Aggression and Violent Behavior 74 (2024) 101889

Contents lists available at ScienceDirect

Aggression and Violent Behavior

journal homepage: www.elsevier.com/locate/aggviobeh

Is it time for the use of pair-matching in all randomized controlled trials of crime and violence prevention? A review of the research

Heather Paterson [a], Brandon C. Welsh [a,b,*]

[a] Northeastern University, Boston, MA, USA
[b] Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Keywords:
Randomized controlled trial
Pair-matching
Evaluation research
Crime and violence prevention
Public policy

ABSTRACT

Pair-matching in randomized controlled trials (RCTs) has received increased attention in criminology, the social sciences more generally, and medicine and public health, with a growing body of research demonstrating the design's benefits over "simple" RCTs. We carry out a review of matched-pair RCTs compared with simple RCTs to address a somewhat provocative yet fair question for evaluation research on crime and violence prevention interventions: Is it time for the use of pair-matching in all RCTs? At the heart of this question is the ability of the design to most efficiently and robustly compare like with like, thereby, improving confidence in observed effects of intervention trials. Several key findings emerge from the review. First, it is inadequate to examine or discuss RCTs as a single, uniform evaluation design. Here, the key organizing construct is the unit of allocation: individuals; groups of individuals (or clusters); and geographical places. Second, the advantages vastly outweigh the disadvantages for the use of matched-pair RCTs compared to simple RCTs, and most of the advantages hold for all three units of allocation. Third, pair-matching can be used with rather small samples (≥6 units) in cluster-based trials without compromising statistical power or degrees of freedom; less is known about individual- and place-based trials. Fourth, pair-matching cannot be used with some types of RCTs (e.g., cross-over) and is less amenable in other contexts (e.g., RCTs that enroll and randomize individuals on a rolling basis). Implications for evaluation research and public policy are discussed.

1. Introduction

In criminology and in the social sciences more generally, critiques of the randomized controlled trial (RCT) come in a variety of flavors, some having more merit than others. They range from a demotion in the evaluation hierarchy owing to weakened internal validity (it is the "bronze standard;" Berk, 2005), to methodological curiosities and overreach (Sampson, 2010), to the ethics of use with some populations or under some conditions (Boruch et al., 2000), to the need to not overlook other rigorous evaluation designs—namely, that high-quality quasi-experiments can also offer a robust measure of causal inference (Nagin & Sampson, 2019; Nagin & Weisburd, 2013). Similar critiques have been leveled at the RCT in medicine and public health (see Bothwell et al., 2016; Smith & Pell, 2003).

These critiques are often directed at the RCT as if it were a single, uniform evaluation design, failing to recognize variability across a range of methodological and operational features. Some of these features include allocation techniques (e.g., alternate, random), unit of allocation (i.e., individuals, groups of individuals, and geographical places), use of one or multiple treatment arms, efficacy or effectiveness trials (Fagan et al., 2019), efforts taken (prospectively and after-the-fact) to mitigate contamination of the control or other treatment conditions, and single or multi-site implementation (with the latter offering a measure of external validity). Acknowledging this variability and examining the RCT with more precision is also crucial to the current article.

Curiously, the body of criticism seems to have had no dampening effect on what has been a rather steady growth of RCTs in criminology over the last several decades (see Braga et al., 2014; Farrington & Welsh, 2006; Mazerolle et al., 2023; Neyroud, 2017). It is important to acknowledge that there have been many writings in defense of the RCT (e.g., Farrington, 2013; Sherman, 2010; Weisburd, 2010; Weisburd & Hinkle, 2014). Whether in defense of the RCT or not, it is fair to ask if the evaluation design can be improved upon. The purpose could not be more pressing: to demonstrate unambiguously that an intervention produced the reported effect on an outcome. In no uncertain terms, this is the

[*] Corresponding author at: School of Criminology and Criminal Justice, Northeastern University, Churchill Hall, 360 Huntington Avenue, Boston, MA 02115, USA.
E-mail address: b.welsh@northeastern.edu (B.C. Welsh).

https://doi.org/10.1016/j.avb.2023.101889
Received 17 April 2023; Received in revised form 9 August 2023; Accepted 17 October 2023
Available online 19 October 2023

central focus of pair-matching in combination with random allocation—otherwise known as a matched-pair RCT. Here, units (individuals, groups of individuals, or geographical places) are matched by pairs on specific covariates and units of each pair are randomly allocated to treatment and control conditions.

The combination of pair-matching and random allocation in prospective clinical trials was first used in 1926 in medicine—in a trial of Sanocrysin as a therapeutic for pulmonary tuberculosis (Amberson et al., 1931)—and in 1935 in criminology—in the Cambridge-Somerville Youth Study (CSYS; de Q. Cabot, 1940; Powers & Witmer, 1951). In recent decades, pair-matching in RCTs has received increased attention in criminology, the social sciences more broadly, and medicine and public health, with a growing body of research demonstrating the design's benefits over "naïve" or "simple" RCTs[1]—and with different applications and units of analysis (see Ariel & Farrington, 2010; Balzer et al., 2015; Chondros et al., 2021; Imai et al., 2009a, 2009b; Weisburd & Gill, 2014).

The main aim of this article is to carry out a review of matched-pair RCTs compared with simple RCTs. In doing so, it seeks to address a somewhat provocative yet fair question for evaluation research on crime and violence prevention interventions: Is it time for the use of pair-matching in all RCTs? By all, we mean for the different types of units of allocation: individuals; groups of individuals (or clusters); and geographical places. We are interested in explicating and assessing the advantages and disadvantages of the design, as well as understanding if advantages and disadvantages differ for the different types of units of allocation. In addition, we are interested in examining the minimum sample size (*N*) threshold for the different applications of the design. This is in recognition that a small *N* is a key motivating factor for pair-matching in some RCTs (see Lipsey, 1990). In keeping with the literature, minimum sample size refers to the total number of matched pairs (and only the matched pairs) in a RCT. This applies to each of the different types of units of allocation: individuals; clusters; and geographical places.

Also important to this review is that it draws on the body of knowledge on pair-matching in RCTs from a number of disciplines: criminology; the social sciences more broadly; and medicine and public health. We deemed it necessary to consider the literature in medicine and public health because of the rigorous research on and debate about the subject (see e.g., Imai et al., 2009a, 2009b; Balzer et al., 2015; Chondros et al., 2021; also see Podolsky et al., 2021; Welsh et al., 2022), as well as the relevance this literature holds for evaluations of interventions with criminological outcomes.

The article is divided into five parts. The second part provides some background on the process of pair-matching and key advantages and disadvantages of the use of matched-pair RCTs. The third part documents the review's methodology. The fourth part reports on the results of our assessment of matched-pair RCTs compared with simple RCTs. The fifth part discusses implications for evaluation research and public policy.

2. Background

2.1. *Process of pair-matching with random allocation*

Matched-pair RCT designs begin with matching units (individuals, groups of individuals, or geographic places) by pairs based on similar pre-determined characteristics, known as covariates. This involves using covariates that are of most importance to the research question under study, and this information can be obtained from a wide range of sources, including police reports, medical records, and through observation.

In a large-scale evaluation of Communities That Care, known as the Community Youth Development Study, Hawkins et al. (2008) used pair-matching followed by random allocation as a way to improve like-with-like comparisons between the treatment and control conditions. This allowed for a more efficient measure of the impact of the prevention intervention on youth behavioral and health outcomes. Across seven states, 24 small, rural communities (average population = 14,646) were recruited and matched by pairs based on "population size, racial and ethnic diversity, economic indicators, and crime rates" (Hawkins et al., 2008, p. 6). One community in each pair was then randomly assigned by coin toss to receive the prevention intervention; control communities received usual services.

2.2. *Key advantages and disadvantages of pair-matching with random allocation*

The social sciences and medicine and public health literatures draw attention to a number of key advantages and disadvantages of matched-pair RCTs. One of the most fundamental advantages has to do with mitigating covariate imbalance. While random allocation is designed to help eliminate confounding, covariate imbalance is still possible, meaning that the treatment and control groups may still differ by chance. This can be especially problematic in small *N* studies. In fact, a small *N* is a key motivator in choosing pair-matching in cluster- and place-based RCTs, due to the difficulty in recruiting large numbers of schools, communities, or high-crime properties, for example.

Pair-matching prior to random allocation can also improve study power when the analysis recognizes the matching and the matching is effective, meaning that there is a positive within-pair correlation on variables that are known to impact the outcome (Wacholder & Weinberg, 1982). Greater study power means that researchers are more likely to accurately reject the null hypothesis, given that there is a true intervention effect. By decreasing variation within matched pairs on known covariates that are correlated to the outcome, matching can improve the precision of estimated treatment effects compared to simple randomization. Another advantage of this technique is that it allows researchers to reduce the loss of statistical efficiency when randomizing clusters. An estimator or experimental design with greater efficiency needs fewer observations to produce a more precise result. In simulations using data from the Universal Mexican Health Insurance Evaluation (Seguro Popular de Salud), a matched-pair, cluster randomization design greatly increased efficiency and reduced statistical error compared to an unmatched, cluster-randomization design (Imai et al., 2009a; see below for more details).

Pair-matching with random allocation also provides a straightforward way to deal with differential attrition, which can present a serious threat to the internal validity of follow-up assessments of prospective trials. Here, the researcher can drop both members of the pair in the event one member is missing. A key concern is that this can result in considerable study attrition and lead to a smaller final *N*. This can be especially problematic in small *N* studies (Donner & Klar, 2004). Another concern is that cluster randomized trials often have imperfect treatment compliance among the individuals that comprise each cluster. For trials that allow non-compliance, like the Universal Mexican Health Insurance Evaluation, intention-to-treat (ITT) estimates can be used to account for noncompliance and estimate the effect of the program only for individuals who follow the experimental protocol. An ITT analysis estimates the causal effects of encouragement to affiliate with an intervention rather than the effect of the intervention itself (Imai et al., 2009a). Compared to individual-based RCTs, application of the ITT

[1] Weisburd and Gill (2014) refer to this type of RCT as the most common design, whereby, the total sample is randomly allocated to treatment and control conditions without restrictions. They add: "Naïve randomization relies on the assumption that there are no systematic reasons for the treatment and control subjects to differ (since every subject had an equal probability of assignment to each condition), a key *raison d'être* for experimental studies in the first place. But it does not guarantee equivalence, simply that there is no reason for non-equivalence" (pp. 100–101, emphasis in original).

principle is more difficult to adhere to in cluster-based trials, because the loss of an entire cluster and its pair can potentially impact statistical power/efficiency. Some proposed remedies for this issue include using propensity score or missing data methods to account for missing outcome data at the individual or cluster level, or only randomizing clusters where the first participant is included to prevent empty clusters (DeSantis et al., 2020).

Other disadvantages of this design include its costly and labor-intensive nature. Study structure and pace of enrollment (e.g., on a rolling basis) may create logistical difficulties that complicate the process of effective matching and, ultimately, result in less analytical flexibility (Campbell et al., 2007). Here again, small *N* studies are more likely to suffer from ineffective matching and poor statistical power, as sample size is a key element of this calculation.

3. Methods

We carried out a narrative review of the literature on pair-matching with random allocation in prospective controlled trials (otherwise known as matched-pair RCTs) to investigate the advantages and disadvantages of this evaluation design compared to simple RCTs. Reviews of the literature, methodological notes and debates, and empirical analyses of matched-pair RCTs versus simple RCTs were of chief interest. Our focus was on the literature in the following disciplines: criminology; social sciences more generally; and medicine and public health. Compared to the more rigorous systematic review method, the narrative review method is well suited to the aims of the article. The chief reason is that we are not attempting to identify every study that has used this design. Specifically, we searched for and examined studies that focused on the advantages and disadvantages of matched-pair RCTs compared to simple RCTs. Moreover, based on a recent review of historical developments and contemporary uses of matched-pair RCTs spanning almost a full century (Welsh et al., 2022), it was known that the body of evaluation studies that have used this design, albeit increasing in recent decades, is still rather limited.

Two main strategies were used to locate relevant papers: (a) searches of three key electronic bibliographic databases (i.e., Criminal Justice Abstracts, Google Scholar, and Medline) and (b) forward-citation searches of relevant papers. Several general terms were used at the outset of our searches of the databases, including “pair-matching”, “matched pairs”, “matched-pair RCT”, “pair-matching with random allocation”, and “matched-pair randomized controlled trial”, along with other variations of “RCT”. This was followed with more specific search terms, such as “pair-matching in individual-based RCTs”, “pair-matching in cluster-based RCTs”, “matched-pair cluster randomization”, “pair-matching in place-based RCTs”, and “block randomization”. For the forward citation searches, this included relevant papers we had previously collected as part of other projects and new papers deemed relevant that we identified in the search of these databases.

4. Pair-matching in RCTs: the state of research

This section examines the advantages and disadvantages of matched-pair RCTs compared with simple RCTs. We focus specifically on the different types of units of allocation: individuals; groups of individuals (or clusters); and geographical places. In addition, we examine the available research on sample size threshold for the different applications of the design.

4.1. Individual-based trials

Founded by Richard Cabot in 1935, the Cambridge-Somerville Youth Study set out to evaluate the impact of a prevention intervention, known as ‘directed friendship,’ on youth delinquency (Powers & Witmer, 1951). Today, this intervention is more closely associated with mentoring. By the start of the study in June 1939, the sample included 650 underprivileged boys, ages 5 to 13 years (median age = 10.5 years), from Cambridge and Somerville, Massachusetts. A team of psychologists matched all the boys based on 142 variables (rated on an 11-point scale) covering a wide range of characteristics, including physical health, mental health, emotional and social adjustment, aggressiveness, acceptance of authority, and delinquency or disruption at home. This resulted in 325 matched pairs, whom the researchers referred to as “diagnostic twins” (de Q. Cabot, 1940, p. 146). Following the matching process, members of each matched pair were randomly allocated—based on a coin toss—to the treatment and control groups. The sample was later (in 1942) reduced to 253 pairs, a result of the United States’ involvement in World War II and the need for rationing.

The research literature, including the CSYS, demonstrates that pair-matching in individual-based RCTs improves covariate imbalance without jeopardizing study power and sample size (Ariel & Farrington, 2010; Balzer et al., 2015). Compared to cluster- and place-based trials, individual-based trials tend to have larger sample sizes with respect to the number of units available for randomization. This may imply that covariate balance is more likely to be achieved with simple randomization, hence militating against the need for pair-matching. However, there is no guarantee that known, measured covariates will be evenly balanced between treatment and control conditions. It is also the case that pair-matching is only beneficial if the covariates are relevant, both theoretically and empirically, to the outcome being investigated. As stated by Chalmers (1989, p. 27), “depending on the choice of variables used to make the statistical adjustments for imbalances, the likelihood of bias may increase rather than decrease.” This also applies to the other units of allocation.

Ariel and Farrington (2010) note that there is little relationship between large sample size and increased statistical power. In short, more participants can create more covariate diversity, resulting in more variance or “noise.” As explained by Lipsey (1990, p. 138), “The relationship between statistical power and sample size is based less on the total number of subjects involved than on the number in each group or cell within the design. This means that, with regard to statistical power, close attention must be paid to the effect of the number of groups over which subjects are distributed and the proportion of subjects within each group.”

Pair-matching can also improve study efficiency by decreasing the variation in outcome within pairs and, therefore, between treatment and control groups (Balzer et al., 2015). Balzer et al. (2015) illustrate this point by comparing the results of pair-matching for estimating the average treatment effect. In testing the estimators’ performance with >5000 simulated data sets, the authors detected an efficiency gain with the use of pair-matching. They found that pair-matching on three covariates reduced variability in the outcomes within matched pairs. The matched-pair coefficient of variation, which measures the variability in outcomes between units in the absence of intervention, was 0.29 in the first simulation and 0.14 in the second, compared to the unmatched-pair coefficient of variation of 0.53 and 0.27, respectively.

Ariel and Farrington (2010) discuss the concept of “relative efficiency” and provide an equation for researchers to measure the efficiency gain (or lack thereof) of using either a matched-pair (fully blocked design) or partially blocked design compared to simple random allocation. The equation allows for the calculation of the estimated relative efficiency by generating a ratio of the improvement of treatment and control group comparisons, which is highly dependent on the variance of each design. If the relative efficiency ratio is larger than 1, the blocking factor is considered efficient, providing support for the use of a blocked design.

Another advantage of pair-matching in individual-based RCTs is that it provides researchers with a straightforward way to deal with differential attrition as part of longer follow-up assessments. Here, the researcher can drop both members of the pair in the event one member is missing (Donner & Klar, 2004). The downside is that this can result in considerable attrition and small final samples. It is not enough to say that better participant tracking needs to be used to overcome this potential limitation. This is because the ability to maintain a high participant retention rate over time often has to do with a number of factors, including the nature of the participants (e.g., high or low risk), the need for parental consent, and whether data is collected through surveys or administrative records. We return to the importance of differential attrition later in the article.

The combination of pair-matching and random allocation in the CSYS greatly improved study power and, even with the large sample size, mitigated any concerns with covariate imbalance. An important potential limitation of using pair-matching, however, is the loss of degrees of freedom (i.e., the number of variables that are free to vary following one or more restrictions placed on the data). In the case of the CSYS, an unmatched or simple randomized design would have resulted in 504 degrees of freedom for statistical tests of significance ($N_1 + N_2 - 2$). Employing a matched-pair design reduces the degrees of freedom to 252 ($N_{\text{pairs}} - 1$). The loss of degrees of freedom changes the distribution of the test statistic, meaning that as the degrees of freedom decrease, the associated *t* value gets larger. Practically, this means that the value of the statistic needed to achieve statistical significance will grow as the degrees of freedom decrease. Weisburd and Gill (2014) refer to this tradeoff as "paying a fine." When the sample size is large, as in the case of the CSYS, and/or when the causal processes underlying the impacts of treatment are well understood, one can afford to lose degrees of freedom in exchange for a substantial gain in efficiency if pair-matching is effective. This issue becomes more crucial in small *N* studies.

With respect to a minimum sample size threshold for pair-matching in individual-based RCTs, to our knowledge there has been no research conducted on this topic. This may have something to do with the view that individual-based trials have modest to sufficiently large *N*s, especially compared with cluster- and place-based trials (see below). It is noteworthy that Ariel and Farrington (2010, p. 438) state that pair-matching in combination with random allocation is "adequate when there are several hundred participants or less." For larger *N*s, any potential benefits to study power and efficiency that could be yielded from improved covariate balance tend to become negligible. Simple randomization seems sufficient. Nevertheless, there does not appear to be any downsides to using matched-pairs for larger *N* trials.

It is also noteworthy that Farrington (1983; see also Farrington & Welsh, 2005, 2006) established a minimum threshold of "about" 50 units per condition for simple RCTs irrespective of the unit of allocation—meaning that there was no differentiation among individual-, cluster-, or place-based trials. The basis of this threshold was the probability—based on coin tosses—of achieving equivalence on extraneous variables between treatment and control conditions.[2] Farrington (1983, p. 263, note 2) provides an informative explanation for the basis of this threshold (which would be a combined total of 100 units when there is one treatment group and one control group):

> …imagine drawing samples of 10, 100, or 1,000 unbiased coins. With 10 coins, just over 10 percent of the samples would include 2 or less, or 8 or more, heads. With 100 coins, just over 10 percent of the samples would include 41 or less, or 59 or more, heads. With 1,000 coins, just over 10 percent of the samples would include 474 or less, or 526 or more, heads. It can be seen that, as the sample size increases, the proportion of heads in it fluctuates in a narrower and narrower band around the mean figure of 50 percent.

4.2. *Cluster-based trials*

While drawn from medicine and public health, a particularly noteworthy example of an evaluation that combined pair-matching with random allocation at the cluster-level is the Universal Mexican Health Insurance Evaluation (Imai et al., 2009a). This study aimed to assess the efficacy of a government program created to provide health insurance coverage to uninsured citizens of Mexico. Researchers developed "health clusters" determined by a particular clinic and its surrounding population. Out of the 12,824 health clusters in Mexico, researchers selected 100 clusters to establish 50 matched-pairs (based on the covariates of census demographics, poverty, education, and health infrastructure). One health cluster per pair was randomly allocated to the treatment group, which involved the immediate implementation of the government program. The control clusters were eligible for the program at a later date. The primary outcome of interest was the level of patient out-of-pocket expenditures. Secondary outcomes of interest included self-reported health behaviors, health self-assessment, and medical service utilization.

For some RCTs the unit of allocation is groups of individuals, such as communities, schools, or health clinics. These groups are known as clusters. Depending upon the research question of interest, it may not be feasible to randomly allocate individuals to treatment and control conditions within these settings. It is important to note that, while groups of individuals are the unit of allocation, analyses of cluster-based trials can occur at the cluster-level (e.g., paired *t*-test on the means of each cluster) or at the individual-level using multi-level modeling/hierarchical linear modeling methods or a generalized estimating equations approach (Donner & Klar, 2000). Similar to the process of pair-matching individuals, clusters are matched on relevant covariate characteristics and then clusters in each pair are randomly allocated to treatment and control conditions. These studies are characterized by the intra-cluster correlation coefficient (ICC), or the outcomes of individuals that belong to the same cluster (Chondros et al., 2021).

There are a number of advantages associated with pair-matching in cluster-based RCTs compared with simple RCTs. These include (a) improved covariate imbalance (Balzer et al., 2015; Ivers et al., 2012); (b) increased statistical power (Imai et al., 2009a; Ivers et al., 2012); and (c) increased efficiency (Balzer et al., 2015; Imai et al., 2009a; Ivers et al., 2012). In randomly assigning clusters to treatment and control conditions (compared to individuals), a natural drop in efficiency occurs due to the loss of degrees of freedom. This happens because, rather than being considered as n/2 independent (or identically distributed pairs of units), the observed data consists of *n* dependent units and a larger test statistic value is needed to achieve statistical significance. Pair-matching compensates by "allow[ing] researchers to obtain the efficiency gains of modeling without risking the statistical advantages of random assignment" (Imai et al., 2009a, p. 70).

Chondros et al. (2021) demonstrated the importance of utilizing prior knowledge in the matching process when they performed a simulation study comparing the efficiency of the matched-pair design with stratified and simple cluster-based RCTs. The authors found that the matched-pair design was more efficient when the correlation between outcomes within pairs was moderate to strong ($r \geq 0.3$), but not more efficient with weaker correlations. However, as the authors explain, the "degrees of freedom used to calculate the confidence interval and P-value for the intervention effect is based on the number of pairs of clusters rather than the total number of clusters," such that pair-matching results in a substantial loss of degrees of freedom compared to

[2] It is important to note that this decision was also based, in part, on the state of RCTs of intervention with criminological outcomes at the time; that is, selecting a minimum *N* threshold any higher (e.g., 100 units per condition) would have resulted in many more RCTs being excluded from the review. This was just as applicable in the update of Farrington's (1983) review 23 years later: "The choice of any minimum sample size to achieve reasonable equivalence on extraneous variables depends on the definition of reasonable equivalence. We felt that the likelihood of large nonequivalence would be too great for samples of fewer than fifty [i.e., in both the treatment and control conditions]. A minimum size of 100 in each sample might have been preferable, but this criterion would have caused the exclusion of too many experiments. Hence, we set a minimum size for inclusion of fifty in each sample, or 100 in total" (Farrington & Welsh, 2006, p. 61, note 2).

simple or stratified designs (Chondros et al., 2021, p. 5766).

The efficiency of a matched-pair study is considerably reduced when the numbers of pairs are less than ten, unless clusters are paired on covariates that are strongly correlated with the outcome. As the sample cluster size increases, closer matches can be achieved on cluster-level outcomes because the observed variability on these observations diminishes. When planning a study with an anticipated intra-cluster correlation coefficient of 0.001, effective matching can potentially capture up to 50 % of the outcome variability between clusters when the sample cluster size is 1000, compared to only 9 % when the sample cluster size is 100. In trials where the number of clusters available is limited, but entire communities are sampled, such as large-scale public policy evaluation and community trials, effective matches are more likely to be achieved than a trial with many clusters with small cluster sizes. For this reason, matched-pair RCTs may be better suited when cluster sizes are large (e.g., >100 participants) and number of clusters are small (Chondros et al., 2021). Additionally, the researchers demonstrated that a higher ICC allows for a larger mean confidence interval width, which increases the likelihood that any observed changes may be attributed to the intervention. Chondros et al. (2021) argue that RCTs with fewer than four clusters per study arm should be disregarded since they face a high risk of covariate imbalance, not enough study power to detect clinically important effect sizes, and prohibitive generalizability of results.

With respect to the Universal Mexican Health Insurance Evaluation, Imai et al. (2009a) calculated the relative efficiency and study power of matched-pair cluster randomization (MPCR) and unmatched cluster randomization (UMCR) designs under four target population quantities of interest: (a) the sample average treatment effect, which is an average of the set of all units in the observed sample; (b) the cluster average treatment effect, which treats observed clusters as fixed and the units within clusters as randomly sampled from the population of units within each cluster; (c) the unit average treatment effect, which treats the clusters as randomly sampled from a larger population; and (d) the population average treatment effect (PATE), which is the average treatment effect of the entire population of units within the population of clusters. Compared to the UMCR design, the MPCR design was far more efficient and contained less standard error, particularly for measuring the PATE. For this measure, as noted by Imai et al. (2009a, p. 41), “The MPCR design for different variables is between 1.8 and 38.3 times more efficient [than the UMCR]. In this situation, our standard errors would have been as much as six times larger if we had neglected to match first.” Study power was also significantly improved through the utilization of the MPCR design due to improved covariate imbalance.

On the matter of sample size threshold for pair-matching in cluster-based trials, a scholarly debate has been underway for several decades. Chondros et al. (2021) recommend no less than four pairs due to concerns with covariate imbalance, study power, and generalizability of results.[3] As the first work to propose a threshold, Martin et al. (1993) argued that, owing to a loss of degrees of freedom, a matched-pair RCT would only be more advantageous than a simple RCT if there were at least ten pairs. It is important to note that these calculations were made with clusters of equal sizes and with a particular assumed parametric model relating the matching and outcome variables. Imai et al. (2009a) challenged this assertion and showed that pair-matching can be beneficial with as few as three pairs. By incorporating weighted cluster means into calculations, efficiency and study power gains were substantial. Imai et al. (2009a, p. 43) explained this point in greater detail:

> When cluster sizes are unequal, the efficiency gain of matching in CR [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 size depending on quantity of interest), not the unweighted correlations used in Martin et al.'s calculations. […] 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 CR 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).

4.3. *Place-based trials*

Like cluster-based trials, place-based trials usually have small sample sizes. As referenced, places, or specific geographic locations, are the unit of allocation and each place-based unit is comprised of measurable events (e.g., 911 calls, violent crimes). An example of a place-based RCT that used pair-matching was carried out by Weisburd et al. (2008) to evaluate a risk-focused policing intervention in Redlands, California. The Redlands Police Department used a survey to identify census block groups that had high scores for risk factors within at least one of the following four domains: school, family, community, and peer/individual. Risk factor scores were tallied and census block groups that scored above an established threshold in one or more domains were eligible to participate in the policing intervention trial. The authors found considerable variability in the characteristics of the 26 eligible census block groups. To address this limitation, the researchers used pair-matching, which resulted in 13 pairs matched according to risk factor scores, calls for police service, population density, and median home value. Units in each matched pair were then randomly allocated to receive risk-focused policing or usual patrol.

Using the Jersey City Drug Market Analysis Experiment (JCE), Weisburd and Gill (2014) illustrated how partial-blocking and full-blocking (also known as pair-matching) combined with random allocation can improve covariate imbalance and study power in place-based trials. The JCE included 56 hot spot locations for drug activity. Partial-blocking was used with 46 of the hot spots. This involved the use of matching to create categories of “high,” “moderate,” and “low” levels of drug activity. For the other ten hot spot locations, pair-matching was used, creating five “very high activity” statistical blocks. Pair-matching was used because these ten places had particularly high arrest and call activity gaps between them (Weisburd & Gill, 2014).

Weisburd and Gill (2014) also demonstrated how blocking (partial and pair-matching) displayed superiority over simple randomization. Using outcome data from the JCE (based on the 56 places), the authors ran 10,000 simulations of both simple randomization and block randomization. For block randomization, a total of 315 simulations produced significantly different outcomes for treatment and control conditions at baseline. In contrast, for simple randomization, a total of 2910 simulations produced significantly different outcomes for treatment and control conditions at baseline. While the authors acknowledged that the latter finding is expected owing to the significance threshold ($p < .10$) used in the simulations, the “important point is that the block randomization approach allowed us to do better” (Weisburd & Gill, 2014, p. 104, note 3).

Weisburd and Gill (2014) ran another simulation model to investigate whether the equivalence gained from block randomization with 28 units per condition (or $N = 56$) would be equal or greater than the equivalence gained from simple randomization with 50 units per condition (or $N = 100$). Results indicated that block randomization was

[3] As the ICC approaches zero, effective matches are difficult to achieve because most of the variation in outcome is within clusters. Chondros et al. (2021, p. 5776) note that, “Matching is intuitively appealing, but we have confirmed … that unless it is possible to create very strong matches it is less useful in practice, especially for studies with small numbers of clusters. Furthermore, there is often little information on the potential strength of matching, and the researchers may be overly optimistic about the effectiveness of their matching.”

again the better approach.

With place-based trials usually involving small sample sizes, this makes the pair-matching benefits of improved covariate imbalance, study power, and efficiency all the more vital. It is also the case that place-based trials that employ pair-matching can overcome concerns about differential attrition. However, in small-*N* studies the dropping of both sets of a pair can present a potential limitation. Additionally, as with cluster-based trials with small sample sizes, pair-matching in place-based trials requires substantial knowledge about the causal process of the intervention in order for places to be matched on meaningful covariates.

To date, there is no definitive sample size threshold for the use of pair-matching in place-based RCTs. While Weisburd and Gill (2014) demonstrated the benefits of pair-matching in the JCE, using the ten highest activity hot spots to create five matched pairs, the authors were not able to make any claims that five pairs was sufficient on its own. Treatment and blocking factors were considered to be fixed effects in this experiment. In studies where treatment varies, treatment might be defined as a random effect. Weisburd and Gill (2014, p. 110) explained, "The statistical blocks in the JCE were defined based on natural breaks in the data, and we think it would be problematic in such cases to extend the statistical analysis inferences to the population of 'blocks' as would be required if we defined the block as a random effect."

5. Discussion and conclusions

Through a review of matched-pair RCTs compared with simple RCTs, this article set out to address a fair but somewhat provocative question for evaluation research on crime and violence prevention interventions: Is it time for the use of pair-matching in all RCTs? The "all" refers to the different types of units of allocation: individuals; groups of individuals (or clusters); and geographical places. At the heart of this question is the ability of the design to most efficiently and robustly compare like with like, thereby, improving confidence in observed effects of intervention trials. Also of interest is an examination of sample size threshold for the different applications of the design.

The motivation for the article grew out of recent works on the history of this evaluation design in criminology and medicine (Welsh et al., 2022), as well as more contemporary interest—also in these fields of study—in the utility of the design coupled with a growing body of research demonstrating its benefits compared to simple RCTs (Ariel & Farrington, 2010; Balzer et al., 2015; Chondros et al., 2021; Imai et al., 2009a, 2009b; Weisburd & Gill, 2014). Also important is the larger context of experimental evaluation research. Specifically, the RCT has been the subject of a great deal of critiques over the years, not to mention a good number of strawman arguments, with much of this directed at the RCT as if it were a single, uniform evaluation design. The consequence of this has been a failure to recognize variability across a range of methodological and operational features and examine the RCT with more precision.

Several key findings emerge from the current review. First, it is inadequate to examine or discuss RCTs as a single, uniform evaluation design. Here, the key organizing construct is the unit of allocation: individuals; groups of individuals (or clusters); and geographical places. Second, the advantages vastly outweigh the disadvantages for the use of matched-pair RCTs compared to simple RCTs, and most of the advantages hold for all three units of allocation. Third, pair-matching can be used with rather small *N*s ($\geq$6 units) in cluster-based trials without compromising statistical power or degrees of freedom; less is known about individual- and place-based trials. Fourth, pair-matching cannot be used with some types of RCTs (e.g., cross-over) and is less amenable in other contexts (e.g., RCTs that enroll and randomize individuals on a rolling basis). Relatedly, there may be more than one choice of trial design appropriate to answer the research question, and advantages and limitations of each trial design will need to be considered.

5.1. Limitations

The current review has some key limitations. First, while the body of knowledge on the utility of pair-matching with RCTs (compared to simple RCTs) is increasing and becoming more methodologically rigorous, there is a paucity of information in some important areas. One of these areas has to do with evaluation of the designs when the units of allocation are either individuals or geographical places. As shown here, the major focus of the research so far has been on cluster-based trials. In some instances, this limited our ability to expound on—drawing on quantitative studies—the advantages and disadvantages of individual- and place-based trials. This was particularly acute in our efforts to examine the minimum *N* threshold for these types of trials. Information is also lacking with respect to diversity of outcomes. Most of the quantitative studies have been conducted in the context of medicine and public health. While this does not present a serious threat to the generalizability of findings across the fields of study under investigation here, it does call attention to the need for increased social science research (see Welsh et al., 2021), a point we return to below.

Second, some may view our sole focus on the RCT as a limitation itself. This is legitimate, especially when the RCT is but one of many internally valid research designs available for evaluating interventions (see Nagin & Weisburd, 2013). In addition, it needs to be emphasized that it is the research question under investigation that should guide the type of evaluation design to be used, not the other way around. As the body of knowledge on matched-pair RCTs likely becomes even more robust in the years ahead, it will be useful to examine the utility of this design in comparison to other experimental as well as quasi-experimental designs.

Another limitation of the article concerns our singular focus on comparing the matched-pair RCT design and the simple RCT design when other methods of randomization are available (see Chondros et al., 2021; Turner, Fan, et al., 2017). As we noted in the beginning of the article, our focus on this comparison is guided by the increased attention and growing body of research on pair-matching in combination with random allocation as an alternative to the simple RCT design, in addition to the focus of this research on the different units of allocation.

Also, while the narrative review method is well suited to the objectives of the current article, it is certainly the case that a systematic review could be carried out on this topic. In many respects, we view our review as a ground clearing exercise (i.e., to examine the advantages and disadvantages of matched-pair RCTs compared to simple RCTs) and as a potential starting point for future systematic reviews. For example, in policing, where there have been hundreds of RCTs (see Braga et al., 2014; Mazerolle et al., 2023; Neyroud, 2017), a systematic review could be carried out on matched-pair RCTs compared to simple RCTs (in addition to other types of RCT designs and for different units of allocation) to investigate the relationship between study design and outcomes.

5.2. Implications for research

As noted above, the current review found that the advantages vastly outweigh the disadvantages for use of matched-pair RCTs compared to simple RCTs in evaluating interventions in criminology and medicine. Moreover, most of these advantages hold for all three units of allocation: individuals; clusters; and geographical places. Put another way, researchers need to give up very little, and sometimes nothing at all, in using matched-pair RCTs compared to simple RCTs. It needs to be reiterated that pair-matching is not feasible with all types of RCTs. The cross-over or repeated cross-over RCT (see Sherman, 2022) is one example. Pair-matching is also less amenable, or altogether impractical, under some conditions in which RCTs are implemented. An example is when RCTs need to enroll and randomize individuals on a rolling basis. This can take place in a critical care medical setting or in juvenile or criminal court where a judge may be less agreeable to any delay in assignment to different dispositions or sanctions (i.e., assignment to treatment and control conditions).

To return to the point about researchers needing to give up little in return for the benefits of this novel design, not all conditions are made equal. For example, where there is a large and homogeneous group of individuals (e.g., 7-year-old Caucasian boys from low SES families and suffering from severe attention deficit hyperactivity disorder), pair-matching may do little to address a small degree of covariate imbalance (between treatment and control conditions) at the expense of limiting sub-group analyses, for example. Conversely, if the same homogeneous group of individuals is rather small in numbers (e.g., <25 per condition), the small degree of covariate imbalance that can be mitigated, first, by pair-matching and, second, by random allocation, would seemingly outweigh the inability to conduct sub-group analyses. Importantly, there are analytical approaches that can be used to adjust for imbalances in trials arms, as well as increase study efficiency and power (see Turner, Prague, et al., 2017). It is important to note that pair-matching is not always a limitation to conducting sub-group analyses, but it can be a consideration. Moreover, in the context of blocking in general (not just full blocking or pair-matching), Ariel and Farrington (2010, p. 449) recommend that "without proper planning and sound rationale for conducting the analysis, subgroup analyses should not be considered as a replacement for prerandomization blocking."[4] Missing from this example is a quantitative understanding of the minimum *N* threshold for pair-matching with individual-level trials. This should be a priority for future research.

The stakes are even higher when the unit of allocation is clusters or geographical places. This has much to do with the difficulty in obtaining a sufficiently large *N* for cluster- and place-based trials. The use of a small or moderate *N* is compounded when there is sizeable heterogeneity across units. Take hot spots policing as an example. In Weisburd and Gill's (2014) reporting of the Jersey City Drug Market Analysis Experiment, a place-based trial, drug activity hot spots were divided into four different levels of emergency calls for service and arrests at pre-test: very high, high, medium, and low. With only 56 hot spots, partial blocking and full blocking (or pair-matching) were used to mitigate some of this heterogeneity prior to random allocation to treatment and control conditions. This was done by using partial blocking for the high, medium, and low groups and full blocking (or pair-matching) for the very high group. Weisburd and Gill (2014, p. 104) describe the particulars of the blocking procedures as follows: "The ten highest activity hot spots were randomized in pairs because of large gaps between them; these five pairs represented the five 'very high activity' statistical blocks. Of the rest of the sample of hot spots, 8 were grouped into a 'high activity' block, 26 hot spots were classified as a medium activity block, and 12 as a low activity block." Importantly, the authors reported no reduction in statistical power or concerns about the loss of degrees of freedom. Even though this example employed both partial and full blocking, it marks a key first step in better understanding the utility of pair-matching with place-based RCTs. Further research is needed to ascertain the range of conditions under which pair-matching is feasible with place-based RCTs, including research on the minimum *N* threshold.

In the case of cluster-based trials, where the knowledge base on the utility of pair-matching is most developed, there is growing recognition of the need to use pair-matching whenever it is possible. For example, Imai et al. (2009a, p. 48) argue 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 reduce efficiency, power and robustness…" This state of affairs by no means negates the need for continued research on specific issues related to the design (see Chondros et al., 2021). Nevertheless, researchers investigating the benefits of pair-matching with individual- and place-based trials would be well served to draw upon the substantial advances in knowledge about pair-matching with cluster-based trials.

5.3. *Implications for public policy*

Pair-matching in combination with random allocation in controlled trials is a tool that can go a long way to helping researchers, policymakers, practitioners, and the wider public to have greater confidence in reported effects of crime and violence prevention interventions—-whether it be anti-bullying programs in schools, improved street lighting, hot spots policing, or cognitive behavioral therapy for offenders (Weisburd et al., 2016; Welsh & Farrington, 2014). The results of these interventions on crime and other outcomes—whatever the results may show—must be those that can be trusted. This begins with using the highest quality evaluation design to address the question under investigation. It bears repeating that the research or policy question being investigated needs to drive the type of evaluation design that will be used, not the other way around. In the current article, we are focused quite narrowly on the randomized controlled trial and how, under certain conditions and in different contexts, it can be improved so that every stakeholder can have even greater trust in the results.

Of course, whether the intervention ever gets translated into policy and routine practice or gets scaled-up for wider dissemination is a highly important but different matter altogether (Fagan et al., 2019). The goal, from the start of the process, must be to evaluate the intervention using the most rigorous design possible and provide results that decision-makers and the public can trust. This is a central issue in an evidence-based approach to policy-making (Haskins, 2018). But the evidence-based movement is well beyond accepting this point. Indeed, this could be viewed, in the words of Weisburd et al. (2016), as a "first generation" issue. The focus now should be about moving to "second generation" studies, which are more attentive to delivering and maintaining effective interventions in specific contexts and for specific groups. We think the use of pair-matching with RCTs, when feasible, can play an important part in this next phase of the evidence-based movement.

Declaration of competing interest

None.

Data availability

No data was used for the research described in the article.

Acknowledgements

We are grateful to the anonymous reviewers and the journal editor for especially helpful comments.

References

Amberson, J. B., McMahon, B. T., & Pinner, M. (1931). A clinical trial of sanocrysin in pulmonary tuberculosis. *American Review of Tuberculosis, 24*, 401–435.

Ariel, B., & Farrington, D. P. (2010). Randomized block designs. In A. R. Piquero, & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 437–454). New York: Springer.

Balzer, L. B., Petersen, M. L., van der Laan, M. J., & the SEARCH Consortium. (2015). Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation. *Statistics in Medicine, 34*, 999–1011.

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology, 1*, 417–433.

Boruch, R., Victor, T., & Cecil, J. (2000). Resolving ethical and legal problems in randomized experiments. *Crime & Delinquency, 46*, 300–353.

Bothwell, L. E., Greene, J. A., Podolsky, S. H., & Jones, D. S. (2016). Assessing the gold standard—Lessons from the history of RCTs. *New England Journal of Medicine, 374*, 2175–2181.

[4] Ariel and Farrington further state that "[s]ubgroup analyses should be justified on theoretical grounds *a priori*, in order to avoid the appearance of improper data-mining" (p. 449, emphasis in original).

Braga, A. A., Welsh, B. C., Papachristos, A. V., Schnell, C., & Grossman, L. (2014). The growth of randomized experiments in policing: The vital few and the salience of mentoring. *Journal of Experimental Criminology, 10*, 1–28.

Campbell, M. J., Donner, A., & Klar, N. (2007). Developments in cluster randomized trials and statistics in medicine. *Statistics in Medicine, 26*, 2–19.

Chalmers, I. (1989). Evaluating the effects of care during pregnancy and childbirth. In I. Chalmers, M. Enkin, & M. C. N. C. Keirse (Eds.), *Effective care in pregnancy and childbirth* (pp. 3–38). Oxford: Oxford University Press.

Chondros, P., Ukoumunne, O. C., Gunnn, J. M., & Carlin, J. B. (2021). When should matching be used in the design of cluster randomized trials? *Statistics in Medicine, 40*, 5765–5778.

de Q. Cabot, P. S. (1940). A long-term study of children: The Cambridge-Somerville Youth Study. *Child Development, 11*, 143–151.

DeSantis, S. M., Li, R., Zhang, Y., Wang, X., Vernon, S. W., Tilley, B. C., & Koch, G. (2020). Intent-to-treat analysis of cluster randomized trials when clusters report unidentifiable outcome proportions. *Clinical Trials, 17*, 627–636.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.

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

Fagan, A. A., Bumbarger, B. K., Barth, R. P., Bradshaw, C. P., Cooper, B. R., Supplee, L. H., et al. (2019). Scaling-up evidence-based interventions in US public systems to prevent behavioral health problems: Challenges and opportunities. *Prevention Science, 20*, 1147–1168.

Farrington, D. P. (1983). Randomized experiments on crime and justice. *Crime and Justice, 4*, 257–308.

Farrington, D. P. (2013). Longitudinal and experimental research in criminology. In M. Tonry (Ed.), *Crime and justice: 1975–2025* (pp. 453–527). Chicago: University of Chicago Press.

Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology, 1*, 9–38.

Farrington, D. P., & Welsh, B. C. (2006). A half century of randomized experiments on crime and justice. *Crime and Justice, 34*, 55–132.

Haskins, R. (2018). Evidence-based policy: The movement, the goals, the issues, the promise. *The Annals of the American Academy of Political and Social Science, 678*, 8–37.

Hawkins, J. D., Catalano, R. F., Arthur, M. W., Egan, E., Brown, E. C., et al. (2008). Testing Communities That Care: The rationale, design and behavioral baseline equivalence of the community youth development study. *Prevention Science, 9*, 178–190.

Imai, K., King, G., & Nall, C. (2009a). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Science, 24*, 29–53.

Imai, K., King, G., & Nall, C. (2009b). Rejoinder: Matched pairs and the future of cluster-randomized experiments. *Statistical Science, 24*, 65–72.

Ivers, N. M., Halperin, I. J., Barnsley, J., Grimshaw, J. M., Shah, B. R., Tu, K., Upshur, R., & Zwarenstein, M. (2012). Allocation techniques for balance at baseline in cluster randomized trials: A methodological review. *Trials, 13*, 1–9 (e120).

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

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

Mazerolle, L., Eggins, E., Hine, L., & Higginson, A. (2023). The role of randomized experiments in developing the evidence for evidence-based policing. In D. Weisburd, T. Jonathan-Zamir, G. Perry, & B. Hasisi (Eds.), *The future of evidence-based policing*. New York: Cambridge University Press (in press).

Nagin, D. S., & Sampson, R. J. (2019). The real gold standard: Measuring counterfactual worlds that matter most to social science and policy. *Annual Review of Criminology, 2*, 123–145.

Nagin, D. S., & Weisburd, D. (2013). Evidence and public policy: The example of evaluation research in policing. *Criminology & Public Policy, 12*, 651–679.

Neyroud, P. W. (2017). *Learning to field test in policing: Using an analysis of completed randomised controlled trials involving the police to develop a grounded theory on the factors contributing to high levels of treatment integrity in police field experiments*. Cambridge, UK: University of Cambridge. https://doi.org/10.17863/CAM.14377 (Doctoral dissertation).

Podolsky, S. H., Welsh, B. C., & Zane, S. N. (2021). Richard Cabot, pair-matched random allocation, and the attempt to compare like with like in the social sciences and medicine. Part 2: The context of medicine and public health. *Journal of the Royal Society of Medicine, 114*, 264–270.

Powers, E., & Witmer, H. L. (1951). *An experiment in the prevention of delinquency: The Cambridge-Somerville Youth Study*. New York: Columbia University Press.

Sampson, R. J. (2010). Gold standard myths: Observations on the experimental turn in criminology. *Journal of Quantitative Criminology, 26*, 489–500.

Sherman, L. W. (2010). An introduction to experimental criminology. In A. R. Piquero, & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 399–436). New York: Springer.

Sherman, L. W. (2022). "Test-as-you-go" for hot spots policing: Continuous impact assessment with repeat crossover designs. *Cambridge Journal of Evidence-Based Policing, 6*, 25–41.

Smith, G. C. S., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: A systematic review of randomised controlled trials. *British Medical Journal, 327*, 1459–1461.

Turner, E. L., Fan, L., Gallis, J. A., Prague, M., & Murray, D. M. (2017). Review of recent methodological developments in group-randomized trials: Part 1—Design. *American Journal of Public Health, 107*, 907–915.

Turner, E. L., Prague, M., Gallis, J. A., Fan, L., & Murray, D. M. (2017). Review of recent methodological developments in group-randomized trials: Part 2—Analysis. *American Journal of Public Health, 107*, 1078–1086.

Wacholder, S., & Weinberg, C. R. (1982). Paired versus two-sample design for a clinical trial of treatments with dichotomous outcome: Power considerations. *Biometrics, 38*, 801–812.

Weisburd, D. (2010). Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: Challenging the folklore in evaluation research in crime and justice. *Journal of Experimental Criminology, 6*, 209–227.

Weisburd, D., Farrington, D. P., & Gill, C. E. (Eds.). (2016). *What works in crime prevention and rehabilitation: Lessons from systematic reviews*. New York: Springer.

Weisburd, D., & Gill, C. E. (2014). Block randomized trials at places: Rethinking the limitations of small N experiments. *Journal of Quantitative Criminology, 30*, 97–112.

Weisburd, D., & Hinkle, J. C. (2014). The importance of randomized experiments in evaluating crime prevention. In B. C. Welsh, & D. P. Farrington (Eds.), *The Oxford handbook of crime prevention* (pp. 446–465). New York: Oxford University Press.

Weisburd, D., Morris, N. A., & Ready, J. (2008). Risk-focused policing at places: An experimental evaluation. *Justice Quarterly, 25*, 163–200.

Welsh, B. C., & Farrington, D. P. (Eds.). (2014). *The Oxford handbook of crime prevention*. New York: Oxford University Press.

Welsh, B. C., Podolsky, S. H., & Zane, S. N. (2021). Richard Cabot, pair-matched random allocation, and the attempt to compare like with like in the social sciences and medicine. Part 1: The context of the social sciences. *Journal of the Royal Society of Medicine, 114*, 212–217.

Welsh, B. C., Podolsky, S. H., & Zane, S. N. (2022). Pair-matching with random allocation in prospective controlled trials: The evolution of a novel design in criminology and medicine, 1926-2021. *Journal of Experimental Criminology*. https://doi.org/10.1007/s11292-022-09520-2