that is the target of the inference. For matched-pair cluster-randomized experimental designs, prior literature

has left the causal quantity of interest vague or unstated; the choice can now be made from the menu offered in this section.

3.3.3 SPS Evaluation

In the SPS evaluation, we would very much like to make an inference about PATE for all of Mexico, but our health clusters were not (and could not due to political and administrative constraints be) randomly selected. This means that, like most medical experiments, any method applied to our data to estimate PATE will be dependent on assumptions about the selection process. One thing we can do is to infer to all the people within the population of clusters like those sampled. Alternatively, a more conservative approach would be to try to estimate one of the other quantities. CATE or SATE are straightforward possibilities, and CATE is probably most apt in this case, since individuals within clusters were randomly selected, and both quantities condition on the clusters we observe. From a public policy perspective, UATE may be a reasonable target quantity, where we try to infer to the individuals who would be sampled in all the health clusters in Mexico that are similar to our observed clusters, and from which our clusters could plausibly have been randomly drawn.

Many of the clusters in our experiment are not contiguous to each other, and many of these are coherent self-contained communities, and so we do not expect a great deal of direct interference between them (within or across pairs). However, nearby clusters not in our experiment could possibly exert influence over those we observe.

4 Estimators

In this section, we define our estimators and those in the literature, and then prove the conditions under which each is unbiased. We also derive the variance of the estimators and offer a valid method of computing standard errors without modeling assumptions. As no formal justification has been given of the existing estimator in the literature, we derive an individual-level model that gives rise to it; this model also reveals the onerous assumptions that data need to meet in order for this estimator to give valid estimates. In contrast, the estimators we introduce involve making

as few assumptions as feasible aside from those based on the design of the experiment.

4.1 Definitions

Imbens (2004, p.6), who first clearly defined the population and sample "average treatment effect" (ATE) in unit-level randomization studies, comments that "a good estimator for one ATE is automatically a good estimator for the other." This is an important insight but of course was not meant to apply to any design other than a unit-level study with simple random sampling. We show here that the insight also does not apply to cluster-randomization studies. In fact, as we showed in Section 3.3, cluster-randomized studies suggest at least four causal quantities of potential interest rather than the usual two in unit-randomized studies.

All point estimators we discuss for the with-interference version of the four quantities of interest introduced in Section 3.3 have the same general mathematical form. They are each the weighted average of within-pair differences between the treated and control clusters, but with different weights. As the weights are crucial for causal inference in cluster randomized experiments, we introduce the following general notation for an estimator:

$$\hat{\psi}(w_k) \equiv \frac{1}{\sum_{k=1}^{m} w_k} \sum_{k=1}^{m} w_k \left\{ Z_k \left(\frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}} - \frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}} \right) + (1 - Z_k) \left(\frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}} - \frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}} \right) \right\}, \tag{5}$$

where the choice of a weight for the kth pair of clusters, denoted by w_k, defines a specific estimator.

The estimator most commonly recommended in the methodological literature is based on a weight using the harmonic mean of sample cluster sizes, which using our notation can be written as $\hat{\psi}(n_{1k}n_{2k}/(n_{1k} + n_{2k}))$ (see e.g., Donner, 1987; Donner and Donald, 1987; Donner and Klar, 1993; Hayes and Bennett, 1999; Bloom, 2006; Raudenbush, 1997; Turner *et al.*, 2007). As we show below, however, this estimator relies on assumptions unlikely to be met in practice, and has undesirable properties if those assumptions are not met. We also prove that the associated variance estimator given in the literature is generally biased unless those restrictive assumptions are met.

To address these problems, we introduce new estimators, based on the arithmetic mean of sample or population cluster sizes, that do not require such modeling assumptions. As shown

	SATE	CATE	UATE	PATE
Point estimator	$\hat{\psi}(n_{1k} + n_{2k})$	$\hat{\psi}(N_{1k} + N_{2k})$	$\hat{\psi}(n_{1k} + n_{2k})$	$\hat{\psi}(N_{1k} + N_{2k})$
Variance	$\text{Var}_a(\hat{\psi})$	$\text{Var}_{au}(\hat{\psi})$	$\text{Var}_{ap}(\hat{\psi})$	$\text{Var}_{aup}(\hat{\psi})$
Identified	no	no	YES	YES

Table 2: Point estimators and variances for the four causal quantities of interest

in Table 2, $\hat{\psi}(n_{1k} + n_{2k})$ is the point estimator we offer for both SATE and UATE, whereas $\hat{\psi}(N_{1k} + N_{2k})$ is the point estimator for both CATE and PATE. This is intuitive, as SATE and UATE are based on those units (which would be) sampled in a cluster whereas CATE and PATE are based on the population of units within clusters. Below, we prove that these estimators have more desirable properties than the existing estimator. Note that our estimator for SATE and UATE differ from the existing estimator based on harmonic mean weights unless the sample cluster sizes within each matched-pair are equal ($n_{1k} = n_{2k}$ for all $k = 1, \ldots, m$), which is often inapplicable in practice.

Table 2 also gives an overview of the variances and their estimation. UATE and PATE have variances that are identifiable, the exact expression for which we give below. SATE and CATE have unidentifiable variances, and so we offer their best possible or sharp upper (and lower) bound, which is identifiable and leads to a confidence interval that is conservative. The variance estimators we provide differ from the existing variance estimator even when sample cluster sizes are matched exactly and yet does not rely on the modeling assumptions required by the latter. In fact, we show that our variance estimator is unbiased regardless of weights used, whereas the existing estimator is generally biased.

SPS Evaluation. Estimates from UATE and PATE (or equivalently SATE and CATE) will differ depending on how sample and population sizes differ across clusters. In the SPS evaluation, we sampled individuals randomly from each selected cluster (one per household according to the Kish tables) until either 380 individuals were sampled from a cluster or all households were sampled. Thus, the sample weights used for SATE and UATE were neither constant across clusters nor proportional to the population sizes used for CATE and PATE. In our data, for 67 key outcome variables, we found that PATE ranged from twice as large as UATE to about half its size, although

this range can be very different for other data sets. We give some numerical examples from SPS evaluation in Section 7.

4.2 Bias

For expository reasons, we first evaluate the bias of our proposed estimator followed by the bias of the method commonly used in the literature. We also discuss the difficulties with a different alternative.

4.2.1 Proposed Estimator Based on Arithmetic Mean Weights

To evaluate the bias of our proposed estimator for various quantities of interest defined in Section 3.3, we first focus on the estimation of SATE. This allows us, following Neyman (1923), to use the randomized treatment assignment mechanism as the sole basis for statistical inference (Author cite redacted).Under this framework, the potential outcomes are assumed fixed, but possibly unknown, quantities. We begin by rewriting $\hat{\psi}(n_{1k} + n_{2k})$ using potential outcome notation:

$$\begin{aligned}\hat{\psi}(n_{1k} + n_{2k}) &= \frac{1}{n}\sum_{k=1}^{m}(n_{1k} + n_{2k})\left\{Z_k\left(\frac{\sum_{i=1}^{n_{1k}} Y_{i1k}(1)}{n_{1k}} - \frac{\sum_{i=1}^{n_{2k}} Y_{i2k}(0)}{n_{2k}}\right)\right. \\ &\qquad \left. +(1 - Z_k)\left(\frac{\sum_{i=1}^{n_{2k}} Y_{i2k}(1)}{n_{2k}} - \frac{\sum_{i=1}^{n_{1k}} Y_{i1k}(0)}{n_{1k}}\right)\right\}.\end{aligned}$$

Then, taking the expectation with respect to Z_k yields:

$$E_a\{\hat{\psi}(n_{1k} + n_{2k})\} - \psi_S = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{2}\left\{\left(\frac{n_{1k} + n_{2k}}{2} - n_{jk}\right)\sum_{i=1}^{n_{jk}}\frac{Y_{ijk}(1) - Y_{ijk}(0)}{n_{jk}}\right\}, \tag{6}$$

where the expectation is taken with respect to the randomization of treatment assignment which we indicate by the subscript "a".

Although the bias does not generally equal zero, there are two common conditions under each of which it can be eliminated. These two conditions motivate our choice of weights ($w_k = n_{1k} + n_{2k}$). First, when cluster sizes are equal within each matched-pair (i.e., $n_{1k} = n_{2k}$ for all k), the bias is always zero. This implies that researchers may wish to form pairs of clusters, at least partially, based on their sample size if the SATE is the estimand. Second, $\hat{\psi}(n_{1k} + n_{2k})$ is also unbiased if matching is effective, so that the within-cluster SATEs are identical for each matched-pair (i.e., $\sum_{i=1}^{n_{1k}}(Y_{i1k}(1) - Y_{i1k}(0))/n_{1k} = \sum_{i=1}^{n_{2k}}(Y_{i2k}(1) - Y_{i2k}(0))/n_{2k}$ for all k). In contrast, bias may remain if cluster sizes are poorly matched

and within each pair cluster sizes are strongly associated with the cluster-specific SATEs. So roughly speaking, if cluster sizes and important confounders are matched well so that pre-randomization matching accomplishes the purpose for which it was designed, this estimator will be approximately unbiased.

A similar bias expression can be derived for our CATE estimator, $\hat{\psi}(N_{1k}+N_{2k})$, where the weights are now based on the arithmetic mean of the population cluster sizes rather than their sample counterparts. Thus, a calculation analogous to the one above yields the following bias expression:

$$E_{au}(\hat{\psi}(N_{1k}+N_{2k}))-\psi_C \;=\; \frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{2}\left\{\left(\frac{N_{1k}+N_{2k}}{2}-N_{jk}\right)E_u(Y_{ijk}(1)-Y_{ijk}(0))\right\}, \qquad (7)$$

where subscript "au" means that the expectation is taken with respect to random treatment assignment and the simple random sampling of units within each cluster. The conditions under which this bias disappears are analogous to the ones we obtained for SATE. For example, if matching is effective so that the cluster-specific average causal effects, i.e., $E_u[Y_{ijk}(1)-Y_{ijk}(0)]$, are constant across clusters within each matched-pair, then the bias is zero. The bias also vanishes if the population cluster sizes are identical within each matched-pair, i.e., $N_{1k}=N_{2k}$ for all k.

Finally, the bias for UATE and PATE can be obtained by taking the expectation of the bias for SATE and CATE, respectively. This expectation is defined with respect to simple random sampling of pairs of clusters. If the within-cluster sample (population) average treatment effects are uncorrelated with cluster sizes within each matched-pair, then the bias for the estimation of UATE (PATE) is zero, regardless of whether one can match exactly on cluster sizes. In general, however, cluster sizes may be correlated with the size of average treatment effects. In such cases, the matching strategies to reduce the bias for the estimation of SATE (CATE) also work for the estimation of UATE (PATE). That is, matched-pairs of clusters should be constructed such that within each pair, cluster sizes and observed pre-treatment covariates (especially those which are strong predictors of the average treatment effects) are similar.

4.2.2 Existing Estimator Based on Harmonic Mean Weights

The weights most commonly discussed and recommended in the methodological literature are based on the harmonic mean of sample cluster sizes: $w_k = n_{1k}n_{2k}/(n_{1k}+n_{2k})$. As we have been unable to find any formal justification given for this choice in the literature, we derive one here in order to clarify the assumptions required for the estimator to work.

Modeling Assumptions. The harmonic mean-based estimator stems from the weighted one-sample t-test for the difference in means: $D_k \overset{\text{indep.}}{\sim} N(\mu, (w_k/\sum_{k'=1}^{m} w_{k'})^{-1}\sigma)$ for $k = 1, 2, \ldots, m$ where w_k is the known weight. In our context, D_k is the observed within-pair difference-in-means, i.e., $D_k \equiv Z_k D_k(1) + (1 - Z_k)D_k(0)$ where $D_k(1) \equiv \sum_{i=1}^{n_{1k}} Y_{i1k}(1)/n_{1k} - \sum_{i=1}^{n_{2k}} Y_{i2k}(0)/n_{2k}$ and $D_k(0) \equiv \sum_{i=1}^{n_{2k}} Y_{i2k}(1)/n_{2k} - \sum_{i=1}^{n_{1k}} Y_{i1k}(0)/n_{1k}$. It is well known that under this model, $\sum_{k=1}^{m} w_k D_k / \sum_{k'=1}^{m} w_{k'}$ is the uniformly minimum variance unbiased estimator. But where does this model come from and how is it justified in cluster-randomized trials?

As it turns out, we can show that the following model of unit-level potential outcomes give rise to the harmonic mean-based estimator: $Y_{ijk}(t) \overset{\text{i.i.d.}}{\sim} N(\mu_t, \tilde{\sigma})$ for $t = 0, 1$ where $\tilde{\sigma} = \sigma \sum_{k=1}^{m} w_k$ (note that $\sum_{k=1}^{m} w_k$ is a known constant since w_k is assumed to be known and fixed). The model is based on the following assumptions: (1) normality; (2) independent and identical distributions across units within each cluster as well as across clusters and pairs (constant means and variances within and across clusters and pairs); and (3) equal variances for the two potential outcomes. Although we focus on t-test here, for binary outcomes the suggested approach in the literature is also based on a homogeneity assumption where a common odds ratio is assumed across clusters (e.g., Donner and Donald, 1987; Donner and Hauck, 1989).

When this model holds, the harmonic mean-based estimator is minimum variance unbiased. However, heterogeneity among clusters almost always exists in field experiments, and only rare cluster-randomized experiments would meet these stringent conditions. And more importantly, the point of matching is to control for heterogeneity, and so an estimator that assumes homogeneity is unlikely to be of much use. In any event, we have already shown that these unrealistic assumptions, and all other modeling assumptions, are unnecessary for estimation of average treatment effects in matched-pair cluster-randomized experiments.

Finally, we also note that although many researchers claim that estimating the intracluster correlation coefficient is of central importance (see Section 4.3.4), an implication of Assumption (2)

Bias Calculation. The harmonic mean weight differs from our proposed weight at least in three ways. First, it gives more weight to matched pairs with well-matched cluster sizes than to pairs whose cluster sizes are unbalanced. That is, if we assume the sum of n_{1k} and n_{2k} is fixed, the harmonic mean is the largest when $n_{1k} = n_{2k}$ and becomes smaller as the difference between n_{1k} and n_{2k} increases. Second, and most importantly, unlike our proposed estimator, the use of this weight does not remove the bias

when within-cluster average treatment effects are identical within matched-pairs. This means that even when matching is effective, the bias may remain large. To see this point formally, we rewrite the bias of the harmonic mean estimator for SATE as:

$$E_a\left\{\hat{\psi}\left(\frac{n_{1k}n_{2k}}{n_{1k}+n_{2k}}\right)\right\} - \psi_S \\ = \sum_{k=1}^{m}\sum_{j=1}^{2}\left\{\left(\frac{n_{1k}n_{2k}}{2(n_{1k}+n_{2k})\sum_{l=1}^{m} n_{1l}n_{2l}/(n_{1l}+n_{2l})} - \frac{n_{jk}}{n}\right)\sum_{i=1}^{n_{jk}}\frac{Y_{ijk}(1)-Y_{ijk}(0)}{n_{jk}}\right\}.$$

The bias is generally not zero even when the cluster-specific SATEs are identical within each matched-pair. (The direction of the bias will depend on the data; see Section 4.3.6.) One condition under which $\hat{\psi}(n_{1k}n_{2k}/(n_{1k}+n_{2k}))$ is unbiased is with exact matching on sample cluster sizes (i.e., $n_{1k}=n_{2k}$ for all k). However, under this condition, this estimator coincides with our proposed estimator, $\hat{\psi}(n_{1k}+n_{2k})$. Finally, since the weight is based on sample cluster sizes, this estimator has little to do with the estimation of CATE and PATE.

4.2.3 An Unbiased But Not Invariant Alternative Estimator

While $\hat{\psi}(w_k)$ is in general biased regardless of the choice of weights, we have derived an unbiased estimator of SATE and UATE (but not the other causal effects because it does not incorporate cluster population weights) with a different form:

$$\hat{\phi}_1 \equiv \frac{2}{n}\sum_{k=1}^{m}\left\{Z_k\left(\sum_{i=1}^{n_{1k}}Y_{i1k} - \sum_{i=1}^{n_{2k}}Y_{i2k}\right) + (1-Z_k)\left(\sum_{i=1}^{n_{2k}}Y_{i2k} - \sum_{i=1}^{n_{1k}}Y_{i1k}\right)\right\}. \quad (8)$$

This estimator coincides with our proposed estimator, $\hat{\psi}(n_{1k}+n_{2k})$, if cluster sizes within each matched-pair are equal, i.e., $n_{1k}=n_{2k}$ for all k. Although $\hat{\phi}_1$ is an unbiased estimator, it has a highly unattractive feature in that it is not invariant to a constant shift of the outcome variable when cluster sizes vary within each matched-pair. That is, if we recode the outcome variable of all units by changing Y_{ijk} to $Y_{ijk}+\alpha$ for some constant α, then $\hat{\phi}_1$ will take a different value. This problem is serious when the number of matches is small and when cluster sizes differ considerably within and across matches.

One way to overcome the invariance problem is to slightly modify the above estimator to the following:

$$\hat{\phi}_2 \equiv \frac{\sum_{k=1}^{m} Z_k\sum_{i=1}^{n_{1k}}Y_{i1k} + (1-Z_k)\sum_{i=1}^{n_{2k}}Y_{i2k}}{\sum_{k=1}^{m} Z_k n_{1k} + (1-Z_k)n_{2k}} \\ - \frac{\sum_{k=1}^{m} Z_k\sum_{i=1}^{n_{2k}}Y_{i2k} + (1-Z_k)\sum_{i=1}^{n_{1k}}Y_{i1k}}{\sum_{k=1}^{m} Z_k n_{2k} + (1-Z_k)n_{1k}}. \quad (9)$$

However, since both numerators and denominators are functions of random variables, the exact calculation of bias is impossible within the design-based inferential framework. The Taylor series expansion yields the following approximate bias:

$$E_a(\hat{\phi}_2) - \psi_S \quad = \quad \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{2}\left\{\left(\frac{n_{1k}+n_{2k}}{2} - n_{jk}\right)\sum_{i=1}^{n_{jk}}\frac{Y_{ijk}(1)-Y_{ijk}(0)}{n}\right\}.$$

The comparison of this expression with the bias of $\hat{\psi}(n_{1k}+n_{2k})$ in Equation 6 suggests that the bias of $\hat{\psi}_2$ may be smaller.

Although both $\hat{\phi}_1$ and $\hat{\phi}_2$ have smaller bias, we prefer $\hat{\psi}(n_{1k}+n_{2k})$ over these two estimators, at least for our Mexico evaluation. This is because, as shown below, the variance estimate of $\hat{\phi}_1$ suffers from an even more serious invariance problem. The exact variance calculation of $\hat{\phi}_2$ is difficult because of the fact that Z_k appears in both numerator and denominator, and yet under the approximation assumption that the total number of units equals its expectation (a reasonable assumption if the number of sample clusters is large like in our Mexico evaluation), the variance of $\hat{\phi}_2$ coincides with that of $\hat{\phi}_1$ and thus suffers from the invariance problem.

4.3 Variance

We now show that nonparametric estimation of the variance of the average treatment effects is possible in matched-pair, cluster- randomized experiments. Contrary to stark claims in the literature we discuss below, these calculations do not require estimates of the intracluster correlation coefficient. Furthermore, in a critical comment on Klar and Donner (1997), Thompson (1998) shows how to obtain valid variance estimates by assuming the linear mixed effects model and the "common effect assumption." In their reply, Klar and Donner (1998) criticize the common effect assumption and as a result maintain their claim of analytical difficulties with the matched-pair design. We prove here how to obtain valid variance estimates without the common treatment effect assumption, any parametric modeling assumptions, or an estimate of the intracluster correlation coefficient.

4.3.1 Our Estimators

Rather than focusing on each of our proposed estimators separately, $\hat{\psi}(n_{1k}+n_{2k})$ and $\hat{\psi}(N_{1k}+N_{2k})$, we consider the variance of the general estimator, $\hat{\psi}(w_k)$ in Equation 5, so that our analytical results apply to any choice of weights. For notational simplicity, we use normalized weights, i.e., $\tilde{w}_k \equiv n w_k / \sum_{k=1}^{m} w_k$

(so that the weights sum up to n as in our estimator of SATE and UATE), and consider the variances of $\hat{\psi}(\tilde{w}_k)$. First, we use potential outcomes notation and write:

$$\hat{\psi}(\tilde{w}_k) \;=\; \frac{1}{n}\sum_{k=1}^{m} \tilde{w}_k \left\{ Z_k D_k(1) + (1 - Z_k) D_k(0) \right\}. \tag{10}$$

Then, our proposed variance estimator is defined by:

$$\begin{aligned}\hat{\sigma}(\tilde{w}_k) \equiv & \frac{m}{(m-1)n^2} \sum_{k=1}^{m} \left[\tilde{w}_k \left\{ Z_k \left(\frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}} - \frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}} \right) \right.\right. \\ & \left.\left. + (1 - Z_k) \left(\frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}} - \frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}} \right) \right\} - \frac{n\hat{\psi}(\tilde{w}_k)}{m} \right]^2. \end{aligned} \tag{11}$$

SATE. We first consider the variance of $\hat{\psi}(\tilde{w}_k)$ for estimating SATE. Taking the expectation of Equation 10 with respect to Z_k, the true variance of $\hat{\psi}(\tilde{w}_k)$ is given by:

$$\mathrm{Var}_a(\hat{\psi}(\tilde{w}_k)) \;=\; \frac{1}{4n^2} \sum_{k=1}^{m} \tilde{w}_k^2 (D_k(1) - D_k(0))^2. \tag{12}$$

From this expression, it is clear that the variance is not identified since we do not jointly observe $D_k(1)$ and $D_k(0)$ for each k. Thus, our strategy is to identify the sharp upper bound of this variance, making no additional assumptions, and to estimate it from the observed data.

The next proposition establishes that although the true variance, $\mathrm{Var}_a(\hat{\psi}(\tilde{w}_k))$, is not identifiable, our proposed variance estimator, $\hat{\sigma}(\tilde{w}_k)$, is unbiased for the sharp (i.e., best possible) upper bound of the true variance.

Proposition 1 (SATE Variance Identification) *Suppose that SATE is the estimand. Then, the true variance of $\hat{\psi}(\tilde{w}_k)$ is not identifiable, but its sharp upper bound can be estimated without bias by $\hat{\sigma}(\tilde{w}_k)$. Formally,*

$$\mathrm{Var}_a(\hat{\psi}(\tilde{w}_k)) \;=\; E_a(\hat{\sigma}(\tilde{w}_k)) - \frac{m}{4n^2} \mathrm{var}\left(\tilde{w}_k \sum_{j=1}^{2} \sum_{i=1}^{n_{jk}} \frac{Y_{ijk}(1) - Y_{ijk}(0)}{n_{jk}} \right),$$

where $\mathrm{var}(\cdot)$ *represents the sample variance with denominator* $m - 1$.

The proof for this proposition is given in Appendix A. The proposition implies that on average $\hat{\sigma}(\tilde{w}_k)$ overestimates the true variance $\mathrm{Var}_a(\hat{\psi}(\tilde{w}_k))$ unless the sample variance of (weighted) within-cluster SATEs across matched-pairs is zero (i.e., matching is highly effective). For example, if the sample average treatment effect is constant across matched-pairs and the cluster sizes are equal, then $\hat{\sigma}(\tilde{w}_k)$ estimates the true variance without bias. However, such a scenario is highly unlikely, and thus $\hat{\sigma}(\tilde{w}_k)$ should be seen as a conservative estimate of the variance.

CATE. Next, we study the identification of variance for CATE. The variance of $\hat{\psi}(\tilde{w}_k)$ for the estimation of CATE can be written as:

$$\begin{aligned}\mathrm{Var}_{au}(\hat{\psi}(\tilde{w}_k)) &= E_u\{\mathrm{Var}_a(\hat{\psi}(\tilde{w}_k))\} + \mathrm{Var}_u\{E_a(\hat{\psi}(\tilde{w}_k))\} \\ &= \frac{1}{4n^2}\sum_{k=1}^{m}\tilde{w}_k^2\left\{E_u(D_k(1)-D_k(0))^2 + \mathrm{Var}_u(D_k(1)+D_k(0))\right\}, \qquad (13)\end{aligned}$$

where the second equality holds because sampling of units is done independently within each cluster. Similar to the variance of SATE, this variance is not identified since we do not jointly observe $D_k(1)$ and $D_k(0)$ for each k. Thus, as before, we derive the sharp bounds of this variance in the next proposition.

Proposition 2 (CATE Variance Identification) *Suppose that CATE is the estimand. The true variance of $\hat{\psi}(\tilde{w}_k)$, $\mathrm{Var}_{au}(\hat{\psi}(\tilde{w}_k))$, is not identifiable, but its sharp bounds are given by:*

$$\frac{1}{2n^2}\sum_{k=1}^{m}\tilde{w}_k^2\left\{\mathrm{Var}_u(D_k(1)) + \mathrm{Var}_u(D_k(0))\right\} \;\leq\; \mathrm{Var}_{au}(\hat{\psi}(\tilde{w}_k)) \;\leq\; E_{au}(\hat{\sigma}(\tilde{w}_k)), \qquad (14)$$

where the difference between $\mathrm{Var}_{au}(\hat{\psi}(\tilde{w}_k))$ and the upper bound equals $\frac{m}{4n^2}\mathrm{var}\left\{\tilde{w}_k E_u(D_k(1)+D_k(0))\right\}$, and the difference between $\mathrm{Var}_{au}(\hat{\psi}(\tilde{w}_k))$ and lower bound equals $\frac{1}{4n^2}\sum_{k=1}^{m}\{\tilde{w}_k E_u(D_k(1)-D_k(0))\}^2$.

The upper bound can be estimated by $\hat{\sigma}(\tilde{w}_k)$ without bias, and an unbiased estimator of the lower bound is given by:

$$\hat{\lambda}(\tilde{w}_k) \;\equiv\; \frac{1}{n^2}\sum_{k=1}^{m}\tilde{w}_k^2\left\{(1-f_{1k})\frac{\mathrm{var}(Y_{i1k})}{n_{1k}} + (1-f_{2k})\frac{\mathrm{var}(Y_{i2k})}{n_{2k}}\right\},$$

where $(1-f_{jk})$ is the finite population correction with $f_{jk} = n_{jk}/N_{jk}$.

Proof of this proposition is given in Appendix B. The proposition implies that our proposed variance estimator, $\hat{\sigma}(\tilde{w}_k)$, is once again an unbiased estimate of the sharp upper bound of the true variance.

UATE and PATE. Unlike in the case of the SATE and the CATE, the variance of $\hat{\psi}$ is identified and can estimated without bias when UATE or PATE is the estimand, which we establish as the following proposition:

Proposition 3 (UATE and PATE Variance Identification) *The variances of $\hat{\psi}(\tilde{w}_k)$ for estimating the UATE and PATE are given by:*

$$\begin{aligned}\mathrm{Var}_{ap}(\hat{\psi}(\tilde{w}_k)) &= \frac{m}{n^2}\mathrm{Var}_p(\tilde{w}_k D_k), \\ \mathrm{Var}_{apu}(\hat{\psi}(\tilde{w}_k)) &= \frac{m}{n^2}\left[E_p\{\tilde{w}_k^2\mathrm{Var}_u(D_k)\} + \mathrm{Var}_p\{\tilde{w}_k E_u(D_k)\}\right],\end{aligned}$$

respectively, where $D_k = Z_k D_k(1) + (1-Z_k)D_k(0)$ and "p" represents the expectation with respect to simple random sampling of matched-pairs of clusters. Both variances can be estimated by $\hat{\sigma}(\tilde{w}_k)$ without bias under their corresponding sampling schemes.

This proposition is proved in Appendix C. The proposition shows that when the estimand is PATE, the variance of $\hat{\psi}(\tilde{w}_k)$ is proportional to the sum of two elements: the mean of within-cluster variances and the variance of within-cluster means. If all units are included in each cluster, then the first term will be zero because the within-cluster means are observed without sampling uncertainty, i.e., $\mathrm{Var}_u(D_k) = 0$ for all k. In either case, however, our proposed estimator $\hat{\sigma}(\tilde{w}_k)$ estimates the variance without bias.

Inference. Given our proposed causal effect estimators and variances, we make statistical inferences by assuming that $\hat{\psi}(\tilde{w}_k)$ is approximately unbiased. We consider three situations:

1. *Many pairs.* When the number of pairs is large (regardless of the number of units within each cluster), no additional assumption is necessary due to the central limit theorem. For PATE and UATE, the level α confidence intervals are given by $[\hat{\psi}(\tilde{w}_k) - z_{\alpha/2}\sqrt{\hat{\sigma}(\tilde{w}_k)},\ \hat{\psi}(\tilde{w}_k) + z_{\alpha/2}\sqrt{\hat{\sigma}(\tilde{w}_k)}]$ where $z_{\alpha/2}$ represents the critical value of two-sided level α normal test. For the SATE and CATE, the confidence level of this interval will be greater than or equal to α.

2. *Few pairs, many units.* For CATE (and PATE), the central limit theorem implies that D_k follows the normal distribution. Since the weights are assumed to be fixed for CATE, $\tilde{w}_k D_k$ is also normally distributed. For the other quantities, we assume $\tilde{w}_k D_k$ is normally distributed. In either case, the level α confidence intervals are given by $[\hat{\psi}(\tilde{w}_k) - t_{m-1,\alpha/2}\sqrt{\hat{\sigma}(\tilde{w}_k)},\ \hat{\psi}(\tilde{w}_k) + t_{m-1,\alpha/2}\sqrt{\hat{\sigma}(\tilde{w}_k)}]$, where $t_{m-1,\alpha/2}$ represents the critical value of the one-sample two-sided level α t-test with $(m-1)$ degrees of freedom. For the SATE and CATE, the confidence level of this interval will be greater than or equal to α.

3. *Few pairs, few units.* When little information is available, a distributional assumption is required for the inferences about all four quantities. We may assume $\tilde{w}_k D_k$ follows the normal distribution as above and construct the confidence intervals and conduct hypothesis tests based on t-distribution.

While statistical inference can be easily conducted as described here under the matched-pair design, confidence interval construction and hypothesis tests under both the complete randomized and stratified designs generally involve approximations as well as distributional assumptions. In particular, the well-known Behrens-Fisher problem prevents exact calculations under either design with the normality assumption unless one is willing to make an optimistic assumption about the equality of variances of two potential outcomes. This problem is important especially when the number of clusters is small, and

it also makes exact power and sample size calculations difficult under these alternative designs. Such problems do not occur with the matched-pair design.

4.3.2 Unbiased Estimator

The variance of the unbiased estimator in Equation 8 can also be derived in an analogous manner to that of our estimators. For example, when the estimand is UATE (recall that this estimator is not generally applicable to CATE or PATE), the true variance is given by:

$$\mathrm{Var}_{ap}(\phi_1) \quad = \quad \frac{4m}{n^2}\mathrm{Var}_p(\widetilde{Y}_{jk}(1) - \widetilde{Y}_{j'k}(0)),$$

where $\widetilde{Y}_{jk}(t) = \sum_{i=1}^{n_{jk}} Y_{ijk}(t)$ for $t = 0, 1$. Calculation similar to the one in Appendix C shows that the unbiased estimator of this variance is:

$$\hat{\xi} \quad \equiv \quad \frac{4m}{(m-1)n^2}\sum_{k=1}^{m}\left\{Z_k(\widetilde{Y}_{1k} - \widetilde{Y}_{2k}) + (1 - Z_k)(\widetilde{Y}_{2k} - \widetilde{Y}_{1k}) - \frac{n\hat{\phi}_1}{m}\right\}^2.$$

However, just like $\hat{\phi}_1$, this variance estimator is not invariant to a constant shift unless cluster sizes are identical within each matched-pair, i.e., $n_{1k} = n_{2k}$ for all k, in which case $\hat{\phi}_1$ equals our proposed estimator $\hat{\psi}(n_{1k} + n_{2k})$.

4.3.3 Existing Estimator Based on Harmonic Mean Weights

In addition to the different choices of weights, our proposed estimator in Equation 11 differs from the variance estimator given in the literature for the general weighted estimator $\hat{\psi}(\tilde{w}_k)$ (see e.g., Donner, 1987; Donner and Donald, 1987; Donner and Klar, 1993). Using our notation and normalized weights, this estimator can be written as:

$$\begin{aligned}\hat{\delta}(\tilde{w}_k) \quad \equiv \quad & \frac{\sum_{k=1}^{m}\tilde{w}_k^2}{n^3}\sum_{k=1}^{m}\tilde{w}_k\left\{Z_k\left(\frac{\sum_{i=1}^{n_{1k}}Y_{i1k}}{n_{1k}} - \frac{\sum_{i=1}^{n_{2k}}Y_{i2k}}{n_{2k}}\right)\right. \\ & \left. +(1 - Z_k)\left(\frac{\sum_{i=1}^{n_{2k}}Y_{i2k}}{n_{2k}} - \frac{\sum_{i=1}^{n_{1k}}Y_{i1k}}{n_{1k}}\right) - \hat{\psi}(\tilde{w}_k)\right\}^2. \qquad (15)\end{aligned}$$

As is the case for the choice of weights, the literature provides no formal justification of this variance estimator. However, unless all of the restrictive modeling assumptions described Section 4.2.2 hold, this variance estimator is invalid. In Appendix D, we show that unlike our proposed estimator, $\hat{\delta}(\tilde{w}_k)$ is biased and the direction of bias is indeterminate: sometimes the standard error computed as recommended in the literature will be too small and sometimes too large (examples are offered in Section 4.3.5). The only

condition under which this estimator is approximately equal to our estimator is when m is large and the weights are identical across matched-pairs, i.e., $\tilde{w}_k = n/m$ for all k. However, such a scenario is highly unlikely, and so it is hard to see any reason to use this variance estimator in practice.

4.3.4 The Irrelevance of the Intracluster Correlation Coefficient

Klar and Donner (1997) list the inability to estimate the intracluster correlation coefficient (ICC) as one of the major disadvantages of the matched-pair design in cluster randomized experiments. Donner (1998) goes further to claim that "an estimate of ρ [ICC] is required to compute the appropriate standard errors for the analyses in question. These consequences do not occur for data arising from either the completely randomized design or the stratified design" (p.99). Similarly, Campbell *et al.* (2001) claim that the ICC is "essential if robust sample size calculations are to be made."

However, as we have already shown in this section, these claim are incorrect: our proposed nonparametric estimator of average treatment effects under the matched-pair design does not require estimation of the ICC. Later in the paper, we also show that efficiency analysis, power comparisons, and sample size calculations can be conducted without the estimation of ICC.

In fact, ICC is of little use for design-based nonparametric analysis of cluster randomized experiments for more general reasons. Most importantly, "The ICC is only defined for clusters of equal sizes" (Lohr, 1999, p.140) whereas most cluster-randomized experiments involve clusters of unequal sizes. When cluster sizes vary, an alternative measure of within-cluster homogeneity sometimes used in the literature is the increase in the coefficient of determination, or R^2, due to fixed effects representing the clusters. However, this alternative measure is applicable only within the framework of linear regression or ANOVA, while maintaining its linearity and other assumptions. In contrast, our proposed design-based nonparametric approach avoids all such assumptions and can produce estimates entirely without the ICC or its substitute.

4.3.5 SPS Evaluation

In Section 4.3.3, we prove that the variance estimator used recommended in the literature is biased, and the direction of the bias is indeterminate, meaning it will differ across data sets and variables. Here we demonstrate that the biases in real data in this variance estimator are large enough to make it unusable in practice.

We begin by computing the standard error (the square root of the estimated variance) based on the (biased) general variance formula of Equation 15 proposed in Donner (1987), Donner and Donald

(1987), and Donner and Klar (1993), as well as the one based on our unbiased variance estimator of Equation 11. We use the arithmetic mean weights for both standard error calculations in order to restrict this comparison to the Donner-Klar variance formula rather than confounding it with additional problems occurring due to their choice of harmonic weights.

We make these computations for a large number of outcome variables from the SPS evaluation survey conducted 10 months after randomization. The outcome variables include some which were binary (e.g., did the respondent suffer catastrophic medical expenditures? Does our blood test indicate that the respondent has high cholesterol? Has the respondent been diagnosed with asthma?) and others denominated in Mexican pesos (e.g., out-of-pocket expenditures for health care, for drugs, etc.). We then computed the ratio: the biased Donner-Klar standard error divided by our unbiased alternative for each variable. The left graph in Figure 1 gives a smoothed histogram of these ratios (plotted on the log scale but labeled in original units). In these real data, the Donner-Klar standard errors range from about two times too small to two times too large. Note that the central tendency of this histogram has no particular meaning, as it is constructed from whatever questions happened to be asked on the survey. The key point is that the deviation from the unbiased estimator for any *one* Donner-Klar standard error can be huge and many in these data are huge. Thus, in addition to the theoretical proof that these standard errors are biased, the results here suggest that in real data the errors, measured by deviations from the unbiased standard error or confidence interval coverage, are too large to make them usable. (We discuss the two right panels in this figure below.)

4.3.6 Monte Carlo Evidence

As shown above, the existing harmonic mean-based estimator is appropriate when the corresponding modeling assumptions hold, but the model applies only when there is no point in doing matching to begin with. Through Monte Carlo simulation based in part on real Mexico data, we address here the specific consequences of violating the assumptions of this model. To construct realistic simulations, we begin with the observed cluster-specific mean for two out-of-pocket health expenditures from the SPS evaluation data (measured in pesos) and use this to set the potential outcomes' true population for the simulation. Finally, we generate the outcome variables via independent normal draws for units within clusters using a set of heterogeneous variances. Thus, the existing estimator's mean and variance constancy assumptions are violated, as in real data, although its normality and independence assumptions are maintained for the first set of simulations.

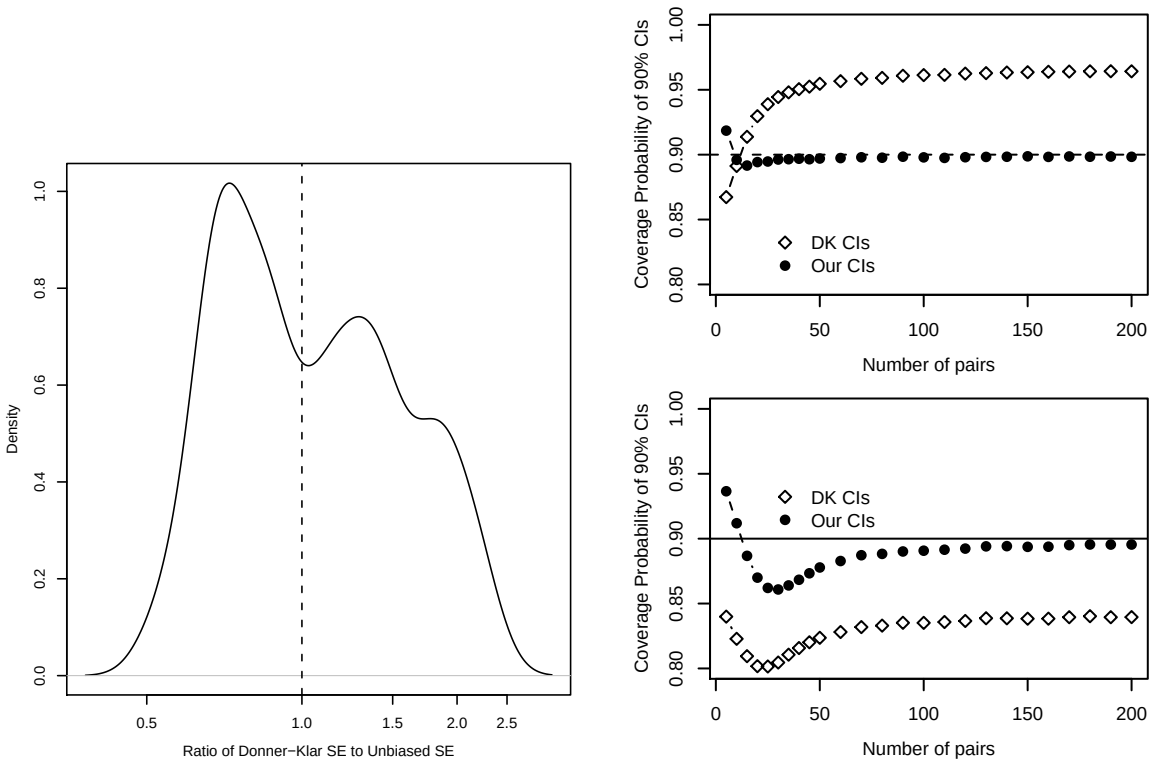

Figure 1: Inference Accuracy. The left panel gives the the ratios of the biased Donner-Klar standard error to our unbiased standard error on the horizontal axis (on the log scale, but labeled as ratios for interpretability). Clearly many of the biased standard error estimates are far from the unbiased figures (noted at the horizontal axis at 1). For the coverage probabilities in the right panel, Monte Carlo simulations demonstrate how our estimator is approximately correct and increasingly so for larger sample sizes, while the Donner-Klar estimator can yield confidence intervals that are either too large (top panel) or too small (bottom panel).

We estimate the bias and root mean square error (RMSE) of the harmonic mean-based estimator by treating these clusters as fixed and thereby focusing on the estimation of CATE. In the two simulated cases, our estimator is unbiased (since the PATEs are the same within each pair). In contrast, the harmonic mean estimator is biased downward by 35 pesos in one case and biased upward by 13 pesos in the other. In the first simulated case, our estimator's RMSE (7.2) is nearly five times smaller than the harmonic mean estimator RMSE (35.7), while in the second case, our RMSE (5.1) is almost three times smaller than that for the harmonic mean (14.2). As shown in Proposition 2, the CATE variance is not identified (and the expectation of our variance estimator equals the sharp upper bound). Thus, our variance estimator as well as the Donner-Klar variance estimator are larger than the true variance of the corresponding estimators, yielding conservative confidence intervals.

Next, we study the PATE and UATE estimators. To do this by staying close to the data, we generate

a population of clusters by bootstrapping the observed pairs of SPS clusters along with their observed means and a set of heterogeneous variances. We then examine the properties of confidence intervals based on both estimators by calculating the coverage probabilities. We draw from the discrete empirical distribution, which is far from Gaussian. The right panels of Figure 1 summarize the results. As expected due to the central limit theorem, both sets of our 90% confidence intervals (solid disks) approach their corresponding nominal coverage probabilities as the number of pairs increase. In contrast, the confidence intervals based on the Donner-Klar variance estimator (open diamonds) are severely biased — too wide in the example in the top graph and too narrow in example in the bottom — and the magnitude of bias does not decrease even as the number of pairs grow.

Although the harmonic mean-based estimator is the uniformly minimum variance unbiased estimator under its restrictive modeling assumptions, our simulations demonstrate that the estimator can perform badly when these assumptions are violated. Our estimator, in contrast, does not rely on these modeling assumptions and can significantly outperform the harmonic mean approach, and Donner-Klar standard errors, in the presence of across-pair heterogeneity.

4.4 Cluster-Level Quantities of Interest

The eight quantities of interest defined in Section 3.3 — SATE, CATE, PATE, and UATE, both with and without interference — are all defined as aggregations of unit-level causal effects. For some purposes, however, analogous quantities of interest can be defined at the cluster level. For example, quantities of interest in the SPS evaluation include the health clinic-level variables. Some of these effects, such as the supply of drugs and doctors, are defined and measured at the health clinic, and so are effectively unit level variables amenable to unit level analyses.

However, for some variables, individual-level survey responses are required to measure the quality of the health clinics. These include the success clinics have in protecting privacy, reduce waiting times, etc. If these latter variables are used to judge the causal effect of SPS on the clinics, we have a cluster-randomized experiment, but a quantity of interest at the cluster level. In this situation, our estimator is a special case of Equation 5, with a constant weight: $\hat{\psi}(1)$. Similarly, the variance of this estimator is a special case of our general formulation in Equation 11: $q\hat{\sigma}(1)$.

Unlike our estimator for quantities defined at the individual level, this estimator for aggregate quantities is unbiased and invariant for all quantities of interest.

5 Comparing Matched-Pair and Other Designs

In this section, we study the relative efficiency and power of the matched-pair and completely randomized (or "unmatched") designs in cluster randomized experiments, and give sample size calculations for matched-pair designs. We also briefly compare the matched-pair design with the stratified design and discuss the consequences of loss of clusters under each design.

5.1 Completely Randomized Design

The complete randomization design is defined as follows. Consider a random sample of $2m$ clusters from a population. We observe a total of n_j units within the jth cluster in the sample, and use n to denote the total number of units in the sample: $n = \sum_{j=1}^{2m} n_j$. Under this design, m randomly selected clusters are assigned to the treatment group with equal probability while the remaining m clusters are assigned to the control group.

We construct an estimator analogous to the one proposed for the matched-pair design as:

$$\begin{aligned} \hat{\tau}(\tilde{w}_j) &\equiv \frac{2}{n}\sum_{j=1}^{2m}\sum_{i=1}^{n_j}\frac{\tilde{w}_j}{n_j}\{Z_jY_{ij} - (1-Z_j)Y_{ij}\} \\ &= \frac{2}{n}\sum_{j=1}^{2m}\sum_{i=1}^{n_j}\frac{\tilde{w}_j}{n_j}\{Z_jY_{ij}(1) - (1-Z_j)Y_{ij}(0)\}, \end{aligned} \tag{16}$$

where Z_j is the randomized binary treatment variable, $Y_{ij}(t)$ is the potential outcome for the ith unit in the jth cluster under the treatment value t for $t = 0, 1$, and $\tilde{w}_j$ is the known normalized weight with $\sum_{j=1}^{2m}\tilde{w}_j = n$. For SATE and UATE, we use $\tilde{w}_j = n_j$. For CATE and PATE, we use $\tilde{w}_j \propto N_j$ where N_j is the population size of the jth cluster. Analysis similar to the one in Section 4.2 shows that this estimator is unbiased for all four quantities in completely randomized (unmatched) cluster-randomized experiments.

The commonly used estimator in the literature for this design takes a form that is slightly different from the one given in Equation 16. Rather, it is similar to $\hat{\phi}_2$ given in Equation 9: $\hat{\kappa} \equiv \sum_{j=1}^{2m} Z_j \sum_{i=1}^{n_j} Y_{ij} / \sum_{j=1}^{2m} Z_j n_j + \sum_{j=1}^{2m}(1-Z_j)\sum_{i=1}^{n_j} Y_{ij}/(n - \sum_{j=1}^{2m} Z_j n_j)$. This estimator is applicable to the estimation of SATE and UATE but not CATE and PATE because it does not incorporate population weights for clusters. The estimator is also biased for SATE and UATE, and the magnitude of bias can be derived using the Taylor series as done for $\hat{\phi}_2$. Without modeling assumptions, the exact variance calculation is difficult within the design-based inferential framework for the same reason as before: the

denominator as well as the numerator is a function of the randomized treatment variable. In addition, the usual approximate variance calculations for such a ratio estimator yield either the same variance as $\hat{\tau}(n_j)$ or the variance estimator that is not invariant to a constant shift. Thus, for the remainder of this section we focus on $\hat{\tau}(\tilde{w}_j)$.

For the rest of this section, we assume that the estimand is UATE. However, the exact same calculations apply when the estimand is PATE since the variance estimator is the same for two estimands. For SATE and CATE, we can interpret these results as the most conservative estimates of efficiency, power, and sample sizes.

5.2 Efficiency

When the estimand is UATE, the variance of $\hat{\tau}(\tilde{w}_j)$ of this estimator is given by:

$$\mathrm{Var}_{ac}(\hat{\tau}(\tilde{w}_j)) \quad = \quad \frac{4m}{n^2}\left\{\mathrm{Var}_c(\tilde{w}_j\overline{Y_j(1)}) + \mathrm{Var}_c(\tilde{w}_j\overline{Y_j(0)})\right\},$$

where $\overline{Y_j(t)} \equiv \sum_{i=1}^{n_j} Y_{ij}(t)/n_j$ for $t = 0, 1$, and the subscript "c" represents the simple random sampling of clusters.

For simplicity and to facilitate the comparison, assume that under the matched-pair design, one is able to match on cluster sizes within each match so that $n_{1k} = n_{2k}$ for all k. Proposition 3 implies that under this condition $\hat{\psi}(\tilde{w}_k)$ is unbiased and its variance is given by:

$$\mathrm{Var}_{ap}(\hat{\psi}(\tilde{w}_k)) \quad = \quad \frac{m}{n^2}\mathrm{Var}_p\left\{\tilde{w}_k(\overline{Y_{jk}(1)} - \overline{Y_{j'k}(0)})\right\},$$

where $\overline{Y_{jk}(t)} \equiv \sum_{i=1}^{n_{jk}} Y_{ijk}(t)/n_{jk}$ and $j \neq j'$. Since the assumption of $n_{1k} = n_{2k}$ means $\tilde{w}_{jk} = 2\tilde{w}_j$, we have $\mathrm{Var}_p(\tilde{w}_k\overline{Y_{jk}(t)}) = 4\mathrm{Var}_c(\tilde{w}_j\overline{Y_j(t)})$ for $t = 0, 1$. Thus, the relative efficiency of the matched-pair design over the completely randomized design is given by:

$$\frac{\mathrm{Var}_{ac}(\hat{\tau}(\tilde{w}_j))}{\mathrm{Var}_{ap}(\hat{\psi}(\tilde{w}_k))} \quad = \quad \left\{1 - \frac{2\mathrm{Cov}_p(\tilde{w}_k\overline{Y_{jk}(1)}, \tilde{w}_k\overline{Y_{j'k}(0)})}{\sum_{t=0}^{1}\mathrm{Var}_p(\tilde{w}_k\overline{Y_{jk}(t)})}\right\}^{-1}.$$

The result implies that the relative efficiency of the matched-pair design depends on the correlation of the observed within-pair cluster mean outcomes (weighted by cluster sizes). If matching induces a positive correlation, as is its purpose and will normally happen in practice, then the matched-pair design is more efficient. Note that under the matched-pair design, we can estimate $\mathrm{Cov}_p(\tilde{w}_k\overline{Y_{jk}(1)}, \tilde{w}_k\overline{Y_{j'k}(0)})$ without bias using the sample covariance between $\tilde{w}_k\overline{Y_{jk}(1)}$ and $\tilde{w}_k\overline{Y_{j'k}(0)}$, which are jointly observed for each k. And thus, under the matched-pair design, the variance one would obtain under the completely

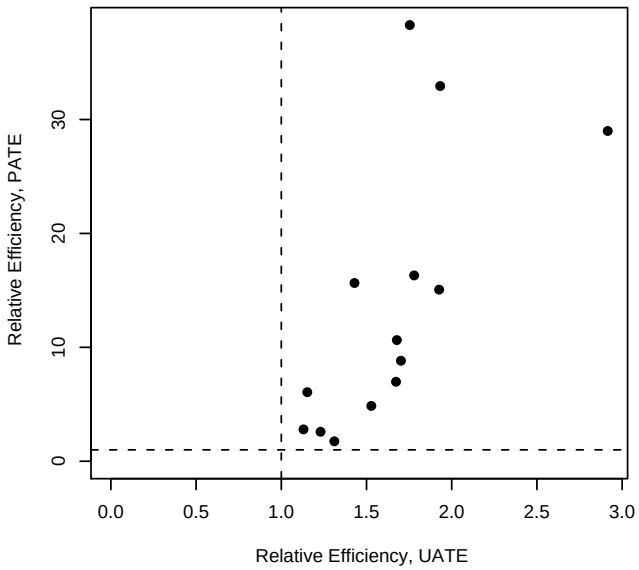

Figure 2: Relative Efficiency of Matched-Pair over Completely Randomized Designs in the SPS Evaluation

randomized design can also be estimated without bias (Author cite redacted) This is another advantage of the matched-pair design since the converse is not true. (If cluster sizes are equal, one can also estimate the ICC nonparametrically and separately for the treated and control groups — there is no reason to assume the ICC is the same for two potential outcomes as done in the literature — although the ICC is not required for efficiency, power, or sample size calculations.) We use this strategy to empirically evaluate the relative efficiency of the matched-pair design in our Mexico evaluation.

SPS Evaluation. Although matched-pair designs have other advantages in public policy evaluations (King *et al.*, 2007), its advantage in statistical efficiency can be considerable. We now use this result to estimate the efficiency of the matched-pair design as used in the SPS evaluation over the efficiency that our experiment would have achieved, if we had used complete randomization without matching.

In Figure 2 we plot the relative efficiency of our estimator of matched pairs over completely randomized designs for UATE and for PATE. We do this for our 14 outcome variables denominated in pesos. For UATE, the estimator based on the matched pair design is between 1.13 and 2.92 times more efficient, which means that our standard errors would have been as much as $\sqrt{2.92} = 1.7$ times larger if we had neglected to match first. The result is even more dramatic for estimating PATE, for which the matched-pair design ranges between 1.8 and 38.3 times more efficient! In this situation, our standard errors would have been more as much as *six* times larger if we had neglected to match first.

5.3 Power

We now use the variance results in Section 4.3 to show how to calculate statistical power, the probability of rejecting the null if it is indeed false, for matched-pair cluster-randomized experimental designs. We consider the power calculation for UATE and PATE, which also represent the minimum power for SATE and CATE, respectively.

5.3.1 Power Calculations under the Matched-Pair Design

We begin with a power calculation for UATE given a null hypothesis of $H_0 : \psi_U = 0$, the alternative hypothesis of $H_A : \psi_U = \psi$, and the level α t-test. Under this general setting, Proposition 3 implies the following power function: $1 + \mathcal{T}_{m-1}(-t_{m-1,\alpha/2} \mid n\psi/\sqrt{m\text{Var}_p\{\tilde{w}_k D_k)} - \mathcal{T}_{m-1}(t_{m-1,\alpha/2} \mid n\psi/\sqrt{m\text{Var}_p\{\tilde{w}_k D_k\}})$, where $\mathcal{T}_{m-1}(\cdot \mid \zeta)$ is the distribution function of the noncentral t distribution with $(m-1)$ degrees of freedom and the noncentrality parameter ζ, and $\tilde{w}_k = n_{1k} + n_{2k}$. In the case of UATE, pairs of clusters (but not units within each cluster) are sampled. Thus, a simpler expression of the power function, which can be more straightforwardly used for sample size calculations, is obtained if we assume cluster sizes are equal. In that case, a researcher may reparameterize the power function by normalizing ψ in terms of the standard deviation of within-pair difference-in-means: $d_U \equiv \psi/\sqrt{\text{Var}(D_k)}$. Then, the power function can be simplified as:

$$1 + \mathcal{T}_{m-1}(-t_{m-1,\alpha/2} \mid d_U\sqrt{m}) - \mathcal{T}_{m-1}(t_{m-1,\alpha/2} \mid d_U\sqrt{m}). \tag{17}$$

Next, suppose that PATE is the estimand and we sample units within each cluster as well as pairs of clusters. The null hypothesis is given by $H_0 : \psi_P = 0$ and the alternative is $H_a : \psi_P = \psi$. Again, for simplicity, we assume sample cluster sizes are equal, i.e., $\bar{n} = n_{jk} = n/(2m)$ for all j and k. Then, Proposition 3 implies the power function of the same form as Equation 17 except that the noncentrality parameter is given by $\psi\sqrt{m}/\sqrt{\sum_{j=1}^{2} E_p\{\tilde{w}_k^2 \text{Var}_u(Y_{ijk})\}/\bar{n} + \text{Var}_p(\tilde{w}_k E_u(D_k))}$ where $\tilde{w}_k \propto N_{1k} + N_{2k}$. Similar to the case of UATE, one can obtain a simpler expression of the power function if population clusters sizes are assumed to be equal. Under this additional assumption, the power function becomes:

$$1 + \mathcal{T}_{m-1}\left(-t_{m-1,\alpha/2} \,\middle|\, \frac{d_P\sqrt{m}}{\sqrt{1+\pi/\bar{n}}}\right) - \mathcal{T}_{m-1}\left(t_{m-1,\alpha/2} \,\middle|\, \frac{d_P\sqrt{m}}{\sqrt{1+\pi/\bar{n}}}\right), \tag{18}$$

where, as in the case of UATE, ψ is normalized by the standard deviation of within-pair (true) difference-in-means, i.e., $d_P \equiv \psi/\sqrt{\text{Var}_p\{E_u(D_k)\}}$, and π represents the ratio of the mean variances of the potential

outcomes and the variance of within-pair differences-in-means by the mean variances of the potential outcomes, i.e., $\pi \equiv \sum_{j=1}^{2} E_p\{\mathrm{Var}_u(Y_{ijk})\}/\mathrm{Var}_p(E_u(D_k))$.

5.3.2 Sample Size Calculations

We can use the above results to estimate the sample size required to achieve the desired precision in a future experiment under the matched-pair design. Suppose that an investigator wishes to specify the desired degree of precision in terms of Type I and Type II error rates in hypothesis testing, denoted by α and β, respectively. In particular, the goal is to calculate the sample size required to achieve a given degree of power, $1-\beta$, against a particular alternative (e.g., Snedecor and Cochran, 1989, Section 6.14). Such a calculation can be conducted by using the power functions derived in Section 5.3.1. For example, suppose that the estimand is UATE and cluster sizes are equal. Then, using Equation 17, the desired number of pairs of clusters is given by the smallest value of m such that $1+\mathcal{T}_{m-1}(-t_{m-1,\alpha/2} \mid d_U\sqrt{m}) - \mathcal{T}_{m-1}(t_{m-1,\alpha/2} \mid d_U\sqrt{m}) \geq 1-\beta$ where $d_U \equiv \psi/\sqrt{\mathrm{Var}(D_k)}$ is specified by a researcher along with the value of α and β. Similarly, for PATE, Equation 18 can be used to determine the number of pairs and units within each cluster.

SPS Evaluation. To illustrate, we use SPS evaluation data on the annualized out-of-pocket health care expenditure that a household spent in the most recent month. Using estimates of π and $\mathrm{Var}_p\{E_u(D_k)\}$ from the SPS data and Equation 18, we calculate the minimal absolute effect size for PATE that can be detected using a two-sided t-test with 0.95 size and power, for any given cluster size and number of cluster pairs. Since the household is the unit of interest in this example, our population count involves the number of households per cluster, instead of the number of individuals.

In the left panel of Figure 3, the number of pairs can be found on the horizontal axis and the number of units within each cluster on the vertical axis. The contour lines represent the minimum detectable size in pesos. The graph shows for example that the matched-pair design with 30 pairs and 100 units within each cluster can detect the true absolute effect size of approximately 450 pesos with the desired precision. The figure displays the obvious result that experiments with more pairs or clusters, can detect smaller sized effects (contour lines are labeled with smaller numbers as we move to the top right of the figure). More importantly, the nearly vertical contour lines (above 50 or so units within each cluster) indicates that adding more pairs of clusters adds more statistical power than adding more units within each pair. However, adding one more pair means that many more units will be added, and in some

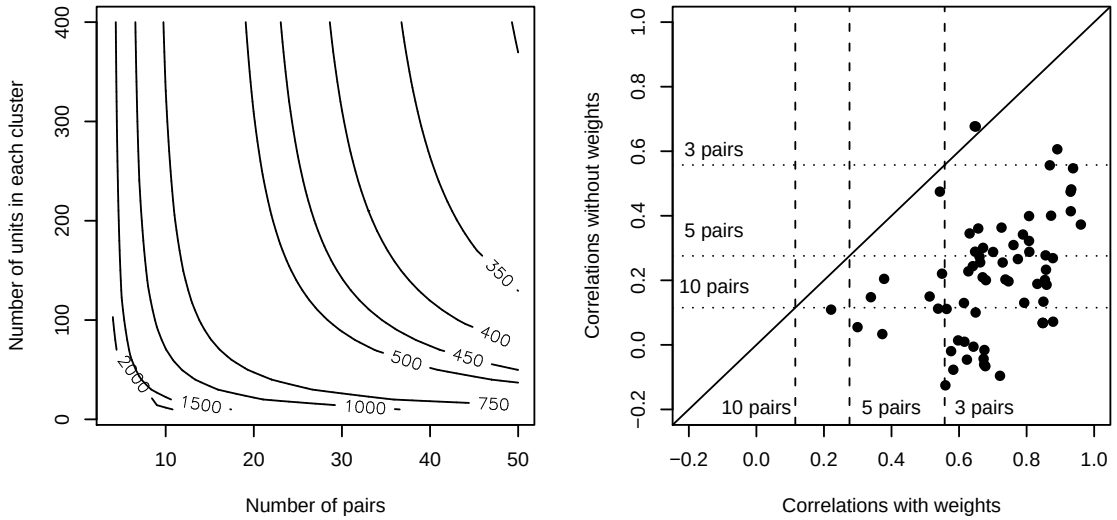

Figure 3: Sample Size Calculations for PATE under the Matched-Pair Design. The left panel plots the smallest detectable absolute effect size of SPS on annualized out-of-pocket expenditures (in pesos) using a 0.05 level two-sided test with 0.95 power, with π estimated from SPS evaluation data. The horizontal and vertical axes plot m and $\bar{n}$, respectively. The right panel compares correlations with and without population weights between treatment and control group cluster-specific means in SPS evaluation data. All but one variable has higher correlation when incorporating weights as seen by a dot appearing below the 45° line. The graph also presents "break-even" correlations (indicated by dashed and dotted lines for with and without weights, respectively), which are the smallest possible correlations matching must induce in order for the matched-pair design to detect smaller effect size than the completely randomized design, given the fixed power (0.8) and size (0.95) of the test. The graph suggests, when weights are appropriately taken into consideration, that the matched-pair design should be preferred (for all but possibly one variable) even when the number of pairs is as small as three.

situations sampling units within new clusters is more expensive than within existing clusters. As such, the exact tradeoff depends on the specifics of each application. (We discuss the right panel of the figure next.)

5.3.3 Power Comparison

Although the matched-pair design is typically more efficient than the completely randomized design regardless of sample size, Martin *et al.* (1993, p.330) point out that when the number of matched pairs is small (fewer than about 10), "the matched design will probably have less power than the unmatched design" due to the loss of degrees of freedom in estimating the variance. Here, we show that this conclusion critically hinges on Martin *et al.*'s assumption of equal cluster population sizes as well as their particular choice of an assumed parametric model relating the matching and outcome variables. Modeling

assumptions are always worrisome since they can be wrong, but the equal cluster size assumption is especially problematic because varying cluster sizes is in fact a fundamental feature of almost all cluster-randomized experiments.

When cluster sizes are unequal, the efficiency gain of matching in cluster randomized trials depends on the correlations of *weighted* cluster means between the treatment and control clusters across pairs (with weights based on sample or population cluster sizes depending on the quantity of interest), not the unweighted correlations used in Martin *et al*'s calculations. Since population cluster sizes are typically observed prior to the treatment randomization, researchers can directly incorporate this variable into their matching procedure. As a result, correlations of weighted outcomes (constructed from clusters with matched weights) will usually be *substantially* higher than those of unweighted outcomes; this is true even when cluster sizes are independent of outcomes. Thus, in cluster-randomized trials with unequal cluster sizes, the efficiency gain due to pre-randomization matching is likely to be considerably greater than the equal cluster size case considered by Martin *et al.* (1993). (Along with the bias reduction, this is another reason to incorporate cluster sizes into one's matching procedure). Any power comparison must take this factor into consideration.

SPS Evaluation. The right panel of Figure 3 illustrates the argument above using the SPS evaluation data, by calculating the across-pair correlations between treatment and control cluster means of 67 outcome variables (ranging from health related variables to household health care expenditure variables), both with and without weights. We use population cluster sizes as weights, which were observed prior to the randomization of the treatment and incorporated into the matching procedure used for the SPS evaluation (King *et al.*, 2007). The graph shows that all but one variable has considerably higher correlations when weights are incorporated (which does not make the assumption of equal cluster size) than when they are ignored (which assumes constant cluster size); this can be seen by all but one of the dots falling below the solid 45° line. In fact, the median of the correlations is more than *three* times larger with (0.68) than without (0.20) weights. In their conclusion, Martin *et al.* (1993, p.336) recommend that if the number of pairs is 10 or fewer, then matching should be used only if researchers are confident that the correlation due to matching is at least 0.2. Indeed, all variables in SPS meet this criteria if the weights are appropriately taken into account, the minimum correlation with weights being 0.22. (If the correlations are calculated incorrectly without weights, then only about half of the variables meet their criteria.)

To illustrate the above result more precisely in terms of power and sample size calculations, the graph also presents the "break-even" matching correlations (indicated by dashed and dotted lines for correlations with and without weights, respectively) that are used by Martin *et al.* (1993, Section 7). As in the original article, we set the power and size of the test to be 0.8 and 0.95, respectively, and derive the smallest correlation matching must induce in order for the matched-pair design to be able to detect smaller effect sizes than the completely randomized design. The result indicates that even with as few as 3 pairs, more than 85% of the variables had a correlation higher than the break-even point, which is 0.56. With 5 pairs, all but one variable exceeds the threshold.

In contrast, if one ignores the weights, by incorrectly assuming that the clusters are equally sized as in Martin et al., then only 4% and 34% of the variables have the correlations higher than the break-even correlations of 3 and 5 pairs, respectively. Martin *et al.* (1993) described the correlation of 0.25 as "difficult to achieve by matching" (p.335). However, as the data from SPS evaluation show, since one can match on cluster sizes, the level of *weighted* correlations is much higher when cluster sizes are different. (It is technically possible to reduce power even when cluster sizes are exactly matched if one pairs clusters that turn out to be maximally different from one another on the outcome variable. However, we would not expect this bizarre situation to occur in practice; and even if it does, exact matching on cluster sizes would still eliminate bias.)

Thus, by dropping the unrealistic assumption that all clusters are equally sized we have shown here that, for practical purposes, the matched pair design may well have more statistical power than the completely randomized design, even for small samples. But if one has fewer than three matched pairs of clusters, its probably time to stop worrying about the properties of statistical estimators and start planning to return to the field to gather more data!

5.4 Lost Clusters and Stratified Designs

We now briefly discuss two additional critiques leveled at matched-pair designs that also turn out to be incorrect. First, Donner and Klar (2000, p.40) recommend the stratified design, where units are matched in blocks of larger than two, in part because they say it avoids "the unique analytical challenges of the matched-pair design." We have shown above that these "analytical challenges" do not exist. In addition, a stratified design is nothing more than a completely randomized design operating within each strata. If all units within a strata have identical values on the important background covariates, then this design

is effectively equivalent to the matched pair design. However, if any heterogeneity on these covariates or cluster sizes remain within any strata, as will almost always be the case if the researcher has collected sufficient pre-treatment measures, then the stratified design may leave some efficiency on the table. Thus, when feasible, switching from a stratified to a matched pair design has the potential to greatly increase efficiency and power. The more heterogeneity that exists within strata, the more additional benefits one would get from switching to a matched pair design.

Finally, Donner and Klar (2000, p.40) claim that a "disadvantage of the matched-pair design is that the loss to follow-up of a single cluster in a pair implies that both clusters in that pair must effectively be discarded from the trial, at least with respect to testing the effect of the intervention. This problem ... clearly does not arise if there is some replication of clusters within each combination of intervention and stratum." In fact, nearly the opposite is true. If a cluster is lost in a matched-pair study for reasons related to pre-treatment variables, then dropping the other member of the pair makes it possible to retain all the benefits of randomization in the remaining pairs (King *et al.*, 2007). For example, one can estimate SATE for this new sample without bias. In contrast, the loss of even a single cluster in a completely randomized design turns an experimental study into an observational study requiring the addition of ignorability assumptions normally abhorrent to experimentalists. The loss of a single cluster within a strata larger than two units means that more than one cluster will need to be dropped in order to retain the benefits of randomization. (Of course, at the cost of some model-dependence, the researcher in any of these designs may wish to impute rather than discard the missing cluster.)

6 Methods for Unit-Level Noncompliance

In the evaluation of SPS described in Section 2, some individuals in treatment clusters did not receive treatment while others in the control cluster received the treatment. The reason is that individuals need to affiliate with the SPS program in order to receive medical services. And although the state governments spend considerable effort trying to affiliate individuals to SPS in treated clusters (and no effort in control clusters), compliance is not perfect. The program is designed primarily for the poor and so the wealthy, who have their own preexisting health care arrangements, often do not sign up. Moreover, since individual action is required, some of the non-wealthy also fail to affiliate, although the task is made easier since those in an existing anti-poverty program are affiliated automatically. Finally, individuals in control clusters are legally entitled to travel to treated clusters to affiliate if they wish, but

they would then have to wait thirty days before traveling back for medical care, and so this last form of noncompliance is likely to be less of an issue.

In fact, cluster-randomized trials often have imperfect treatment compliance problems at the unit level. For many studies, this problem may even describe the essence of the reason for the cluster-randomized design in the first place. For example, in SPS, randomly assigning the program to different individuals was politically infeasible because their family members, neighbors, and coworkers, who might have been assigned to the control group, would have objected that they did not have the same rights. So at the individual level, we were unable to make random assignments, but even if we could interference among the individuals would have lead to serious noncompliance problems. In contrast, in an experiment where individuals were isolated from each other, interference that generated noncompliance would not be an issue but then individual-level randomization would likely be politically feasible.

Because of this close connection between cluster-randomization and unit-noncompliance problems, most analyses of this type of data need to take compliance into account. Thus, we consider an extension of our approach to cluster-randomized trials under the *matched-pair cluster randomized encouragement design*, where the encouragement to receive a treatment, rather than the receipt of the treatment itself, is randomized at the cluster-level.

Angrist *et al.* (1996) showed how an instrumental variable method can be used to analyze unit-randomized experiments with noncompliance under the completely randomized design. Here, we extend their approach to analyze cluster-randomized experiments with unit-level noncompliance under the matched-pair design. To compliment the parametric Bayesian approach to this problem (under the completely randomized design) by Frangakis *et al.* (2002), we consider a nonparametric analysis based on the estimator introduced in Section 4. We now discuss design and notation, new quantities of interest, and their estimators.

6.1 Design and Notation

The setup is the same as described in Section 3.2 except that T_{jk} now represents whether the jth cluster in the kth matched-pair was encouraged to receive the treatment rather than whether it received the treatment. Recall that $T_{1k} = Z_k$ and $T_{2k} = 1 - Z_k$. Now, let $R_{ijk}(T_{jk})$ represent the potential treatment receipt indicator variables for the ith unit in the jth cluster of the kth pair under the encouragement ($T_{jk} = 1$) and control ($T_{jk} = 0$) conditions. The observed treatment variable is, then, defined

as $R_{ijk} \equiv T_{jk}R_{ijk}(1) + (1 - T_{jk})R_{ijk}(0)$. Similar to the potential outcomes, these potential treatment variables depend on the cluster-level encouragement variable rather than the unit-level encouragement variable, requiring a different interpretation of the resulting causal effects. Finally, we write the potential outcomes as functions of randomized encouragement and actual receipt of treatment, $Y_{ijk}(R_{ijk}, T_{jk})$. This formulation makes the following assumption, which is an extension of Assumption 1, given in Section 3.2:

Assumption 3 (No Interference between Units) *Let $R_{ijk}(\mathbf{T})$ be the potential outcomes for the ith unit in the jth cluster of the kth matched-pair where $\mathbf{T}$ is an $(m \times 2)$ matrix whose (j, k) element is T_{jk}. Furthermore, let $Y_{ijk}(\mathbf{R}, \mathbf{T})$ be the potential outcomes for the ith unit in the jth cluster of the kth matched-pair where $\mathbf{R}$ is an $(n_{jk} \times m \times 2)$ ragged array whose (i, j, k) element is R_{ijk}. Then,*

1. *If $T_{jk} = T'_{jk}$, then $R_{ijk}(\mathbf{T}) = R_{ijk}(\mathbf{T}')$.*
2. *If $T_{jk} = T'_{jk}$ and $R_{ijk} = R'_{ijk}$, then $Y_{ijk}(\mathbf{R}, \mathbf{T}) = Y_{ijk}(\mathbf{R}', \mathbf{T}')$.*

In other words, this assumption requires that one person's decision to comply has no effect on any other person's outcomes within the same cluster, and as such the requirements are more demanding than for the ITT effects in previous sections of this paper. For example, this assumption might be violated for certain health outcomes in the SPS evaluation: If all of one's neighbors comply with encouragement to affiliate to SPS, the vaccines and health care they receive may reduce the prevalence of infectious diseases and so might independently improve that person's health outcomes. However, this is the case only if the noncompliers would not have received health care in absence of affiliation with SPS; if instead, as is likely the case, most are relatively wealthy citizens who have other forms of health care, then interference is not likely to be much of an issue. For another example, financial outcomes would likely also satisfy this assumption unless neighborhoods self-insure; however, a neighborhood with enough social capital to make produce institutions to reliably self-insure would probably also have much higher degrees of compliance.

The no interference assumption allows us to write $R_{ijk}(\mathbf{T}) = R_{ijk}(T_{jk})$ and $Y_{ijk}(\mathbf{R}, \mathbf{T}) = Y_{ijk}(R_{ijk}, T_{jk})$. Since $T_{1k} = Z_k$ and $T_{2k} = 1 - Z_k$, both $R_{ijk}(T_{jk})$ and $Y_{ijk}(T_{jk})$ depend on Z_k alone.

Extending the framework of Angrist *et al.* (1996) to cluster-randomized trials, we assume the exclusion restriction so that cluster-level encouragement affects the unit-level outcome only through the unit-level receipt of the treatment.

Assumption 4 (Exclusion Restriction) *$Y_{ijk}(r, 0) = Y_{ijk}(r, 1)$ for $r = 0, 1$ and all i, j, and k.*

These assumptions together simplify the problem by enabling us to write the potential outcomes as functions of T_{jk} (or Z_k) alone, i.e., $Y_{ijk}(R_{jk}, T_{jk}) = Y_{ijk}(T_{jk})$.

Finally, following Angrist *et al.* (1996), we call the units with $R_{ijk}(T_{jk}) = T_{jk}$ *compliers* (and denote them by $C_{ijk} = c$), those with $R_{ijk}(T_{jk}) = 1$ *always-takers* ($C_{ijk} = a$), those with $R_{ijk}(T_{jk}) = 0$ *never-takers* ($C_{ijk} = n$), and the units with $R_{ijk}(T_{jk}) = 1 - T_{jk}$ *defiers* ($C_{ijk} = d$). The monotonicity assumption is made to exclude the existence of defiers.

Assumption 5 (Monotonicity) *There exists no defier. That is,* $R_{ijk}(1) \geq R_{ijk}(0)$ *holds for all* $i, j,$ *and* k.

For example, in our Mexico evaluation, never-takers are those who would not affiliate with SPS regardless of whether the government encourages them to do so or not. Since SPS was designed for the poor, many wealthy citizens with their own preexisting health care arrangements may well be never-takers. We expect a substantial proportion of the population to qualify as never-takers, and in fact the proportion of never-takers in our sample is estimated to be 56%.

Always-takers are those who would affiliate no matter what, even if they happen to be in a control cluster. These are more uncommon, and would likely be the poor without access to health care who nevertheless have the information and financial resources necessary to travel to the place to sign up for SPS and to travel back to receive care. (The estimated proportion of always-takers is only 7%.) The last possibility are defiers, which should be rare: these are people who would affiliate with SPS if not encouraged to do so but would not affiliate if encouraged. While children and many others are known to rebel in the face of authority, we would expect few if any to do so for a program like SPS.

6.2 Causal Quantities of Interest

We consider the two types of causal quantities of interest in matched-pair cluster-randomized encouragement designs – the intention-to-treat (ITT) effect and the complier average causal effect (CACE) (Angrist *et al.*, 1996). The ITT effect is the average causal effect of encouragement (rather than treatment) and is equivalent to the various versions of the average treatment effect in Section 3.3 (i.e., SATE, CATE, UATE, and PATE, with or without interference).

In contrast, the CACE estimand is the average treatment effect (for SATE, CATE, UATE, or PATE, with or without interference) among compliers only. Note that compliers are not those merely observed to affiliate among those in encouragement clusters and those observed not to affiliate in clusters not

encouraged, since the former includes always-takers and the latter includes never-takers. In addition, groups defined this way would be consequences of the treatment. In contrast, the compliers we seek are those who would comply with randomized encouragement only if they were encouraged and would not comply only if they were not encouraged, and so this group is defined at least in principle prior to treatment but identifying its members is itself an estimation problem. Since our assumptions imply that the causal effect of encouragement on always-takers and never-takers is zero, the causal effect for any given quantity is larger for CACE than for ITT.

6.3 Estimation

If we assume sampling of both pairs of clusters and units within each cluster, then the ITT causal effect can be defined as ψ_P in Equation 4. Thus, $\hat{\psi}(N_{1k} + N_{2k})$ can be used to estimate this ITT effect, and the unbiased estimation of its variance is possible using the results given in Section 4.3.

Next, we consider *population* CACE. Under the assumption of simple random sampling of both clusters and units within each cluster, this estimand can be defined as:

$$\gamma \equiv \mathbb{E}_{\mathcal{P}}(Y(1) - Y(0) \mid C = c) = \frac{\mathbb{E}_{\mathcal{P}}[Y(1) - Y(0)]}{\mathbb{E}_{\mathcal{P}}[R(1) - R(0)]}, \tag{19}$$

where the equality follows from the direct application of the argument of Angrist *et al.* (1996) to cluster-randomized trials under the assumptions stated above. If we only assume the simple random sampling of clusters as in the case of UATE, then the expectation in Equation 19 is taken with respect to the set $\mathcal{U}$ rather than $\mathcal{P}$.

Thus, the instrumental variable estimator based on the general weighted estimator in Equation 5 is given by:

$$\hat{\gamma}(w_k) \equiv \frac{\hat{\psi}(w_k)}{\hat{\tau}(w_k)}, \tag{20}$$

where $\hat{\tau}(w_k)$ is the estimator of the ITT effect on the receipt of the treatment:

$$\hat{\tau}(w_k) \equiv \frac{1}{\sum_{k=1}^{m} w_k} \sum_{k=1}^{m} w_k \left\{ Z_k \left(\frac{\sum_{i=1}^{n_{1k}} R_{i1k}}{n_{1k}} - \frac{\sum_{i=1}^{n_{2k}} R_{i2k}}{n_{2k}} \right) + (1 - Z_k) \left(\frac{\sum_{i=1}^{n_{2k}} R_{i2k}}{n_{2k}} - \frac{\sum_{i=1}^{n_{1k}} R_{i1k}}{n_{1k}} \right) \right\}.$$

When matching is effective or when cluster sizes are equal within each matched-pair, this estimator is consistent and approximately unbiased. Using the Taylor series expansion, the variance of this estimator

can be approximated by:

$$\begin{aligned}\mathrm{Var}_{apu}(\hat{\gamma}(w_k)) \;\approx\; & \frac{1}{\{E_{apu}(\hat{\tau}(w_k))\}^4}\Big[\{E_{apu}(\hat{\tau}(w_k))\}^2\mathrm{Var}_{apu}(\hat{\psi}(w_k)) + \{E_{apu}(\hat{\psi}(w_k))\}^2\mathrm{Var}_{apu}(\hat{\tau}(w_k)) \\ & -2E_{apu}(\hat{\psi}(w_k))E_{apu}(\hat{\tau}(w_k))\mathrm{Cov}_{apu}(\hat{\psi}(w_k),\hat{\tau}(w_k))\Big], \qquad (21)\end{aligned}$$

where if simple random sampling of pairs of clusters alone is assumed, then the subscript "*apu*" (for assignment, pairs, and units) is replaced with "*ap*". Furthermore, the argument given in Section 4.3 implies, for example, that the variance of $\hat{\gamma}(\tilde{w}_k)$ for estimating the *sample* CACE is on average less than the variance for the *population* CACE given in Equation 21.

Finally, to estimate the variance consistently, Proposition 3 has already shown how to estimate $\mathrm{Var}_{apu}(\hat{\psi}(w_k))$ and $\mathrm{Var}_{apu}(\hat{\tau}(w_k))$ (or $\mathrm{Var}_{ap}(\hat{\psi}(w_k))$ and $\mathrm{Var}_{ap}(\hat{\tau}(w_k))$) without bias. Thus, we only need to know how to estimate the covariance between $\hat{\psi}(w_k)$ and $\hat{\tau}(w_k))$) from the observed data. Using the normalized weights $\tilde{w}_k$, Appendix E proves that the following estimator is unbiased for both $\mathrm{Cov}_{apu}(\hat{\psi}(w_k),\hat{\tau}(w_k))$ and $\mathrm{Cov}_{pu}(\hat{\psi}(w_k),\hat{\tau}(w_k))$ under their respective sampling assumptions:

$$\begin{aligned}&\hat{\nu}(\tilde{w}_k) \\ &\equiv \frac{m}{(m-1)n^2}\sum_{k=1}^{m}\left[\tilde{w}_k\left\{Z_k\left(\frac{\sum_{i=1}^{n_{1k}}Y_{i1k}}{n_{1k}}-\frac{\sum_{i=1}^{n_{2k}}Y_{i2k}}{n_{2k}}\right)+(1-Z_k)\left(\frac{\sum_{i=1}^{n_{2k}}Y_{i2k}}{n_{2k}}-\frac{\sum_{i=1}^{n_{1k}}Y_{i1k}}{n_{1k}}\right)\right\}-\frac{n\hat{\psi}(\tilde{w}_k)}{m}\right] \\ &\qquad\times\left[\tilde{w}_k\left\{Z_k\left(\frac{\sum_{i=1}^{n_{1k}}R_{i1k}}{n_{1k}}-\frac{\sum_{i=1}^{n_{2k}}R_{i2k}}{n_{2k}}\right)+(1-Z_k)\left(\frac{\sum_{i=1}^{n_{2k}}R_{i2k}}{n_{2k}}-\frac{\sum_{i=1}^{n_{1k}}R_{i1k}}{n_{1k}}\right)\right\}-\frac{n\hat{\tau}(\tilde{w}_k)}{m}\right].\end{aligned}$$

7 SPS Evaluation

In this section, we present an analysis of the causal effect of SPS on the probability of a household suffering catastrophic health expenditures — that is, out-of-pocket health care expenditures totaling more than 30% of a household's annual post-subsistence (or disposable) income. The reduction in this probability is a major aim of the SPS program. As nearly 10% of households suffer catastrophic health expenditures in a year, it is easy to see why this would be a priority of the country.

7.1 Estimated Causal Effects

For our present purposes, we give causal effect estimates of SPS on catastrophic expenditures for the four target population quantities of interest (SATE, CATE, UATE, and PATE). We compute these estimates both for the intention to treat (ITT) effect of encouragement to affiliate as well as the average causal

effect among compliers (CACE). Although in most applications, substantive interest would narrow this list to one or a few of these quantities, for our methodological purposes we present all eight estimates (and standard errors) in Table 3.

		SATE	CATE	UATE	PATE
All	ITT	$-.014$ $(\leq .007)$	$-.023$ $(\leq .015)$	$-.014$ $(.007)$	$-.023$ $(.015)$
	CACE	$-.038$ $(\leq .018)$	$-.064$ $(\leq .024)$	$-.038$ $(.018)$	$-.064$ $(.024)$
Male-Headed	ITT	$-.016$ $(\leq .008)$	$-.025$ $(\leq .018)$	$-.016$ $(.008)$	$-.025$ $(.018)$
	CACE	$-.042$ $(\leq .020)$	$-.070$ $(\leq .031)$	$-.042$ $(.020)$	$-.070$ $(.031)$

Table 3: Estimates of Eight Causal Effect of SPS on the Probability of Catastrophic Health Expenditures for all households and male-headed households (standard errors in parentheses)

A table like this will always have some of the same features, no matter what variable is analyzed. Recall, for example, that point estimates of SATE and UATE are the same, as are CATE and PATE. In addition, standard error estimates of UATE and PATE also happen to be the sharp upper bound of the standard errors for SATE and CATE, respectively. In addition, as expected, CACE estimates are always larger than those for ITT.

For the specific estimates in Table 3, consider first the two top lines of the table corresponding to all households. For these data, the CACE estimates are about about 2.7 times larger than that for ITT. The large difference is because of all the people who had preexisting health care and so were largely never-takers. Overall, these results indicate that SPS was clearly successful in reducing the most devastating type of medical expenditures: In other words, the money did make it to the people. The differences among the columns indicate that the average causal effect of encouragement to affiliate to SPS (the ITT effect) is somewhat larger in the population of individuals represented by our sample (-0.023) than among the individuals we directly observe (-0.014). The same is also true among compliers, but at a higher level (-0.038 vs. -0.064).

Substantively, these numbers are quite large. Since those who suffer from catastrophic health expenditures are mostly the poor without access to health insurance, they are likely to be disproportionately represented among compliers as compared to the wealthy with preexisting health care arrangements. As such, this analysis indicates that the causal effect of rolling out the policy reduces by about 23% the proportion of those who experience catastrophic expenditures (i.e., $-.023$ of the 10% with catastrophic

expenditures).

7.2 Fallacies about the Analysis of Matched-Pair Data

We now turn to the claim that matched-pair designs restrict "prediction models to cluster-level baseline risk factors (for example, cluster size)" (Donner and Klar, 2004). (In private communication, Donner and Klar explained to us that they meant this quote to indicate the straightforward point that cluster-level fixed effects cannot be included in regression models with matched pair designs, even though it has unfortunately been interpreted to mean that prediction models in general cannot include any baseline risk factors.) In fact, results can be analyzed within strata defined by any individual or cluster level variable, so long as it is pre-treatment (for PATE and CATE it also must have known population totals). For example, the bottom two rows of Table 3 repeat the same analysis as the top two rows but only for male-headed households, a variable measured only at the unit-level and used to separate the sample at that level. The results for each quantity of interest in this case appear only slightly larger than for the entire sample. (Detailed analyses of these and other substantive results from the SPS evaluation appear in King et al., In progress.) A matched-pair design does make regression models with fixed effects for clusters unidentified, although substituting in random effects works fine.

Donner and Klar (2004) also claim that matched-pair designs are to be faulted because of their "inability to test for homogeneity" of causal effects across clusters. This claim is also false; matched-pair designs make it possible to estimate the effect of a pre-treatment variable on the causal effect. For example, the causal effect is easy to measure at the pair level by merely taking the difference in means between the two clusters. This may be a noisy estimate if matching is poorly conducted, but it serves as a perfectly acceptable dependent variable for subsequent analyses. We can see how it varies as a function of any variable measured at the unit level and then aggregated to the cluster-pair level, or measured directly at the aggregate level from existing data, such as from an existing census. Of course, hypothesis tests cannot be conducted for the difference between two matched-pairs, but by pooling more pairs we can easily do such tests. For example, since the point of SPS was to help poor families, we could examine whether the causal effect of rolling out SPS on various outcome variables increases as the wealth of an area drops. This can be done by a simple plot of the pair-level causal effect by wealth, or fitting some regression models.

8 Concluding Remarks

The methods developed here are designed for researchers lucky enough to be able to randomize treatment assignment, but stuck because of political or other constraints with having to randomize clusters of individuals rather than the individuals themselves. Field experiments in particular very frequently require cluster-randomization. Individual-level randomization was impossible in our evaluation of the Mexican SPS program; in fact, negotiations with the Mexican government began with the presumption that no type of randomization would be politically feasible, but it eventually concluded by allowing cluster-level randomization to be implemented.

When clusters of individuals are randomized rather than the individuals themselves, the best practice should involve the following steps. First, researchers should choose their causal quantity of interest, such as from among those defined above, a step which has been skipped in most prior work. They should then identify available pre-treatment covariates likely to affect the outcome variable, and if possible pair clusters based on the similarity of these covariates as well as cluster sizes; this step is common but severely underutilized and would have translated into considerable research resources saved and numerous observations gained. Finally, researchers should randomize by choosing one treated and one control cluster within each matched-pair. We have shown here that claims in the literature about problems with matched-pair designs are fallacious, misguided, or irrelevant: when clusters can be matched prior to randomization on cluster sizes and variables that affect the outcome, efficiency will be improved by doing so.

The remarkable advantage of random treatment assignment is that it obviates the need to make some unverifiable assumptions as is necessarily the case in observational work. Although statistical modeling has obviously been highly productive in many areas, adding unnecessary modeling or other assumptions to the analysis of experimental data would seem to violate the raison d'etre of this type of research, as it makes conclusions dependent on unknowns rather than empirical data. We thus extend this advantage of experiments by developing methods for the analysis of matched-pair cluster-randomized experiments that do not add superfluous assumptions about functional forms, distributions, constant treatment effects, homoskedasticity, or equal group sizes. Wherever possible, we have followed Neyman's (1923) idea of basing statistical inference only on the experimental design in order to infer average treatment effects. When feasible, our methods avoid parametric, distributional, or other modeling assumptions. Yet, the resulting calculations are simple, require no simulation or numerical optimization, and for example only

a specific type of weighted mean for the point estimate.

Cornfield (1978, pp.101–2) concludes his now classic study by writing that "Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception,... and should be avoided," and an enormous literature has grown in many fields echoing this warning. We can now add that randomization by cluster without prior construction of matched pairs, when pairing is feasible, is an exercise in self-destruction. Failing to match can greatly and unnecessarily reduce efficiency and power, and thus it is equivalent to discarding a large portion of experimental data. This result should affect practice, especially in literatures like political science where experimental analyses routinely use cluster-randomization but examples of matched-pair designs have almost never been used, as well as community consensus recommendations for best practices in the conduct and analysis of cluster-randomized experiments, which closely follow current methodological literature. These include the extension to the "CONSORT" agreement among the major biomedical journals (Campbell *et al.*, 2004), the Cochrane Collaboration requirements for reviewing research (Higgins and Green, 2006, sec. 8.11.2), the prominent Medical Research Council (2002) guidelines, and the education research What Works Clearinghouse (2006). Each would seem to require modifications in light of the results given here.

A Proof of Proposition 1

This proof uses the calculation similar to the one used in the proof of Proposition 1 of (Author cite redacted). First, we rewrite $\hat{\sigma}(\tilde{w}_k)$ as follows:

$$
\begin{aligned}
& \frac{(m-1)n^2}{m}\,\hat{\sigma}(\tilde{w}_k) \\
= & \sum_{k=1}^{m}\left[\tilde{w}_k\left\{Z_kD_k(1)+(1-Z_k)D_k(0)\right\}-\frac{1}{m}\sum_{k'=1}^{m}\tilde{w}_{k'}\{Z_{k'}D_{k'}(1)+(1-Z_{k'})D_{k'}(0)\}\right]^2 \\
= & \sum_{k=1}^{m}\tilde{w}_k^2\left\{Z_kD_k(1)+(1-Z_k)D_k(0)\right\}^2 \\
& -\frac{1}{m}\sum_{k=1}^{m}\sum_{k'=1}^{m}\tilde{w}_k\tilde{w}_{k'}\{Z_kD_k(1)+(1-Z_k)D_k(0)\}\{Z_{k'}D_{k'}(1)+(1-Z_{k'})D_{k'}(0)\} \\
= & \frac{m-1}{m}\sum_{k=1}^{m}\tilde{w}_k^2\left\{Z_kD_k(1)^2+(1-Z_k)D_k(0))^2\right\}-\frac{1}{m}\sum_{k=1}^{m}\sum_{k'\neq k}\tilde{w}_k\tilde{w}_{k'}\{Z_kZ_{k'}D_k(1)D_{k'}(1) \\
& +Z_k(1-Z_{k'})D_k(1)D_{k'}(0)+(1-Z_k)Z_{k'}D_k(0)D_{k'}(1)+(1-Z_k)(1-Z_{k'})D_k(0)D_{k'}(0)\}.
\end{aligned}
$$

Assumption 2 implies $E_a(Z_k) = 1/2$ and $E_a(Z_k Z_{k'}) = 1/4$ for $k \neq k'$. Thus, taking the expectation with respect to Z_k and rearranging terms, we have:

$$E_a(\hat{\sigma}(\tilde{w}_k)) = \frac{1}{2n^2}\left\{\sum_{k=1}^{m}\tilde{w}_k^2(D_k(1)^2 + D_k(0)^2) - \frac{1}{2(m-1)}\sum_{k=1}^{m}\sum_{k'\neq k}\tilde{w}_k\tilde{w}_{k'}(D_k(1)+D_k(0))(D_{k'}(1)+D_{k'}(0))\right\} \quad (22)$$

Finally, we compare this with the true variance expression in Equation 12.

$$\begin{aligned}
& E_a(\hat{\sigma}(\tilde{w}_k)) - \mathrm{Var}_a(\hat{\psi}(\tilde{w}_k)) \\
= & \frac{1}{4n^2}\left\{\sum_{k=1}^{m}\tilde{w}_k^2\{D_k(1)+D_k(0))\}^2 - \frac{1}{m-1}\sum_{k=1}^{m}\sum_{k'\neq k}\tilde{w}_k\tilde{w}_{k'}(D_k(1)+D_k(0))(D_{k'}(1)+D_{k'}(0))\right\} \\
= & \frac{m}{4(m-1)n^2}\left\{\sum_{k=1}^{m}\tilde{w}_k^2\{D_k(1)+D_k(0))\}^2 - \frac{1}{m}\sum_{k=1}^{m}\sum_{k'=1}^{m}\tilde{w}_k\tilde{w}_{k'}(D_k(1)+D_k(0))(D_{k'}(1)+D_{k'}(0))\right\} \\
= & \frac{m}{4n^2}\mathrm{var}\{\tilde{w}_k(D_k(1)+D_k(0))\}.
\end{aligned}$$

This variance cannot be identified because $D_k(1)$ and $D_k(0)$ are not jointly observed for a given k. Thus, $E_a(\hat{\sigma})$ is the sharp upper bound of $\mathrm{Var}_a(\hat{\psi}(\tilde{w}_k))$. □

B Proof of Proposition 2

First, we derive the lower bound by rewriting the variance in Equation 13 as:

$$\begin{aligned}
& \mathrm{Var}_{au}(\hat{\psi}(\tilde{w}_k)) \\
= & \frac{1}{4n^2}\sum_{k=1}^{m}\tilde{w}_k^2\left\{E_u(D_k(1)^2 + D_k(0)^2) + \mathrm{Var}_u(D_k(1)) + \mathrm{Var}_u(D_k(0)) - 2E_u(D_k(1))E(D_k(0))\right\}, \\
= & \frac{1}{2n^2}\sum_{k=1}^{m}\tilde{w}_k^2\left[\mathrm{Var}_u(D_k(1)) + \mathrm{Var}_u(D_k(0)) + \frac{1}{2}\{E_u(D_k(1) - D_k(0))\}^2\right]. \quad (23)
\end{aligned}$$

Since we do not jointly observe $D_k(1)$ and $D_k(0)$, the third term in Equation 23 is not identifiable. The sharp lower bound of this term is zero, and thus the desired lower bound follows.

To show that this lower bound can be estimated without bias, we write $\hat{\lambda}(\tilde{w}_k)$ using the potential outcome notation:

$$\frac{1}{n^2}\sum_{k=1}^{m}\tilde{w}_k^2\left\{(1-f_{1k})\frac{Z_k\mathrm{var}(Y_{i1k}(1)) + (1-Z_k)\mathrm{var}(Y_{i1k}(0))}{n_{1k}} + (1-f_{2k})\frac{(1-Z_k)\mathrm{var}(Y_{i2k}(1)) + Z_k\mathrm{var}(Y_{i2k}(0))}{n_{2k}}\right\}.$$

Then, taking the expectation with respect to Z_k and then the sampling of units yields:

$$
\begin{aligned}
& E_{au}(\hat{\lambda}(\tilde{w}_k)) \\
= \;& E_u\{E_a(\hat{\lambda}(\tilde{w}_k))\} \\
= \;& \frac{1}{2n^2}\sum_{k=1}^{m}\tilde{w}_k^2 E_u\left\{(1-f_{1k})\frac{\mathrm{var}(Y_{i1k}(1))+\mathrm{var}(Y_{i1k}(0))}{n_{1k}}+(1-f_{2k})\frac{\mathrm{var}(Y_{i2k}(1))+\mathrm{var}(Y_{i2k}(0))}{n_{2k}}\right\} \\
= \;& \frac{1}{2n^2}\sum_{k=1}^{m}\tilde{w}_k^2\left\{\frac{\mathrm{Var}_u(Y_{i1k}(1))+\mathrm{Var}_u(Y_{i1k}(0))}{n_{1k}}+\frac{\mathrm{Var}_u(Y_{i2k}(1))+\mathrm{Var}_u(Y_{i2k}(0))}{n_{2k}}\right\} \\
= \;& \frac{1}{2n^2}\sum_{k=1}^{m}\tilde{w}_k^2\left\{\mathrm{Var}_u(D_k(1))+\mathrm{Var}_u(D_k(0))\right\}.
\end{aligned}
$$

Finally, we derive the upper bound. Applying the law of iterated expectations to Equation 22, we have:

$$
\begin{aligned}
E_{au}(\hat{\sigma}(\tilde{w}_k)) =& \frac{1}{2n^2}\left[\sum_{k=1}^{m}\tilde{w}_k^2 E_u\left\{D_k(1)^2+D_k(0))^2\right\}\right. \\
& \left.-\frac{1}{2(m-1)}\sum_{k=1}^{m}\sum_{k'\neq k}\tilde{w}_k\tilde{w}_{k'}E_u(D_k(1)+D_k(0))E_u(D_{k'}(1)+D_{k'}(0))\right], \qquad (24)
\end{aligned}
$$

where the equality follows from the assumption that sampling of units is done independently across clusters. Together with the expression of $\mathrm{Var}_{au}(\hat{\psi}(\tilde{w}_k))$ given above, this yields:

$$
\begin{aligned}
& E_{au}(\hat{\sigma}(\tilde{w}_k)) - \mathrm{Var}_{au}(\hat{\psi}(\tilde{w}_k)) \\
= \;& \frac{1}{4n^2}\left[\sum_{k=1}^{m}\tilde{w}_k^2\left\{E_u(D_k(1)^2+D_k(0)^2)-\mathrm{Var}_u(D_k(1))-\mathrm{Var}_u(D_k(0))+2E_u(D_k(1))E_u(D_k(0))\right\}\right. \\
& \left.-\frac{1}{m-1}\sum_{k=1}^{m}\sum_{k'\neq k}\tilde{w}_k\tilde{w}_{k'}E_u(D_k(1)+D_k(0))E_u(D_{k'}(1)+D_{k'}(0))\right] \\
= \;& \frac{1}{4n^2}\left[\sum_{k=1}^{m}\tilde{w}_k^2\left[\{E_u(D_k(1))\}^2+\{E_u(D_k(0))\}^2\right]+2E_u(D_k(1))E_u(D_k(0))\right] \\
& -\frac{1}{m-1}\sum_{k=1}^{m}\sum_{k'\neq k}\tilde{w}_k\tilde{w}_{k'}E_u(D_k(1)+D_k(0))E_u(D_{k'}(1)+D_{k'}(0))\Bigg] \\
= \;& \frac{m}{4(m-1)n^2}\left[\sum_{k=1}^{m}\tilde{w}_k^2\{E_u(D_k(1)+D_k(0))\}^2-\frac{1}{m}\sum_{k=1}^{m}\sum_{k'=1}^{m}\tilde{w}_k\tilde{w}_{k'}E_u(D_k(1)+D_k(0))E_u(D_{k'}(1)+D_{k'}(0))\right] \\
= \;& \frac{m}{4(m-1)n^2}\sum_{k=1}^{m}\left\{\tilde{w}_k E_u(D_k(1)+D_k(0))-\frac{1}{m}\sum_{k'=1}^{m}\tilde{w}_k E_u(D_k(1)+D_k(0))\right\}^2 \\
= \;& \frac{m}{4n^2}\mathrm{var}\left\{\tilde{w}_k E_u(D_k(1)+D_k(0))\right\}.
\end{aligned}
$$

Since we do not observe $D_k(1)$ and $D_k(0)$ jointly, this sample variance is not identified. The smallest possible value of this variance is zero, and hence $E_{au}(\hat{\sigma}(\tilde{w}_k))$ is the sharp upper bound of $\mathrm{Var}_{au}(\hat{\psi}(\tilde{w}_k))$. □

C Proof of Proposition 3

Since UATE can be seen as a special case of PATE where all units within each cluster are observed, i.e., $n_{jk} = N_{jk}$, we first derive the variance of $\hat{\psi}(\tilde{w}_k)$ for PATE. Define $\widetilde{D}_k(t) = \tilde{w}_k D_k(t)$, $\tilde{\mu}_k(t) = E_u(\widetilde{D}_k(t))$, and $\tilde{\eta}_k(t) = \mathrm{Var}_u(\widetilde{D}_k(t))$ for $t = 0, 1$. Then, the assumption that the order of clusters within each matched-pair is randomized implies $E_c(\tilde{\eta}_k) = E_c(\tilde{\eta}_k(1)) = E_c(\tilde{\eta}_k(0))$ and $\mathrm{Var}_c(\tilde{\mu}_k) = \mathrm{Var}_c(\tilde{\mu}_k(1)) = \mathrm{Var}_c(\tilde{\mu}_k(0))$. Then, the true variance can be written as:

$$
\begin{aligned}
& \mathrm{Var}_{apu}(\hat{\psi}(\tilde{w}_k)) \\
= \; & \frac{m}{2n^2} E_p\left[\mathrm{Var}_u(\widetilde{D}_k(1)) + \mathrm{Var}_u(\widetilde{D}_k(0)) + \frac{1}{2}\{E_u(\widetilde{D}_k(1) - \widetilde{D}_k(0))\}^2\right] + \frac{m}{4n^2}\mathrm{Var}_p(E_u(\widetilde{D}_k(1) + \widetilde{D}_k(0))), \\
= \; & \frac{m}{4n^2}\left[4E_p(\tilde{\eta}_k) + 2E_p(\tilde{\mu}_k^2) - 2E_p(\tilde{\mu}_k(1)\tilde{\mu}_k(0)) + 2\mathrm{Var}_p(\tilde{\mu}_k) + 2\mathrm{Cov}_p(\tilde{\mu}_k(1), \tilde{\mu}_k(0)))\right], \\
= \; & \frac{m}{n^2}\left\{E_p(\tilde{\eta}_k) + \mathrm{Var}_p(\tilde{\mu}_k)\right\}.
\end{aligned}
$$

When the estimand is UATE, $\tilde{\eta}_k = 0$ for all k since within-cluster means are observed without sampling variability. Thus, $\mathrm{Var}_{apu}(\hat{\psi}(\tilde{w}_k)) = \frac{m}{n^2}\mathrm{Var}_p(\tilde{\mu}_k)$.

Next, we show that $\hat{\sigma}(\tilde{w}_k)$ is an unbiased estimator of the variance by applying the law of iterated expectations to Equation 24:

$$
\begin{aligned}
E_{apu}(\hat{\sigma}(\tilde{w}_k)) &= \frac{m}{2n^2}\left[E_p\{w_k^2 E_u(D_k(1)^2 + D_k(0)^2)\} - \frac{1}{2}[E_p\{w_k E_u(D_k(1) + D_k(0))\}]^2\right] \\
&= \frac{m}{n^2}\left[E_p\left\{w_k^2 E_u(D_k^2)\right\} - \{E_p(\tilde{\mu}_k)\}^2\right] \\
&= \frac{m}{n^2}\left[E_p\{E_u(w_k^2 D_k^2)\} - E_p(\tilde{\mu}_k^2) + \mathrm{Var}_p(\tilde{\mu}_k)\right] \\
&= \frac{m}{n^2}\left[E_p\{\mathrm{Var}_u(\widetilde{D}_k)\} + \mathrm{Var}_p(\tilde{\mu}_k)\right].
\end{aligned}
$$

where $E_u(D_k^2) = E_u(D_k(0)^2) = E_u(D_k(1)^2)$ holds because the order of clusters within each matched-pair is randomized. When PATE is the estimand $\mathrm{Var}_u(D_k) = 0$ for all k since within-cluster means are observed without sampling uncertainty. Thus, $E_{ap}(\hat{\sigma}(\tilde{w}_k)) = \frac{m}{n^2}\mathrm{Var}_p(\tilde{\mu}_k)$. □

D Bias of the Existing Variance Estimator

In this section, we derive the bias of the standard variance estimator used in the literature. To do this, we rewrite $\hat{\delta}(\tilde{w}_k)$ as follows:

$$
\begin{aligned}
&\frac{n^3}{\sum_{k=1}^m \tilde{w}_k^2}\hat{\delta}(\tilde{w}_k) \\
&\qquad = \sum_{k=1}^m \tilde{w}_k \left[Z_k D_k(1) + (1-Z_k) D_k(0) - \frac{1}{n}\sum_{k'=1}^m \tilde{w}_{k'} \{ Z_{k'} D_{k'}(1) + (1-Z_{k'}) D_{k'}(0) \} \right]^2 \\
&\qquad = \sum_{k=1}^m \tilde{w}_k \left[Z_k D_k(1)^2 + (1-Z_k) D_k(0)^2 \right. \\
&\qquad\quad - \frac{2}{n}\sum_{k'=1}^m \tilde{w}_{k'} \{ Z_k D_k(1) + (1-Z_k) D_k(0) \} \{ Z_{k'} D_{k'}(1) + (1-Z_{k'}) D_{k'}(0) \} \\
&\qquad\quad \left. + \frac{1}{n^2}\sum_{k'=1}^m \sum_{k''=1}^m \tilde{w}_{k'}^2 \tilde{w}_{k''}^2 \{ Z_{k'} D_{k'}(1) + (1-Z_{k'}) D_{k'}(0) \} \{ Z_{k''} D_{k''}(1) + (1-Z_{k''}) D_{k''}(0) \} \right],
\end{aligned}
$$

where the final term within the bracket can be written as:

$$
\begin{aligned}
&\frac{1}{n^2}\left[\sum_{k'=1}^m \tilde{w}_{k'}^2 \left\{ Z_{k'} D_{k'}(1)^2 + (1-Z_{k'}) D_{k'}(0)^2 \right\} \right. \\
&\qquad\qquad \left. + \sum_{k'=1}^m \sum_{k''\neq k'} \tilde{w}_{k'} \tilde{w}_{k''} \{ Z_{k'} D_{k'}(1) + (1-Z_{k'}) D_{k'}(0) \} \{ Z_{k''} D_{k''}(1) + (1-Z_{k''}) D_{k''}(0) \} \right].
\end{aligned}
$$

Taking the expectation with respect to Z_k yields:

$$
\begin{aligned}
& E_a(\hat{\delta}(\tilde{w}_k)) \\
= \; & \frac{\sum_{k=1}^m \tilde{w}_k^2}{2n^3} \sum_{k=1}^m \tilde{w}_k \left[\left(1 - \frac{2\tilde{w}_k}{n} \right) (D_k(1)^2 + D_k(0)^2) - \frac{1}{n}\sum_{k'\neq k} \tilde{w}_{k'} (D_k(1) + D_k(0))(D_{k'}(1) + D_{k'}(0)) \right. \\
& \left. + \frac{1}{n^2} \left\{ \sum_{k'=1}^m \tilde{w}_{k'}^2 (D_{k'}(1)^2 + D_{k'}(0)^2) + \frac{1}{2}\sum_{k'=1}^m \sum_{k''\neq k'} \tilde{w}_{k'} \tilde{w}_{k''} (D_{k'}(1) + D_{k'}(0))(D_{k''}(1) + D_{k''}(0)) \right\} \right] \\
= \; & \frac{\sum_{k=1}^m \tilde{w}_k^2}{2n^3} \left[\sum_{k=1}^m \left\{ \left(1 - \frac{2\tilde{w}_k}{n} \right) \tilde{w}_k (D_k(1)^2 + D_k(0)^2) - \frac{1}{n}\sum_{k'\neq k} \tilde{w}_k \tilde{w}_{k'} (D_k(1) + D_k(0))(D_{k'}(1) + D_{k'}(0)) \right\} \right. \\
& \left. + \frac{1}{n} \left\{ \sum_{k=1}^m \tilde{w}_k^2 (D_k(1)^2 + D_k(0)^2) + \frac{1}{2}\sum_{k=1}^m \sum_{k'\neq k} \tilde{w}_k \tilde{w}_{k'} (D_k(1) + D_k(0))(D_{k'}(1) + D_{k'}(0)) \right\} \right] \\
= \; & \frac{\sum_{k=1}^m \tilde{w}_k^2}{2n^3} \sum_{k=1}^m \left\{ \left(1 - \frac{\tilde{w}_k}{n} \right) \tilde{w}_k (D_k(1)^2 + D_k(0)^2) - \frac{1}{2n}\sum_{k=1}^m \sum_{k'\neq k} \tilde{w}_k \tilde{w}_{k'} (D_k(1) + D_k(0))(D_{k'}(1) + D_{k'}(0)) \right\}
\end{aligned}
$$

The comparison of this expression with $E_a(\hat{\sigma}(\tilde{w}_k))$ in Equation 22 shows that they are in general different, and the difference remains even after taking the expectation with respect to the simple random sampling

of pairs of clusters and/or units within clusters. Since $\hat{\sigma}(\tilde{w}_k)$ is an unbiased estimate of the variance for UATE and PATE, this implies that $\hat{\delta}(\tilde{w}_k)$ is generally biased. □

E Unbiased Covariance Estimation

This appendix derives unbiased estimates of $\mathrm{Cov}_{auc}(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k))$ and $\mathrm{Cov}_{ac}(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k))$ using the analytical strategies of the proofs of Propositions 1–3. First, we derive the true covariance between $\hat{\psi}(\tilde{w}_k)$ and $\hat{\tau}(\tilde{w}_k)$. Define $G_k(1) = \sum_{i=1}^{n_{1k}} R_{i1k}(1)/n_{1k} - \sum_{i=1}^{n_{2k}} R_{i2k}(0)/n_{2k}$ and $G_k(0) = \sum_{i=1}^{n_{2k}} R_{i2k}(1)/n_{2k} - \sum_{i=1}^{n_{1k}} R_{i1k}(0)/n_{1k}$. Taking the expectation of with respect to Z_k yields:

$$\mathrm{Cov}_a(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k)) \;=\; \frac{1}{n^2}\sum_{k=1}^{m} \tilde{w}_k^2 (D_k(1) - D_k(0))(G_k(1) - G_k(0)).$$

Then, we have:

$$\begin{aligned}
&\mathrm{Cov}_{ap}(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k)) \\
=\;& E_p\{\mathrm{Cov}_a(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k))\} + \mathrm{Cov}_p\{E_a(\hat{\psi}(\tilde{w}_k)), E_a(\hat{\tau}(\tilde{w}_k))\} \\
=\;& \frac{m}{4n^2}\left[E_p\{(\widetilde{D}_k(1) - \widetilde{D}_k(0))(\widetilde{G}_k(1) - \widetilde{G}_k(0))\} + \mathrm{Cov}_p\{\widetilde{D}_k(1) + \widetilde{D}_k(0), \widetilde{G}_k(1) + \widetilde{G}_k(0)\}\right] \\
=\;& \frac{m}{4n^2}\left[E_p(\widetilde{D}_k(1)\widetilde{G}_k(1) + \widetilde{D}_k(0)\widetilde{G}_k(0)) - E_p\{(\widetilde{D}_k(1) + \widetilde{D}_k(0))\}E_p\{(\widetilde{G}_k(1) + \widetilde{G}_k(0))\}\right] \\
=\;& \frac{m}{n^2}\mathrm{Cov}_p(\widetilde{D}_k, \widetilde{G}_k),
\end{aligned}$$

where $\widetilde{G}_k(t) = \tilde{w}_k G_k(t)$ for $t = 0, 1$, and the last equality follows from the fact that $E_p(\widetilde{D}_k) = E_p(\widetilde{D}_k(t))$, $E_p(\widetilde{G}_k) = E_p(\widetilde{G}_k(t))$ and $E_p(\widetilde{D}_k\widetilde{G}_k) = E_p(\widetilde{D}_k(t)\widetilde{G}_k(t))$ for $t = 0, 1$. Similarly,

$$\begin{aligned}
&\mathrm{Cov}_{au}(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k)) \\
=\;& E_u\{\mathrm{Cov}_a(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k))\} + \mathrm{Cov}_u\{E_a(\hat{\psi}(\tilde{w}_k)), E_a(\hat{\tau}(\tilde{w}_k))\} \\
=\;& \frac{1}{2n^2}\sum_{k=1}^{m}\Big[\mathrm{Cov}_u(\widetilde{D}_k(1), \widetilde{G}(1)) + \mathrm{Cov}_u(\widetilde{D}_k(0), \widetilde{G}_k(0)) \\
&\qquad + \frac{1}{2}\{E_u(\widetilde{D}_k(1)) - E_u(\widetilde{D}_k(0))\}\{E_u(\widetilde{G}_k(1)) - E_u(\widetilde{G}_k(0))\}\Big].
\end{aligned}$$

And thus,

$$\begin{aligned}
\mathrm{Cov}_{apu}(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k)) \;=\;& \frac{m}{n^2}\Big[\mathrm{Cov}_u(\widetilde{D}_k, \widetilde{G}_k) + \frac{1}{4}E_p\{E_u(\widetilde{D}_k(1)) - E_u(\widetilde{D}_k(0))\}\{E_u(\widetilde{G}_k(1)) - E_u(\widetilde{G}_k(0))\} \\
&\qquad + \frac{1}{4}\mathrm{Cov}_p\{E_u(\widetilde{D}_k(1) + \widetilde{D}_k(0)), E_u(\widetilde{G}_k(1) + \widetilde{G}_k(0))\}\Big] \\
=\;& \frac{m}{n^2}\left[E_p\{\mathrm{Cov}_u(\widetilde{D}_k, \widetilde{G}_k)\} + \mathrm{Cov}_p\{E_u(\widetilde{D}_k), E_u(\widetilde{G}_k)\}\right].
\end{aligned}$$

Then, calculations analogous to the ones above shows that $E_{ap}(\hat{\nu}(\tilde{w}_k)) = \mathrm{Cov}_{ap}(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k))$ and $E_{apu}(\hat{\nu}(\tilde{w}_k)) = \mathrm{Cov}_{apu}(\hat{\psi}(\tilde{w}_k), \hat{\tau}(\tilde{w}_k))$. □

References

Angrist, J. and Lavy, V. (2002). The effect of high school matriculation awards: Evidence from randomized trials. Working Paper 9389, National Bureau of Economic Research.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* **91**, 434, 444–455.

Arceneaux, K. (2005). Using cluster randomized field experiments to study voting behavior. *The Annals of the American Academy of Political and Social Science* **601**, 1, 169–179.

Ball, S. and Bogatz, G. A. (1972). Reading with television: An evaluation of the electric company. Tech. Rep. PR-72-2, Educational Testing Service, Princeton, N.J.

Bloom, H. S. (2006). The core analytics of randomized experiments for social research. Tech. rep., MDRC.

Box, G. E., Hunger, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. Wiley-Interscience, New York.

Braun, T. M. and Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association* **96**, 456, 1424–1432.

Campbell, M., Elbourne, D., and Altman, D. (2004). CONSORT statement: extension to cluster randomised trials. *BMJ* **328**, 7441, 702–708.

Campbell, M., Mollison, J., and Grimshaw, J. (2001). Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. *Statistics in Medicine* **20**, 3, 391–399.

Campbell, M. J. (2004). Editorial: Extending consort to include cluster trials. *BMJ* **328**, 654–655. http://www.bmj.com/cgi/content/full/328/7441/654

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology* **108**, 2, 100–102.

Cox, D. R. (1958). *Planning of Experiments*. John Wiley & Sons, New York.

Donner, A. (1987). Statistical methodology for paired cluster designs. *American Journal of Epidemiology* **126**, 5, 972–979.

Donner, A. (1998). Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics* **47**, 1, 95–113.

Donner, A. and Donald, A. (1987). Analysis of data arising from a stratified design with the cluster as unit of randomization. *Statistics in Medicine* **6**, 43–52.

Donner, A. and Hauck, W. (1989). Estimation of a common odds ration in paired-cluster randomization designs. *Statistics in Medicine* **8**, 599–607.

Donner, A. and Klar, N. (1993). Confidence interval construction for effect measures arising from cluster randomization trials. *Journal of Clinical Epidemiology* **46**, 2, 123–131.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Oxford University Press, New York.

Donner, A. and Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health* **94**, 3, 416–422.

Feng, Z., Diehr, P., Peterson, A., and McLerran, D. (2001). Selected statistical issues in group randomized trials. *Annual Review of Public Health* **22**, 167–187.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, London.

Frangakis, C. E., Rubin, D. B., and Zhou, X.-H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms (with discussion). *Biostatistics* **3**, 2, 147–164.

Frenk, J., Sepúlveda, J., Gómez-Dantés, O., and Knaul, F. (2003). Evidence-based health policy: three generations of reform in Mexico. *The Lancet* **362**, 9396, 1667–1671.

Gail, M. H., Byar, D. P., Pechacek, T. F., and Corle, D. K. (1992). Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT). *Controlled Clinical Trials* **13**, 16–21.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* **15**, 11, 1069–1992.

Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics* **5**, 2, 263–275.

Hayes, R. and Bennett, S. (1999). Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* **28**, 319–326.

Higgins, J. and Green, S., eds. (2006). *Cochrane Handbook for Systematic Review of Interventions 4.2.5 [updated September 2006]*, no. 4 in The Cochrane Library, Chichester, UK. John Wiley and Sons.

Hill, J. L., Rubin, D. B., and Thomas, N. (1999). *Research Designs: Inspired by the Work of Donald Campbell (eds. Leonard Bickman)*, chap. The Design of the New York School Choice Scholarship Program Evaluation, 155–180. Sage, Thousand Oaks.

Imai, K. (2007). experiment: R package for designing and analyzing randomized experiments. available at The Comprehensive R Archive Network (CRAN). `http://cran.r-project.org`.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* **86**, 1, 4–29.

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

Klar, N. and Donner, A. (1997). The merits of matching in community intervention trials: A cautionary tale. *Statistics in Medicine* **16**, 15, 1753–1764.

Klar, N. and Donner, A. (1998). Author's reply. *Statistics in Medicine* **17**, 18, 2151–2152.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press, New York.

Martin, D. C., Diehr, P., Perrin, E. B., and Koepsell, T. D. (1993). The effect of matching on the power of randomized community intervention studies. *Statistics in Medicine* **12**, 329–338.

Medical Research Council (2002). Cluster randomized trials: Methodological and ethical considerations. Tech. rep., MRC Clinical Trials Series. http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002406.

Murray, D. M. (1998). *Design and Analysis of Community Trials*. Oxford University Press, Oxford.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science* **5**, 465–480.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster-randomized trials. *Psychological Methods* **2**, 2, 173–185.

Raudenbush, S. W., Martinez, A., and Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis* **29**, 5–29.

Rubin, D. B. (1990). Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5**, 472–480.

Small, D., Ten Have, T., and Rosenbaum, P. (In-press). Randomization inference in a group-randomized trial of treatments for depression: covariate adjustment, noncompliance and quantile effects. *Journal of the American Statistical Association* .

Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods*. Iowa State University Press, Ames, Iowa, 8th edn.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association* **101**, 476, 1398–1407.

Sommer, A., Djunaedi, E., Loeden, A. A., Tarwotjo, I. J., West, K. P., and Tilden, R. (1986). Impact of vitamin A supplementation on childhood mortality, a randomized clinical trial. *Lancet* **1**, 1169–1173.

Thompson, S. G. (1998). Letter to the editor: The merits of matching in community intervention trials: a cautionary tale by N. Klar and A. Donner. *Statistics in Medicine* **17**, 18, 2149–2151.

Turner, R. M., White, I. R., and Croudace, T. (2007). Analysis of cluster-randomized cross-over data. *Statistics in Medicine* **26**, 274–289.

Varnell, S., Murray, D., Janega, J., and Blitstein, J. (2004). Design and Analysis of Group-Randomized Trials: A Review of Recent Practices. *American Journal of Public Health* **93**, 9, 393–399.

What Works Clearinghouse (2006). Evidence standards for reviewing studies. Tech. rep., Institute for Educational Sciences. http://www.whatworks.ed.gov/reviewprocess/standards.html.