Advance Access publication September 17, 2012

Political Analysis (2012) 20:460–479
doi:10.1093/pan/mps025

Multivariate Continuous Blocking to Improve Political Science Experiments

Ryan T. Moore
University of California–Berkeley and University of California–San Francisco, 50 University Hall, MC7360, Berkeley CA 94720-7360; Department of Political Science, Washington University in St. Louis, 241 Seigle Hall, Campus Box 1063, One Brookings Drive, St. Louis MO 63130
e-mail: rtm@wustl.edu

Edited by R. Michael Alvarez

Political scientists use randomized treatment assignments to aid causal inference in field experiments, psychological laboratories, and survey research. Political research can do considerably better than completely randomized designs, but few political science experiments combine random treatment assignment with blocking on a rich set of background covariates. We describe high-dimensional multivariate blocking, including on continuous covariates, detail its statistical and political advantages over complete randomization, introduce a particular algorithm, and propose a procedure to mitigate unit interference in experiments. We demonstrate the performance of our algorithm in simulations and three field experiments from campaign politics and education.

1 Introduction

Political scientists use randomized treatment assignments to aid causal inference in many research settings. Laboratory and survey randomizations offer precise control in well-defined environments; Bullock (2011), for example, manipulates whether survey respondents are randomly told that Democratic legislators support or oppose a new policy. On the other hand, field experimentation allows measurement of effects of randomly assigned interventions in natural (but often challenging to control) settings: Gerber et al. (2011) randomized the amount and timing of paid political advertising in conjunction with Rick Perry's 2006 Texas gubernatorial campaign. Much of the experimental political science literature relies on *complete randomization*, wherein each unit is assigned to treatment or control based on an independently generated random number, or *random allocation*, the random selection of a given proportion of the units for each treatment condition (Lachin 1988). Occasionally, designs consider a few characteristics of the respondents, precincts, or schools being randomized, but rarely do experimenters exploit more than one or two discrete covariates prior to randomization.

Though advantages of blocking a few discrete covariates have been known for some time, we describe methods for multivariate blocking on continuous covariates, and we demonstrate the benefits of our methods in a variety of simulated and real political science data. We show how blocking political experiments on several covariates can improve the precision of causal estimates and facilitate clean comparisons between treatment and control groups by guaranteeing balance. Blocked designs can also make experimental estimates more robust to unlucky randomizations or inadequate parametric adjustment, produce local causal estimates, and preserve power, even in relatively small group-randomized trials.

Author's note: Thanks to Keith Schnakenberg and Jack Frasier for research assistance; thanks to Michael L. Anderson, Jake Bowers, Thad Dunning, Jeff Gill, Donald P. Green, Jens Hainmueller, Gary King, Kevin Quinn, Jas Sekhon, Laura Stoker, and anonymous reviewers for suggestions on previous versions; thanks to Daniel Rubenson, Peter John Loewen, Larry Schweinhart, and Zongping Xiang for generously sharing data and answering questions. The replication archive is available as Moore (2012b). Supplementary materials for the article are available on the *Political Analysis* Web site.

Blocking can improve designs across political research settings. We focus on field experiments, as their conjunction of random treatment assignment and genuine political environments have led to increasing recent popularity. Field experimentation in political science dates at least to the 1920s, but the past decade has seen an explosion of randomized political field trials. Get-out-the-vote (GOTV) experiments have dominated this resurgence, but political communication, election monitoring, social capital, institution building, and other experiments have taken root across the subfields of political science. These experiments complement the extensive related literature from public policy analysis, where public health, criminology, education, and social welfare policy evaluations abound.

The next section introduces some preliminary concepts; we also offer Supplementary Materials with more examples and definitions. Section 3 introduces multivariate blocking and discusses when, why, and what to block on. Thereafter, we detail our optimal-greedy blocking algorithm, discuss variants and alternatives for practical multivariate blocking, and demonstrate their advantages. Using data from actual field experiments in early childhood education, elite political persuasion, and political advertising, we show how our methods improve balance and precision in a variety of settings.

2 Preliminary Concepts and Notation

Consider a sample of experimental units indexed by $i \in \{1,\ldots,2n\}$. A binary treatment T_i takes the value 1 for treatment or 0 for control for each unit, and is applied to each unit at a certain point in time. The observed outcome Y_i is a function of T_i and the unit's potential outcomes under treatment (Y_{i1}) and control (Y_{i0}): $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$. The unit treatment effect is $TE_i \equiv Y_{i1} - Y_{i0}$, which cannot be observed in practice—the "fundamental problem of causal inference" (Holland 1986). The (unobservable) true average treatment effect is $\overline{TE} \equiv \frac{1}{2n} \sum_{i=1}^{2n}(Y_{i1} - Y_{i0})$. In experiments, we select which of Y_{i1} or Y_{i0} to observe by randomizing which treatment i will receive. We do so, as Holland (1986) describes, to buttress the assumption that T is independent of Y_1 and Y_0, and thus that $E(Y_1) = E(Y_1|T=1)$, $E(Y_0) = E(Y_0|T=0)$, and the observed difference in treatment and control group means, $\overline{Y}_{1|T=1} - \overline{Y}_{0|T=0}$, gives an unbiased estimate of $\overline{TE}$.

Observed covariates are pretreatment variables, with vector $\mathbf{x}_i$ containing their values for unit i. With p covariates, we can stack the $\mathbf{x}_i'$ into the $2n \times p$ matrix $\mathbf{X}$, with covariance matrix Σ. Distances d_{ij} represent how different the covariate profiles are between units i and j. Among the multivariate dissimilarity metrics we discuss are the Euclidean distance, $ED_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}$, and the Mahalanobis distance (MD), $MD_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)}$. The latter represents a scaled version of the former, with Σ^{-1} usually estimated by the inverse of $\hat{\boldsymbol{\Sigma}} = \frac{1}{2n-1}\sum_{i=1}^{2n}(\mathbf{x}_i - \hat{\mu}(\mathbf{X}))(\mathbf{x}_i - \hat{\mu}(\mathbf{X}))'$, where $\hat{\mu}(\mathbf{X}) = \frac{1}{2n}\sum_{i=1}^{2n}\mathbf{x}_i$ is the vector of sample means.

The stable unit treatment value assumption, SUTVA, is a central consideration. Violations of its first component, cases with different "versions of each treatment varying in effectiveness" (Rubin 1990, 282), can be partially addressed by particular algorithms. Different versions of the treatment might be informational campaigns of different intensities or drugs with different concentrations of active ingredients. We also offer a framework for supporting the second component, non-interference between units. Interference occurs when the potential outcomes of one unit depend on the treatment assigned to other units; for example, a subject may be more likely to turn out to vote if a neighbor is contacted by a party, even if she herself is not contacted.

3 Blocking and Its Advantages

Blocking is an element of experimental design in which a researcher uses observed covariates to create pre-assignment groups of similar units. Since the covariate measurements $\mathbf{X}$, the distance calculations d_{ij}, and the blocking that employs these distances all occur before randomization (which precedes the application of treatment), experimentalists need not worry about the usual post-treatment bias that emerges when ex post regression adjustment in an observational study includes a variable that is affected by the treatment (Epstein and King 2002), such as a mediator (Imai et al. 2011). Randomization takes place within blocks, usually via random allocation.

Researchers can block any experiment with at least one measured covariate for the randomization units. Covariate data may come from many sources, including administrative, cluster, pretest, or baseline survey data. For example, in a GOTV experiment, administrative voting records will provide turnout in prior elections. The Census provides rich data for many place-randomized trials. Laboratory and survey experimentalists often collect information before assigning individuals to treatments, and that information can be used to block before randomization. Clinical trials may incorporate individuals' long medical histories.

Blocking should focus on covariates that affect or condition the outcome. Not blocking on outcome-influencing variables can lead to fragile analyses involving extreme extrapolation, particularly with small samples, outliers, or unlucky randomizations. Instead of parametrically adjusting for large differences after an experiment ends, researchers should block to control those differences before randomization ever takes place; we return to this point in Section 6. When an analyst has too few experimental units, however, neither blocking nor randomization may help much. A two-unit experiment cannot be balanced; a certain level of imbalance exists and cannot be changed by randomization, blocking, or other design features.

Though information from many covariates, whether discrete or continuous, can be used in blocked designs, most political science experiments that block employ a limited set of categorical covariates. For example, a survey experiment sorts respondents into six groups based on a binary sex indicator and a three-level declaration of voting intention (Horiuchi, Imai, and Taniguchi 2007). However, the abundant, often continuous, pretreatment information in modern experiments is frequently ignored at the design stage. When continuous covariates include information that can improve inferences, they, too, should be employed. Further, when several covariates predict outcomes, blocking can and should be *high-dimensional*.

Prominent examples of high-dimensional blocking typically come from statistics (Lu et al. 2001; Barnard et al. 2003; Greevy et al. 2004). Very few political science examples of high-dimensional blocking or blocking on continuous covariates exist, though this may be changing (Goldstein et al. 2010). Examining all field experiments reported in the *American Political Science Review, American Journal of Political Science*, and *Journal of Politics* from 2000 to 2010, among others, we found 17 unblocked randomizations, nine that used a single discrete covariate to block, and four that used a small covariate set. In Section 5, we show concrete improvements resulting from adding covariates to several designs.

Continuous and discrete blocking can be combined; GOTV experimentalists, for example, may want to compare households of the same size or precinct, but also incorporate continuous measures like age or net worth. We see three main advantages higher-dimensional blocking possesses over limited exact blocking. First, often a main variable of design interest is nearly continuous, and no exact blocks will exist. Second, discretizing important continuous variables may throw away some of the very information that blocking intends to exploit. If a covariate warrants attention in design, then losing information about that covariate could be counterproductive. Third, limited exact blocking can quickly encounter a dimensionality curse. For example, blocking on four three-category covariates implies up to $3^4 = 81$ different profiles. If any unit has a unique covariate profile, limited exact blocking must drop it or otherwise address the absence of blockmates for singletons. While recent work describes benefits of covariate coarsening approaches (Iacus, King, and Porro 2011, 2012), here we offer aid for experimentalists with variables they wish to not coarsen, or with coarsened strata within which they want to block further.

Blocking can ensure covariate balance between treatment and control groups, improve the efficiency of causal inferences, allow alternative causal estimates to be produced, and provide guidance when resources are constrained. Others discuss benefits such as triple-robustness and preserved power in blocked designs.[1]

[1] See King et al. (2007) and Bowers (2011), for example. We discuss unlucky randomizations (showing, e.g., that blocked randomization produces covariate imbalances unlikely to be worse than no blocking, even when blocking is done poorly) in our Supplementary Materials. For a general discussion of randomization, see Duflo, Glennerster, and Kremer (2008). Others describe related advantages of low-dimensional discrete blocking (Mason et al. 1989; Murray 1998; Casella 2008; Tamhane 2009). Imai, King, and Stuart (2008) summarize the asymptotics related to various design

3.1 *Balance and Efficiency*

Complete randomization promotes covariate balance as the number of units increases to infinity, but blocking helps ensure balance in the finite sample researchers are actually given. For a population of size N, the average treatment effect in the population (PATE) is $\frac{1}{N}\sum_{i=1}^{N} TE_i$, and the difference between the PATE and the observed difference in sample means—the estimation error—can be decomposed into four terms, two of which represent the portion of the estimation error that comes from unrepresentative sampling on observed and unobserved covariates, and two of which represent the estimation error that comes from covariates unbalanced across the treatment conditions (Imai, King, and Stuart 2008). By producing covariate balance, blocking directly controls the estimation error due to differing levels of observed covariates in the treatment and control groups. Blocking also balances unobserved covariates to the extent that they correlate with observed ones.

Covariate balance between treatment conditions reduces bias, promotes efficiency, can reduce dependence on analysis model choices, and can be controlled without access to outcomes. Indeed, Rubin (2001) advocates balancing with "no outcome variables in sight" as a critical first step in observational designs as well (p. 171).

Blocking on covariates that affect the outcome increases the precision of treatment effect estimates. Blocked designs represent improvement both in finite samples and asymptotically with the number of experimental units (Imai, King, and Stuart 2008). The asymptotic result is deterministic; the difference between the asymptotic variances under classical randomization and blocking is always positive. In finite samples, the efficiency increase depends on the relationship between the number of blocks and the number of units, and in practice, blocking should increase efficiency.

3.2 *Assessing Effect Heterogeneity and Effects for Particular Units*

Blocking allows the analyst to detect and estimate heterogeneity in treatment effects. If a treatment negatively affects men, but positively affects women, blocking will ensure that both men and women receive treatment and control so that the different effects can be efficiently estimated. Without blocking, too many men or too few women may receive treatment, and average inferences will obfuscate the true effects.

Experimentalists can calculate causal effects within each block, and these block effects can constitute both a statistical and a political advantage. Statistically, if blocks representing different covariate combinations yield similar effect estimates, then these within-block "treatment comparisons . . . made across a wide variety of situations" (Casella 2008, 98) allow an analyst to make a case for homogeneous or heterogeneous treatment effects.[2]

Additionally, some blocks may include units of particular interest, and causal estimates can be calculated directly for those units. Block-level effect estimates can satisfy local implementers' interest in local estimates, but these estimates are produced without sacrificing larger project goals. For example, if a field experiment's units are geographic areas, precincts, schools, health clinics, or administrative departments, then local policy implementers often have interest in the causal estimate for the unit under their jurisdiction. Instead of designing a more limited experiment to estimate a single unit's treatment effect, block-level estimates from a larger-scope experiment provide each local implementer with such an estimate, as well as a sense of the quality of that estimate relative to others. In an unblocked experiment, analogs to this local estimate require modeling assumptions.

Further, analysts may trust estimates that come from more-similar blocks more, and an experiment's causal estimate may be thought of as a meta-analytic estimate derived from several randomized subexperiments, one in each block. Weights for such a "meta-analysis" could come

choices; for sample size asymptotics in both levels of cluster randomizations, we suggest Imai, King, and Nall (2009), the subsequent discussion, and Imbens (2011).

[2] See Green and Kern (2012) on automated searching for treatment heterogeneity.

from the block quality or block-quality rank.[3] Generally, analysts should specify ex ante whether the estimand of interest is the sample average treatment effect, a subgroup effect, a population-weighted effect, etc.

3.3 *Allocating Limited Resources*

Ranking blocks by their similarity can suggest to politicians and bureaucrats where resources are best deployed scientifically, allowing them to make principled experimental decisions when facing budgets that are limited or change midstream.

One experiment only had resources to conduct a household-level survey in 100 of the 148 randomization areas. Block quality ranks were one of the two criteria used to determine where the survey would be conducted (King et al. 2007, 494). Focusing implementation efforts on the best blocks reduces implementers' discretion in selecting which treatment units to treat or survey; this, in turn, hampers their ability to hand-pick outcomes. Focusing on the best blocks does not inherently select on covariate levels; instead, it selects on similarity of covariate levels. The first pair could be the two richest areas, but the second pair could be the two poorest ones.

4 How to Block: Decisions Made in Practice

Experimentalists make several choices when creating blocks. One selects a blocking algorithm, decides how to weight covariates relative to one another, and elects whether to impose substantive restrictions via particular variables. Researchers can combine their choices to suit their exact needs.[4] Below, we first detail three sets of alternatives for creating blocks, focusing on the advantages of the *optimal*-*greedy* algorithm we introduce. We then evaluate a set of covariate weightings, highlighting benefits of methods resistant to covariate outliers. Finally, we present *Goldilocks blocking*, offering an ex ante technique to restrict blocks in ways that are often described ex post or informally in randomized designs. Applied researchers can implement the algorithms and alternatives in any combination using our blockTools package (Moore 2012a) for R (R Development Core Team 2012).

4.1 *The Optimal-Greedy and Related Blocking Algorithms*

Optimal-greedy blocking makes the best-ranked matches as good as possible. In many political science field experiment settings, such as those with limited or uncertain resources, this attribute is highly desirable. With two treatment conditions, we first calculate the $2n(2n-\ 1)/2$ distances d_{ij} (collectively, $\mathbf{D}^{[1]}$) between units i and j, $\forall i, j$ s.t. $i < j$. The optimal-greedy blocking algorithm's first step finds $i^{[1]}, j^{[1]}$ such that $d_{i^{[1]}, j^{[1]}} = min(\mathbf{D}^{[1]})$ and places the two units into a block. Subsequent steps $s \in \{2, \ldots, n\}$ find $i^{[s]}, j^{[s]}$ from the set of available distances $\mathbf{D}^{[s]}$ such that $d_{i^{[s]}, j^{[s]}} = min(\mathbf{D}^{[s]})$, where the available distances at each step are defined recursively: $\mathbf{D}^{[\mathbf{s}]} = \mathbf{D}^{[s-1]} \setminus \{d_{i^{[s-1]},j}, d_{i,i^{[s-1]}}, d_{i,j^{[s-1]}}, d_{j^{[s-1]},j}\}$. We start by letting $d_{ij} = MD_{ij}$, discussing alternatives below.

As the name implies, optimal-greedy blocking shares attributes with optimal algorithms and with greedy algorithms.[5] The algorithm is *optimal* in that it considers every available distance at once, and it outperforms naive greedy algorithms on the global measures used to evaluate pure optimal algorithms. The algorithm is *greedy* in that it selects a particular block at each step rather than an entire set of blocks at once.[6] Greedy algorithms solve problems of a smaller order than do

[3] Under perfect exact blocking, all blocks are of the same quality by design. However, in other cases, the inverse of the multivariate distance within the block can represent the block quality; with blocks of more than two units, the inverse of the maximum pairwise distance is a conservative alternative.

[4] Furthermore, experimenters often have several treatment conditions, have nested levels on which to block (such as blocking states by their most similar cities or schools by their most similar classrooms), and need replication and output capabilities (King 1995), all of which our software accommodates.

[5] Hansen (2004) offers an introduction to optimal and greedy algorithms.

[6] Suboptimal greedy algorithms select the best blockmate for a particular unit at each step. The *naive* algorithm takes the units in the order presented in the data; the *sorted* algorithm allows the researcher to create the best blocks including

pure optimal algorithms, so they have shorter computational times to solution. Moderately large experiments can be infeasible to block optimally, and even smaller experiments can require selecting from a substantial number of blockings.[7] In political science field experiments, this computational time may be a small part of research costs, but when an experiment is being conducted with real policymakers under tight deadlines, quick design turnaround can be needed.

To demonstrate the optimal side of the optimal-greedy algorithm, we reanalyze data from the seventy-four blocked pairs in a large-scale randomized health policy application (King et al. 2007, 2009).[8] The optimal-greedy algorithm creates better blocks overall than a naive greedy algorithm. Specifically, sixty-five of seventy-four experimental blocks have smaller or identical Mahalanobis distances than if the units had been blocked with a naive greedy algorithm. Of the nine experimental blocks in which the naive algorithm created smaller Mahalanobis distances, most of these occur among the blocks with the largest multivariate distances.

Figure 1 compares the performance of the optimal-greedy and naive greedy algorithms in these data. The figure's eight rows show every combination of two global criteria (the mean and median MD across blocks), two unit types (rural and urban), and two subsets of the experimental units (the entire set of blocks and the best half of each states' blocks). In all eight comparisons, the optimal-greedy algorithm outperforms the naive greedy algorithm.

An optimal-greedy design ensures that block-level estimates from the most similar blocks are as good as possible, assuming a uniform treatment. When treatment efforts are nonuniform, such as when budgets force encouragements to be of different intensities or advertising campaigns to vary in penetration, optimal-greedy blocks can be used to determine resource allocations. While any scheme that can rank blocks permits resources to be allocated to the best blocks first, optimal-greedy blocks maximize the quality of exactly those blocks where estimates will be least harmed by covariate imbalance. As Boruch et al. (2004) note, when experimental treatments are nonuniform, levels of implementation effort should be recorded.

Additionally, the optimal-greedy algorithm solves the exact problem posed in some research scenarios. In the National Cancer Institute–sponsored Community Intervention Trial for Smoking Cessation (COMMIT), interested research institutions had to propose not one, but a pair of "similar communities that agreed to be randomized" if the institution was chosen for the study (National Cancer Institute 1995, 28). The optimal-greedy algorithm, along with the Goldilocks blocking described in Section 4.3, could have aided the COMMIT evaluation by allowing research institutions with several candidate communities to formally select the best blocked pair.

4.2 *Covariate Weightings*

Suppose a study has some subjects who are rich and old, and others who are poor and young. With whom should a rich, young person be blocked? Whom is she more like? To block, one must formalize whether units are similar or different, and this requires weighting covariate differences relative to one another, implicitly or explicitly.

For both discrete and continuous measures, a variety of multivariate distance measures exist, each with its own weighting assumptions. The Euclidean distance weights variable differences equally, so a difference of $100 in income receives weight equal to a difference of 100 years of age. Clearly these differences should not be equally weighted if one wants substantively homogeneous blocks. Below we describe four ways to weight variables that covary and are on different scales.

First, Mahalanobis distances weight differences in pretreatment values by the inverse of the covariates' covariance matrix. This accounts for the scales and linear relationships between variables, and implies that differences of one standard deviation (SD) on two different variables

particular units; the *random* algorithm allows researchers to shuffle data sets sorted in undesirable ways, such as by last name or time of entry into a trial.

[7] Optimally blocking 148 units, as in King et al. (2007), requires consideration of $\approx 4.1 \times 10^{128}$ possible blockings. Paluck and Green (2009) have fourteen units, requiring consideration of 135,135 blockings for optimality.

[8] The replication archive is available as Moore (2012b).

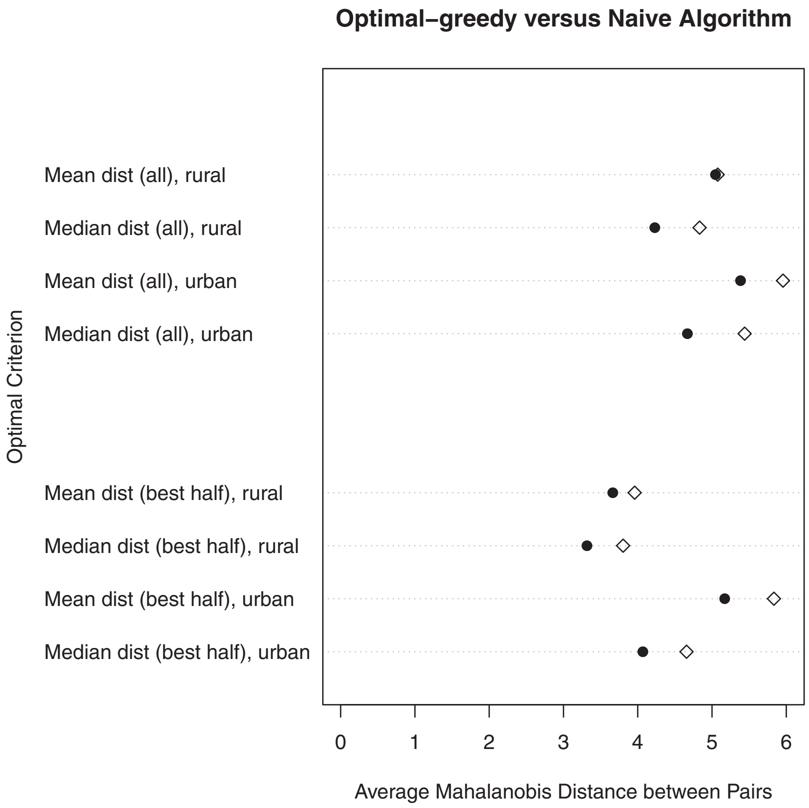


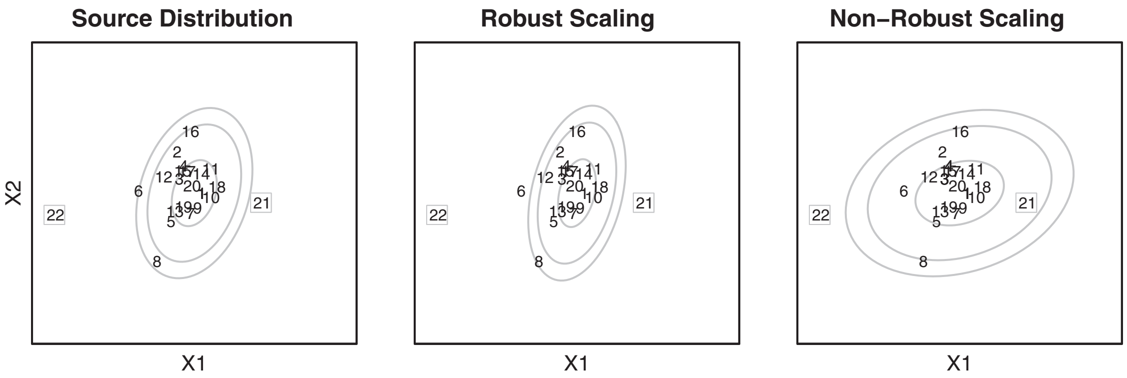
Fig. 1 Optimal-greedy algorithm outperforms naive greedy algorithm on optimal criteria. Symbols represent average Mahalanobis distance across blocks using King et al.'s (2007) data. Every optimal-greedy value (filled circle) is less than the corresponding naive value (open diamond). Eight rows display the eight combinations of (mean, median MD) × (all, best-ranked half of experimental units) × (rural, urban) health clusters.

contribute the same amount to the distance (assuming zero covariances). Important Mahalanobis-metric reductions in statistical bias have been shown in theory and in Monte Carlo simulations for observational data (Rubin 1980).

Second, in the presence of outliers, we show the practical benefit, block stability, of replacing the Mahalanobis distance's nonresistant estimate with a resistant estimate of the covariance matrix. Resistant measures of multivariate location and spread include rank-based Mahalanobis distances (Rosenbaum 2010), and the minimum volume ellipsoid estimator (MVE) and minimum covariance determinant (MCD) of Rousseeuw (1985).[9]

Figure 2 displays twenty experimental units to be paired using two covariates. The left panel's ellipses represent 50%, 90%, and 97.5% quantile regions for the bivariate normal distribution from which the covariates have been drawn. Units 21 and 22 (outlined in gray) are outliers in X_1. The second and third panels show the same units with quite different ellipses, demonstrating the sensitivity to outliers of some estimates of the covariance matrix. In the second panel, the ellipses represent the MVE estimate of the covariance matrix calculated using all twenty-two units, an estimate resistant to the outliers and very similar to the source distribution. The third panel's

[9] The MVE estimate is a positive semidefinite matrix $\hat{\boldsymbol{\Sigma}}$ whose determinant is minimized subject to $\#\{i;\, (\mathbf{x}_i - \hat{\mu}(\mathbf{X}))'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \hat{\mu}(\mathbf{X})) \le a^2\} \ge h$, where $h = [(2n+p+1)/2]$ and a^2 is a fixed value drawn from $a^2 \sim \chi^2_{p,50}$ (Rousseeuw and van Zomeren 1990). The MVE estimator uses at least half the observations to minimize the multivariate ellipsoid's volume while satisfying the constraint. The MCD estimator seeks the half of the data that generates the covariance matrix with the smallest determinant, subject to no other constraints.


Source Distribution

Robust Scaling

Non–Robust Scaling

Fig. 2 Resistant estimates of scaling matrix create more stable blocks than nonresistant estimates. Panels' ellipses represent original, resistant, and nonresistant estimates of variance–covariance scaling matrix. Outlying units in squares distort scaling matrix under nonresistant scaling, thus altering many blocks. Resistant scaling matrix reproduces blocks of source distribution.

ellipses are from the nonresistant estimate of the covariance matrix using all twenty-two units, incorporating the outlying observations equally alongside the others.

We use these three covariance matrix estimates to scale covariate differences and create sets of multivariate distances that we use to optimal-greedy block the units (see the Supplementary Materials for the blocks). Between the first and second panels, the blocks are stable; the resistant estimate preserves the initial blocks, and adds a worst block containing the outliers. On the other hand, the third panel's quantile regions are distorted by the outliers, and the nonresistant scaling matrix jumbles all but two of the blocks from those created using the resistant (or source distribution) scaling matrix. Though the resistant matrix slightly underestimates X_1's variance, it does so much less than the nonresistant estimate overestimates this variance, and it preserves the relative scaled locations of the data points.

While the resistant estimate correctly identifies units 21 and 22 as outliers, the nonresistant estimate makes these units look much closer to the main surface. The resistant estimate correctly puts these units outside the primary data cloud, while the nonresistant estimate locates these units considerably closer to the center of the surface. All blocks, not just those that include nearby units, are sensitive to the resistance of the scaling matrix estimation procedure. We include another example in the Supplementary Materials. When faced with outliers, we recommend using scaling matrices that are resistant to extreme values.[10]

Similarly, trialists can use a larger sample (or the population) of units to calculate the scaling matrix if sample variances will distort meaningful covariate differences. If the experimental sample has a small estimate of the variance of X_1, then (assuming zero covariances) small differences across units on X_1 will substantially increase a pair's MD. However, this small variance may be idiosyncratic to the experimental units, resulting in overweighting X_1. Then, using the larger sample of units to create the covariate difference scaling matrix allows the researcher to de-emphasize sample differences on X_1 that she deems less important.

Fourth, analysts with substantive knowledge about relative covariate weights can employ that information directly. Using a diagonal matrix of covariate weights $\mathbf{W}$, the usual Mahalanobis distance scaling matrix, Σ^{-1}, can be replaced by $(\mathbf{C}'\mathbf{W}\mathbf{C})^{-1}$, where $\mathbf{C} = ([\text{chol}(\Sigma)]')^{-1}$ and chol(Σ) is the Cholesky decomposition of the covariance matrix. Covariate differences will be smaller on covariates given higher weights.

[10] Note that generating a resistant measure of multivariate spread is different than generating the best possible scaling matrix for balancing covariates. Optimal scaling for covariate balance could be directly pursued, for example, through genetic algorithms such as GENOUD (Mebane and Sekhon 1988).

4.3 *Goldilocks Blocking*

We introduce Goldilocks blocking,[11] a general restriction on how similar or different a selected covariate can be within the same block. Such restrictions enable experimentalists to exploit substantive knowledge about covariate differences that one deems "just right," and to incorporate principled blocking and randomization into designs where units can be "too similar."

We implement the Goldilocks restriction for a particular covariate x_g by setting the distance to infinity for two units whose covariate difference is outside a range [*a,b*] set by the analyst:

$$d_{ij} = \begin{cases} d_{ij} & \text{if } |x_{ig} - x_{jg}| \in [a, b], \\ \infty & \text{if } |x_{ig} - x_{jg}| \notin [a, b]. \end{cases} \tag{1}$$

A blocking caliper akin to the typical observational matching caliper can be implemented by setting $a = 0$ and $b < \infty$ (Cochran and Rubin 1973).[12]

In applications, proximity between units may introduce interference, motivating Goldilocks blocking. If a vaccine prevents the spread of a contagious disease, and individuals A and B are in contact, then B's potential outcome under control will depend on A's assignment. If A is vaccinated, B is less likely to get sick; if A is not, B is more likely to catch the disease. Similar reasoning applies to political communication, GOTV prompts, and electoral monitoring, and has been described in applications.[13] While some procedures restrict how different units can be, restrictions can also prevent blockmates from being too similar. While recent work develops estimates and inference under interference (Sobel 2006; Rosenbaum 2007; Hudgens and Halloran 2008), Goldilocks blocking can increase the plausibility of the noninterference assumption in common causal inference models.

Units near one another are often similar on both observed and unobserved covariates on which analysts would like to block, but if units with different treatment assignments are too close, then analysts expect flows of information, contagion, etc., and poor compliance with experimental protocols. Donner and Klar (2000, 4), for example, write that the threat of contamination or noncompliance "can sometimes be further minimized, however, by implementing the study in a geographic area in which only distinct and well separated clusters of subjects are recruited."

Two political field experiments evince similar concerns. Gosnell (1927) splits districts to "avoid possible contacts between the experimental and the control groups" (p. 17), while Wantchekon (2003) notes that "sixteen of the twenty selected villages were at least twenty-five miles apart with seven to ten villages separating them.. . .The risk of contagion between the two treatment groups was thereby minimized" (p. 407). Similarly, in the COMMIT trial, an informal restriction on media market overlap was partially enforced.[14] The particular thresholds described could be more systematically integrated into these designs. Many applications could benefit from preventing "too close" temporal blocks as well.[15]

[11] In the fable "Goldilocks and the Three Bears," a young girl trespasses into three bears' home. She tests their porridge, chairs, and beds, each time finding two of the items too extreme, and the third "just right."

[12] Goldilocks matching could also aid causal inference in observational settings. In *caliper matching*, matches too far apart (usually on the propensity score) are prohibited (Hansen and Klopfer 2006; Ho et al. 2007; Sekhon 2011). Goldilocks matching could prevent data from being "too close," and could help create strong instruments in observational designs. Strong instruments divide treated and control units that are very similar on all relevant background characteristics except the instrument. On the instrument, they differ as much as possible (Keele and Morgan 2011).

[13] For example, in Gail et al. (1992), Gosnell (1927), Hyde (2010), Ichino and Schündeln (2012), and Wantchekon (2003).

[14] The study blocked "on factors such as population size, geographic proximity, age and sex composition, degree of urbanization, and socioeconomic factors. The paired communities were geographically close enough to permit monitoring and intervention by the investigators, but not so close that educational activities in the intervention community would affect the control community" (Gail et al. 1992, 7). Researchers gave "priority" in selecting experimental units to pairs without shared media markets, but at least one included pair did share a media market (Pechacek 2006).

[15] For example, Linday et al. (2001) evaluate a child autism treatment wherein unlucky randomization could assign all treatment periods to the trial's beginning, and all control periods to the end. If the ten-week study period is developmentally significant for the three-year-old subjects (as individuals become less exchangeable with themselves over time), this unlucky randomization could yield faulty inferences. Goldilocks blocking could prevent this and account for the wear-off period of the drugs. Similar precaution could be built into political communication experiments, where messages have effects that wear off (Gerber et al. 2011), or in experiments that occur within survey windows.

In addition to pretreatment Goldilocks blocking, we provide functionality for researchers to diagnose whether units are likely to interfere with one another post-treatment. Experimental and observational researchers can detect units of differing treatment assignments that are within a specified range of one another to assess their confidence in unit interaction assumptions.

5 Demonstrating Benefits of Blocking

5.1 *Monte Carlo Simulations*

We demonstrate that blocking improves balance in simulation studies using the *p*-value of the d^2 statistic (Hansen and Bowers 2008).[16] Throughout these demonstrations, we employ optimal-greedy pair-blocking relying on the nonresistant Mahalanobis distance.

For each of 1000 replications, we simulate an experiment with 100 units, each of which has three covariate values drawn from three different distributions: one standard normal (X_1), one uniform on [0,1] (X_2), and one χ^2_2 (X_3). Thus, we include covariates that are symmetric and asymmetric, skew and nonskew, and have varying degrees of kurtosis. We assign treatments using complete randomization, random allocation, and blocking. For each replication, we calculate the d^2 *p*-value using the implementation of Bowers, Fredrickson, and Hansen (2010). We display these values as a boxplot for each assignment method in Fig. 3, where 0 represents perfect imbalance and 1 perfect balance. Whereas balance across complete randomizations and random allocations are approximately uniform on [0,1] (as is expected), the blocked *p*-values represent considerably better balance on average. The mean and median *p*-values for the complete randomization and random allocation assignments center around 0.5, while the blocked values are greater than 0.9. In fact, the worst blocked replication *p*-value is 0.47, nearly as great as the average for the other assignment mechanisms.

These *p*-values reflect a fair comparison across simulated replications. Though comparisons of covariate *t*-test *p*-values from different matched or pruned samples may inadvertently incent analysts to throw away valuable data by dropping units (Imai, King, and Stuart 2008), our simulations fix the sample sizes and the units to be included *a priori* and use the *p*-values to compare these fixed samples' properties.

We also compare the balance across assignment methods on each of the three variables independently. The three panels of Fig. 4 display the *p*-values from Kolmogorov–Smirnov, *t*, and Wilcoxon rank sum tests, respectively, for differences in the distributions between treatment and control units for the three covariates under each of the three assignment mechanisms. The pattern of Fig. 3 holds here: for every variable and test, the blocked *p*-values tend to be larger than the values under complete randomization or random allocation. This suggests that the overall balance obtained by blocking displayed in Fig. 3 is not the product of outstanding performance on one type of variable, overshadowing inferiority on others. Though relying on the nonresistant MD for these optimal-greedy blocks might be expected to handle the normal variable better than the other two, the uniform and skewed covariates appear balanced as well.

We model the potential outcome under control as $Y_{i0} = 1 + X_{i1} + 2X_{i2} + 3X_{i3} + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$, and assume a true treatment effect of one for all units. We calculate treatment effect estimates under blocked randomization, complete randomization, and random allocation for each of the 1000 sets of covariates drawn as described above.

Figure 5 displays the estimates. As expected, all three methods' estimates of the difference in means are unbiased, centering around the true treatment effect. The clear improvement in efficiency under blocking is evident in Fig. 5, as the top boxplot has standard and mean absolute deviations less than half of those of the two below, and a similarly tighter range. We obtain similar patterns for a variety of models for the potential outcomes, including interacting the treatment effect with covariate values. We provide an example of efficiency gains under effect heterogeneity in the Supplementary Materials.

[16] Our Supplementary Materials also include a demonstration using an intuitive balance measure.

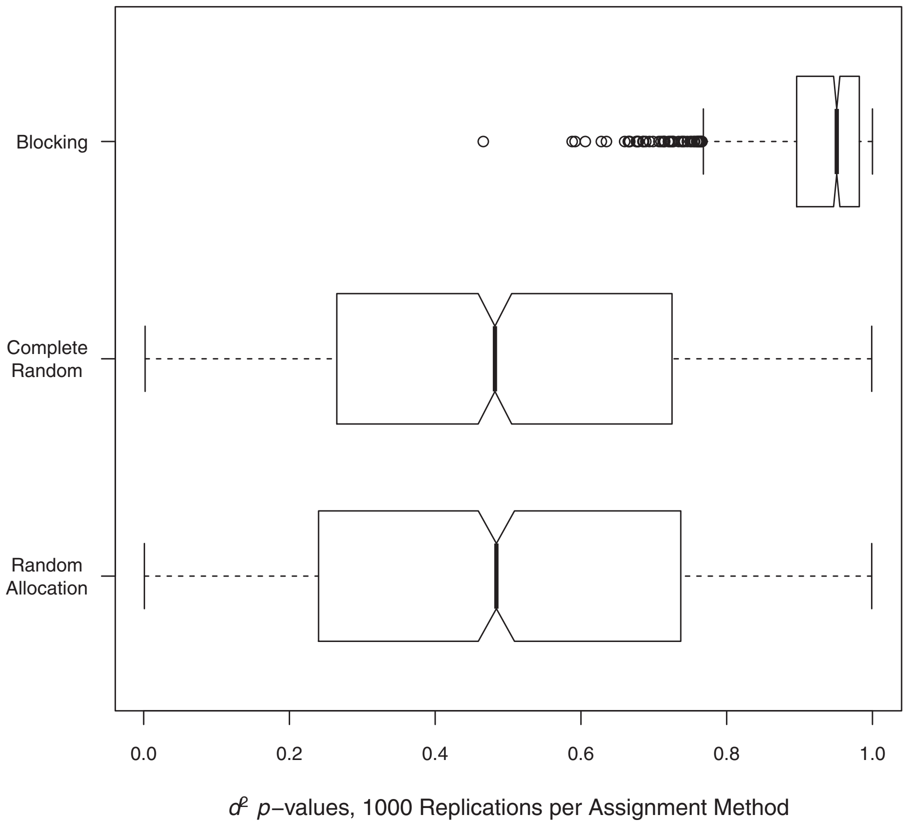


Fig. 3 Blocked experiments better balanced than unblocked ones overall. Boxplots represent overall balance in 1000 sets of covariates common across the methods.

5.2 *Applications*

This section demonstrates two types of improvements: blocked designs generate more balance and efficiency than unblocked ones, and higher-dimensional blocking can improve on lower-dimensional designs.

In each of the three field experiments we reanalyze below, we use the original data and perform optimal-greedy blocking, complete randomization, and random allocation. We attempt to employ the same pretreatment variables used by the original authors before showing how our methods embrace an expanded set of covariates. These experiments in education and elite and mass campaign politics each contribute to larger bodies of work. Rather than attempting to replicate these literatures or the particular findings of these studies, we aim to demonstrate how our blocking methods compare to the actual procedure employed, and how they compare to the usual default procedures. In each case, we show that balance and efficiency could be improved by blocking the experiment.

5.2.1 Perry Preschool Experiment

The Perry Preschool project randomly assigned 123 disadvantaged African American children to attend or not attend preschool during the mid-1960s. Reports from follow-up measures of the children's academic, economic, health, and social trajectories have appeared since the 1980s, and measurement is ongoing. Heckman et al. (2009) detail the randomization procedure fully. This study has been used to estimate downstream rates of return to preschool spending (Belfield et al. 2006; Hansen et al. 2010), and to develop methods for analysis of compromised experiments (Heckman et al. 2009) and multiple outcomes (Anderson 2008).

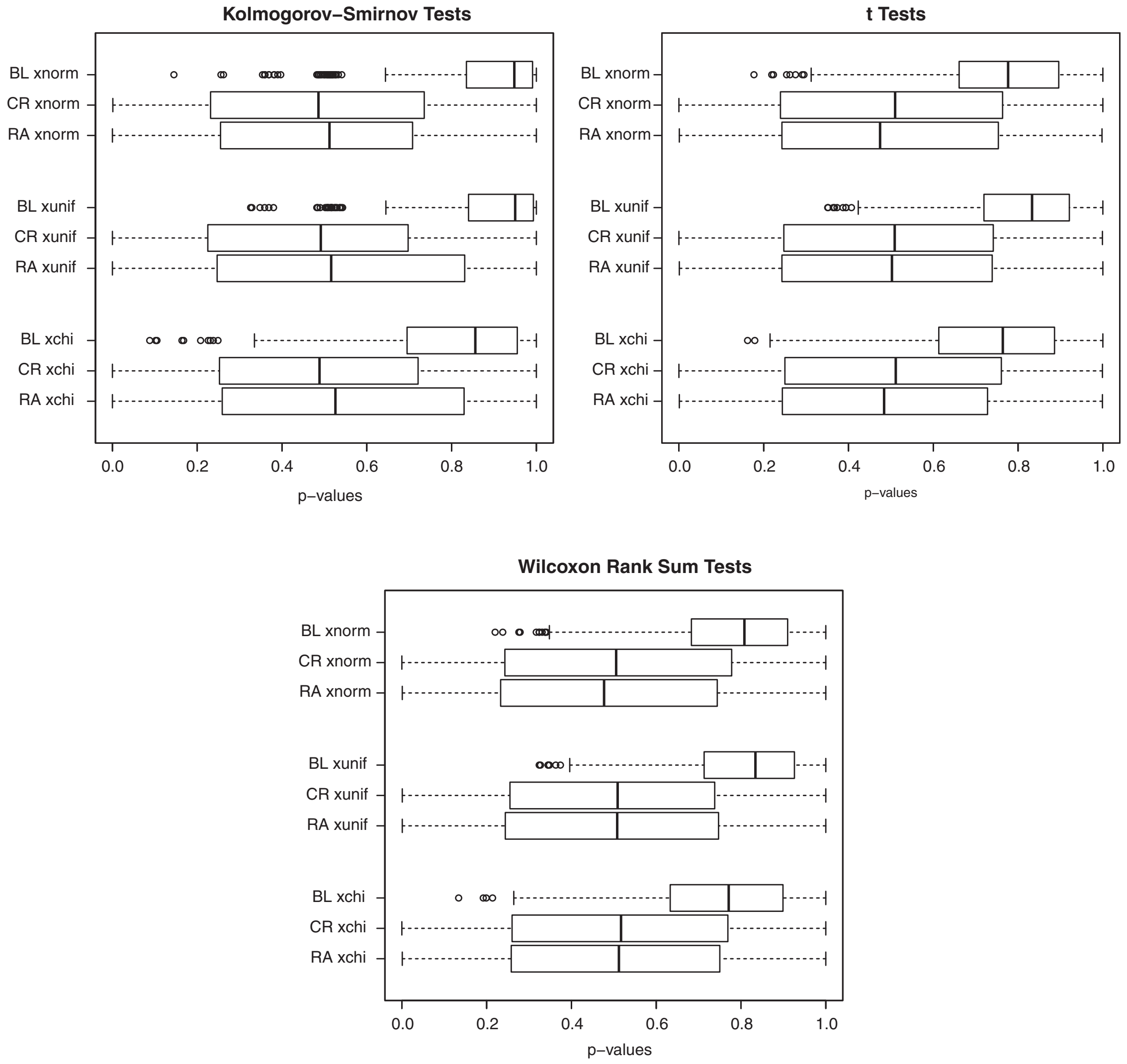


Fig. 4 Blocked experiments better balanced than unblocked ones on each covariate. Boxplots represent balance in 1000 sets of covariates common across the methods. BL = blocked assignment; CR = complete randomization; RA = random allocation. Covariate "xnorm" ~ $\mathcal{N}(0, 1)$; "xunif" ~ $\mathcal{U}[0, 1]$; "xchi" ~χ^2_2.

The project's original randomization incorporated an IQ score rank, sex, and an index of socioeconomic status (SES) at entry. The top left panel of Fig. 6 compares the balance on these three covariates in 100 optimal-greedy blocked, completely randomized, and randomly allocated simulated experiments. The x-axis of this quantile–quantile plot gives the d^2 omnibus p-value from an unblocked randomization of the original Perry data, while the y-axis gives the same value for a randomization that blocks on SES, sex, and IQ rank.

The Perry design yielded good balance in the actual study, with a d^2 p-value for these three variables of about 0.8 (displayed as a horizontal dotted line). Every one of our blocked simulations is more balanced than the actual study. Our method also produces much better balance (ranging from $0.87 \leq p \leq 1$) than both types of unblocked randomizations as well (which are roughly uniform over [0,1], as expected).

To estimate and compare treatment effects from our simulated blocked and unblocked random assignments, we need potential outcomes for each subject under both treatment and control. Thus, we assume that the unobserved Y_{i1} for control units is equal to their observed Y_{i0}, and the unobserved Y_{i0} for treated units is equal to their observed Y_{i1}, a *sharp null* assumption of zero treatment

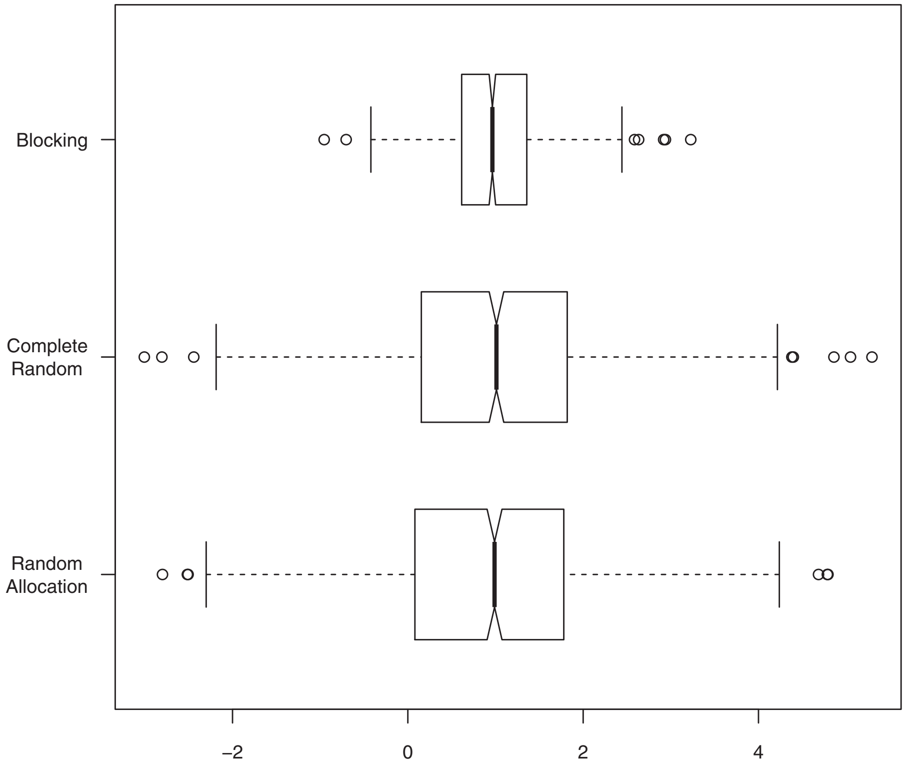


Fig. 5 Blocked experiment estimates more efficient than unblocked ones. Boxplots represent differences in means in 1000 sets of covariates common across the methods. True treatment effect equals one.

effect for every subject.[17] Under this assumption, we estimate the effect for each simulated assignment of the preschool program on the total arrest count at age forty.

The blocked data effect estimates represent an improvement over the estimates from complete randomization and random allocation, whether the estimates are differences in means or the linear regression coefficient of the treatment variable. For differences in mean arrests, the mean and median blocked estimates are about 0.02 and 0.01 from the assumed zero, while the mean and median complete randomization and random allocation estimates range from 0.19 to 0.35 from zero. Thus, the average blocked estimates are 87%–97% closer to zero than the other estimates, and the standard deviation of the blocked estimates is 15%–19% smaller. For linear regression estimates including the covariates (and block indicators for the blocked data), the average estimates are about 39%–67% closer to zero (with the exception of the median random allocation estimate) and the SD of the estimates is 5%–9% smaller. Comparing the blocked difference in means to the regression estimates from complete randomization and random allocation yields similar results: the average blocked estimates all are closer to the assumed zero and have smaller SDs.[18] The right panels of Fig. 6 show the distributions of the regression estimates for total arrests.

[17] Equivalently, $Y_{i1} = Y_{i0} + \tau_i$ with $\tau_i = 0$ for all i. We note that our findings below hold if we assume instead that $\tau_i = c$ for $c \neq 0$; c could be set to the average treatment effect measured in an actual application, for example. This alternative assumption would shift the distributions of treatment effect estimates in Figs. 6 and 7 to the left or right by c. If both (1) our simulations were to assume that $c = 0$ when c is actually large relative to the order of magnitude of the Y_{i0}, and (2) the original experimental assignment successfully approximated exact blocking, then we could have observed slightly more precision for unblocked assignments than blocked ones. However, this observation would be reversed by simply setting c to something near its true value in our simulations. We do not imply that the true effect of the interventions is exactly zero, but rather assume that it is zero and calculate and compare the distributions of estimates under blocked versus unblocked designs.

[18] Duflo, Glennerster, and Kremer (2008) note that either including or excluding block indicators is "acceptable." Omitting the block indicators usually "leads to the exact same point estimates for β but a higher residual variance." However, as they note, if one includes the block indicators, "in a given sample, [the residual variance] could be higher" (p. 3926).

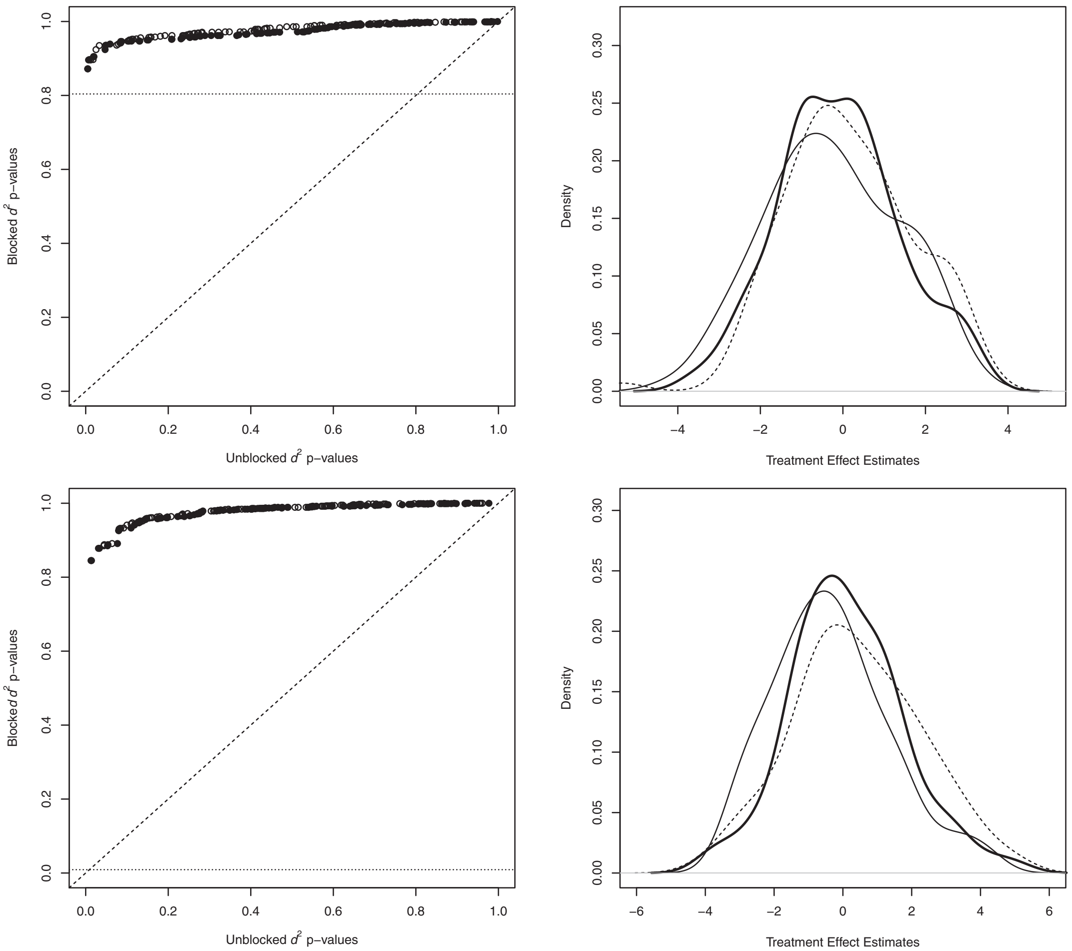


Fig. 6 Perry Preschool application, balance (left) and estimated treatment effects under a sharp null assumption of no effect on total arrests (right). Top row uses SES, sex, and IQ; bottom row uses larger covariate set. Complete randomizations (open circle); random allocations (filled circle). Blocked experiments better balanced than unblocked randomizations (above dashed $y = x$ line) and actual balance obtained in Perry study (dotted line); also more efficient (thick density is narrowest).

Next, we optimal-greedy block on more variables than the Perry study used, but on which comparable treatment and control groups could be desirable. These additional pretreatment measures include the number of siblings, whether the family was receiving AFDC, whether the mother was employed, the occupational skill level of the father, housing density, whether the father lived with the child, the mother's education level, and the raw IQ score.

The bottom row of Fig. 6 shows that the blockings incorporating these variables again create better balance and efficiency than the unblocked experiments. Most striking is the poor aggregate balance in the actual experiment ($p \approx 0.01$), owing to the differences between treatment and control groups in welfare receipt, housing density, father's skill level, and especially mother's employment. The blocked p-values range from about 0.84 to 1, while the unblocked values expectedly cover the zero to one interval uniformly. As the bottom right panel of Fig. 6 shows, the mean linear regression treatment effect estimate from the blocked data is more than 70% closer to zero, and the standard deviation is 3%–12% smaller when compared to the complete randomization and random allocation estimates.

5.2.2 Delegate Persuasion Experiment

During the 2006 race for the leadership of the Canadian Liberal Party, Loewen and Rubenson (2011) conducted a randomized study of direct mail's ability to persuade delegates in conjunction with the onetime front-running campaign of Michael Ignatieff. Considering the pretreatment variables of interest in the original study, the overall balance in the whole experiment and in the smaller respondent sample is strong, with p-values of about 0.75 and 0.44, respectively. This compares well to one hundred unblocked simulated experiments with uniform p-values on [0,1].

Optimal-greedy blocking the data would generate considerably better balance. The top row of Fig. 7 shows that the blocked p-values are always at least 0.98, using seventeen indicator variables representing the respondent delegates' provinces, pledged candidates, and representation of special constituencies. Adding in variables for delegate campaign attention and interest (which are actually measured post-treatment) changes the unblocked and blocked balance very little ($p \approx 0.48$ and $p \geq 0.95$, respectively), as shown in the bottom row of Fig. 7. With about 30% more units, the minimum blocked p-values in this example are even higher than those in the Perry experiment.

Assuming a sharp null treatment effect, we consider the effect of the delegate mailing on the rank delegates later gave to Ignatieff. As in the Perry case, under the blocked design, both the differences in means and the linear regression coefficient estimates are improvements over the complete randomization and random allocation. For differences in mean rank, the SD of the blocked estimates is 3%–18% smaller and the range is 15%–25% smaller. For linear regression estimates including the covariates (and block indicators for the blocked data), the SD of the blocked estimates is 7%–21% smaller and the range is 8%–31% smaller. Comparing the blocked difference in means to the regression estimates from complete randomization and random allocation yields similar results: the average blocked data estimates have smaller ranges and SDs, with improvements of 9%–33%. These results hold when the two additional measures are introduced, with the blocked SDs 12%–21% smaller for the difference in means and 18%–24% smaller for the regression estimate. The right panels of Fig. 7 show the distributions of the regression estimates.

5.2.3 Radio Advertising Experiment

Panagopoulos and Green (2008) provide experimental evidence that nonpartisan radio advertisements can provide greater benefit to challengers than incumbents in local elections. Panagopoulos and Green (2008) buy one of four levels of advertising in forty-nine cities during the campaigns before November 2005 and 2006. As in the previous applications, we compare the balance obtained in one hundred simulated blocked, completely randomized, and randomly allocated experiments as well as that in the actual experiment.

Considering the four advertising levels as manifestations of a plausibly continuous treatment, we use the p-value from the t-statistic of a (generalized) linear regression of each covariate of interest on the treatment variable (Imai and van Dyk 2004). Thus, we compare balance among our simulated experiments one variable at a time. Fig. 8 shows the balance on three pretreatment variables, turnout in the prior election and whether the election was partisan and/or part of a statewide ballot. For the blocked simulations, we first exact blocked on strata defined by year and relative cost of local radio advertising, then used the optimal-greedy algorithm within strata. In this case, like in the Perry and Ignatieff experiments, the blocked p-values tend to be greater than the unblocked ones. The original experiment obtained good balance, with p-values of 0.4, 0.47, and 0.73 on the three variables. The mean and median blocked values range from 0.61 to 0.75, while the mean and median unblocked values range from 0.4 to 0.51. Adopting an approach not available to Panagopoulos and Green (who conducted the original experiments over two years and bought more radio time in less expensive markets), we also replicated this analysis, treating the year-cost strata as indicator variables weighted along with the covariates represented in Fig. 8. We obtained the same patterns of balance, with the primary difference that we nearly perfectly balance the two low-cost stratum indicators.

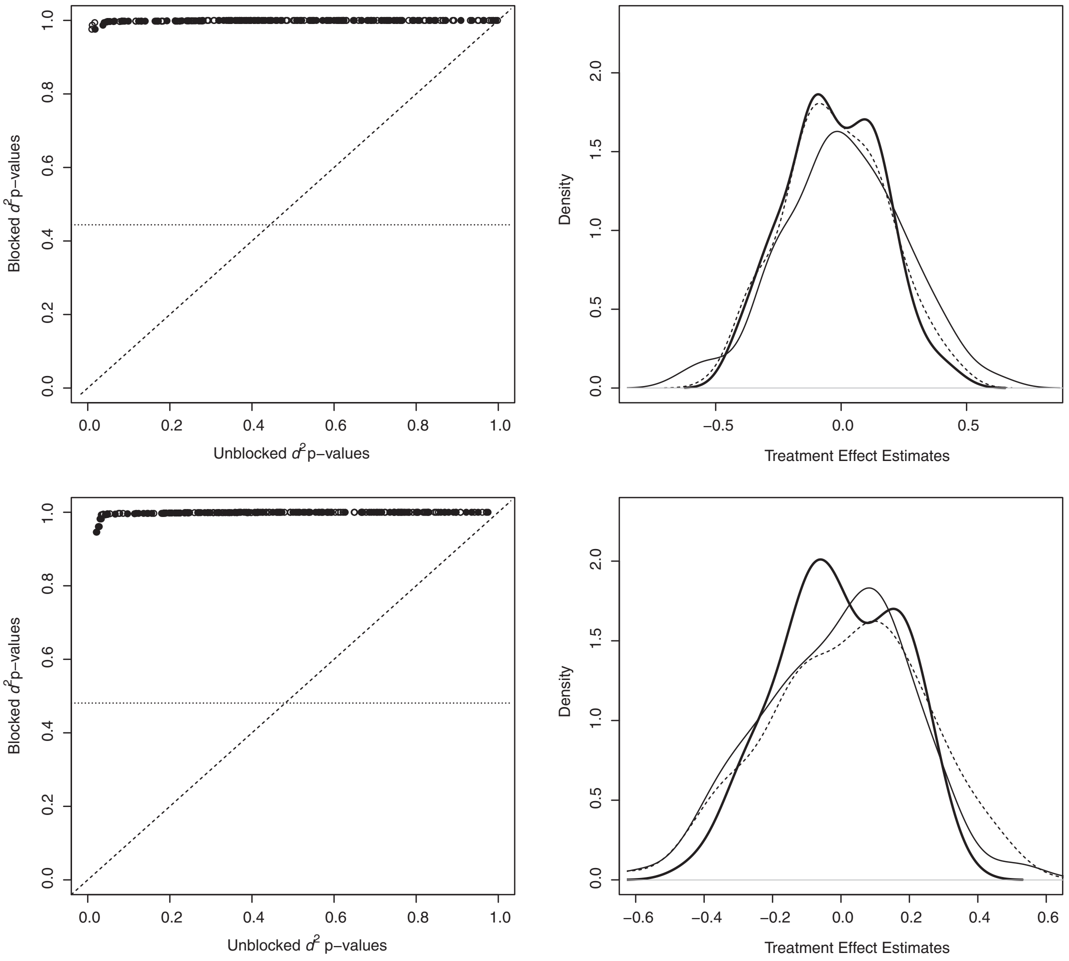


Fig. 7 Ignatieff delegate application, balance (left) and estimated treatment effects (right). The top row uses province, pledged candidate, and special constituency representation; the bottom row adds delegate attention and interest. Complete randomizations (open circle); random allocations (filled circle). Blocked experiments better balanced than unblocked randomizations (above dashed $y=x$ line) and actual balance obtained in Ignatieff delegate study (dotted line); also more efficient (thick density is narrowest).

6 Discussion

Political science has a long and important tradition of conducting randomized field, laboratory, and survey experiments to estimate causal effects of political and policy interventions. Despite its high quality, this literature could be improved going forward by implementing a simple, intuitive design feature—blocking. In addition to the applications in which we show improvement, several group-randomized trials favorably compare their blocked designs to unblocked ones (Freedman et al. 1997).

Compared to ex post conditioning via regression, blocking has favorable properties. As Duflo, Glennerster, and Kremer (2008) write, "blocking is more efficient than controlling ex post for [background] variables, since it ensures an equal proportion of treated and untreated unit [sic] within each block and therefore minimizes variance" (p. 3925). In our applications, difference-in-means and regression treatment effect estimates from the blocked data are more precise than comparable estimates from the completely randomized and random allocation data. We also note that regression adjustment cannot remove substantial biases from comparisons

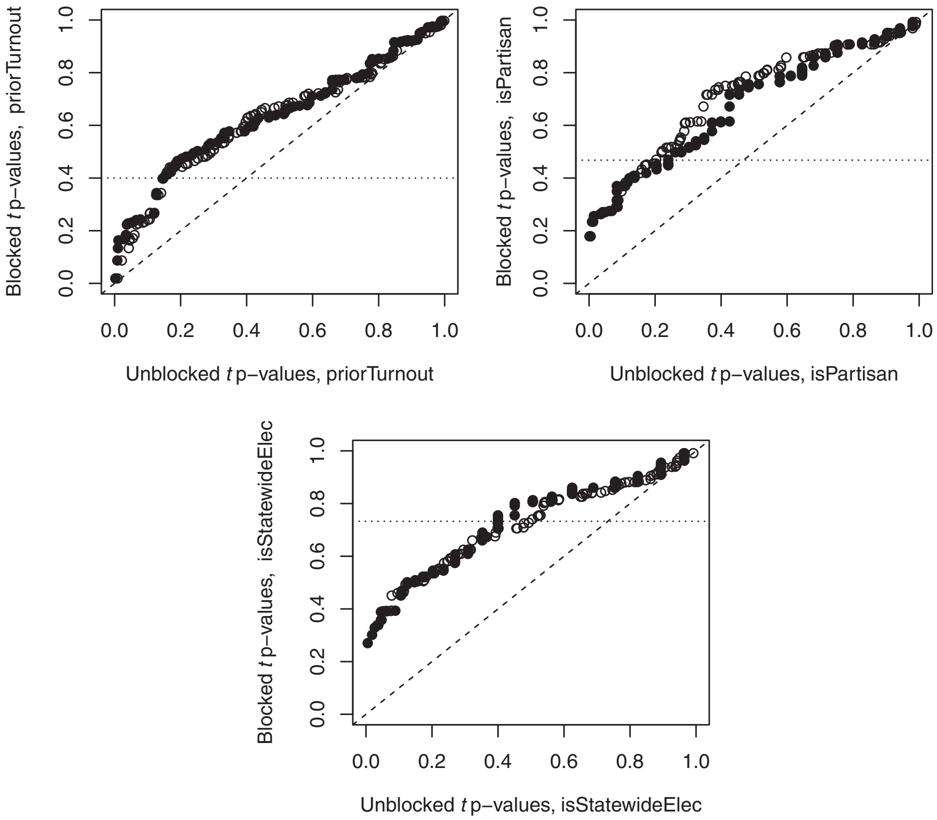


Fig. 8 Radio advertising application, blocked experiments better balanced than unblocked randomizations (above dashed $y = x$ line) and actual balance obtained in radio advertising study (dotted line). Complete randomizations (open circle); random allocations (filled circle).

of very different treatment and control distributions—comparisons that blocking can render unnecessary (Rubin 2001). A similar case has been made regarding confounding and generalized augmented inverse propensity-weighted estimators (Hangartner and Moore 2011). Model dependence also can render inferences from unmatched observational data quite fragile (Ho et al. 2007). In ANCOVA analyses, Casella (2008) notes that ex post adjustment factors should be unrelated to the treatment so that regression adjustment does not increase the variance of the treatment effect estimate over the unadjusted estimate. Blocking helps ensure that the levels of the covariates one wants to include in an ex post regression are, in fact, unrelated to the treatment assignment.

Blocking can incorporate a rich amount of background information into experimental design, whether trialists assign treatments sequentially (Moore and Moore 2012) or at a single point in time. As with any element of design, several choices must be made in practice, including choices of algorithm, covariate weighting, outlier resistance, blocking level, and distances that may contribute to or impede interference between units. Even after treatments are carefully defined, covariates are collected, and blocks are created, subsequent study-specific decisions will always remain. Among these decisions are whether assignment probabilities should be equal across or within blocks and whether implementation efforts should be uniform across units assigned to treatment.

Funding

National Science Foundation, IGERT training grant 0333403.

References

Anderson, Michael L. 2008. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and early training projects. *Journal of the American Statistical Association* 103:1481–95.

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. 2003. Principal stratification approach to broken randomized experiments. *Journal of the American Statistical Association* 98:299–323.

Belfield, Clive R., Milagros Nores, Steve Barnett, and Lawrence Schweinhart. 2006. The high/scope Perry Preschool Program. *Journal of Human Resources* 41:162–90.

Boruch, Robert, Henry May, Herbert Turner, Julia Lavenberg, Anthony Petrosino, Dorothy De Moya, Jeremy Grimshaw, and Ellen Foley. 2004. Estimating the effects of interventions that are deployed in many places. *American Behavioral Scientist* 47:608–33.

Bowers, Jake. 2011. Making effects manifest in randomized experiments. In *Cambridge handbook of experimental political science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, 459–80. Cambridge, UK: Cambridge University Press.

Bowers, Jake, Mark Fredrickson, and Ben Hansen. 2010. *RItools: Randomization inference tools*. R package version 0.1–11, http//www.jakebowers.org/RItools.html (accessed August 16, 2012).

Bullock, John G. 2011. Elite influence on public opinion in an informed electorate. *American Political Science Review* 105:496–515.

Casella, George. 2008. *Statistical design*. New York: Springer.

Cochran, William G., and Donald B. Rubin. 1973. Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A* 35:417–46.

Donner, Allan, and Neil Klar. 2000. *Design and analysis of cluster randomization trials in health research*. London: Arnold Publishers.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. Using randomization in development economics research: A toolkit. In *Handbook of Development Economics*, ed. T. Paul Schultz, Vol. 4, 3895–962. Amsterdam: Elsevier, B.V.

Epstein, Lee, and Gary King. 2002. The rules of inference. *University of Chicago Law Review* 69:1–133.

Freedman, Laurence S., Mitchell H. Gail, Sylvan B. Green, and Donald K. Corle. 1997. The efficiency of the matched-pairs design of the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials* 18:131–9.

Gail, Mitchell H., David P. Byar, Terry F. Pechacek, and Donald K. Corle. 1992. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials* 13:6–21.

Gerber, Alan S., James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2011. How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. *American Political Science Review* 105:135–50.

Goldstein, Daniel G., Kosuke Imai, Anja S. Göritz, and Peter M. Gollwitzer. 2010. Nudging turnout: Mere measurement and implementation planning of intentions to vote. Manuscript.

Gosnell, Harold F. 1927. *Getting out the vote: An experiment in the stimulation of voting*. Chicago, IL: University of Chicago Press.

Green, Donald P., and Holger L. Kern. Forthcoming 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly*.

Greevy, Robert, Bo Lu, Jeffrey H. Silber, and Paul Rosenbaum. 2004. Optimal multivariate matching before randomization. *Biostatistics* 5:263–75.

Hangartner, Dominik, and Ryan T. Moore. 2011. Generalizing and stabilizing the augmented inverse propensity weighted estimator. Proceedings of the Midwest Political Science Association Annual Meeting.

Hansen, Ben B. 2004. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* 99:609–18.

Hansen, Ben B., and Jake Bowers. 2008. Covariate balance in simple, stratified, and clustered comparative studies. *Statistical Science* 23:219–36.

Hansen, Ben B., and Stephanie Olsen Klopfer. 2006. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* 15:609–27.

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2009. A reanalysis of the high-scope Perry Preschool Program, Unpublished manuscript.

———. 2010. The rate of return to the high-scope Perry Preschool Program. *Journal of Public Economics* 94:114–28.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15:199–236.

Holland, Paul. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81(396):945–60.

Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. Designing and analyzing randomized experiments: Application to a Japanese election survey experiment. *American Journal of Political Science* 51:669–87.

Hudgens, Michael G., and M. Elizabeth Halloran. 2008. Toward causal inference with interference. *Journal of the American Statistical Association* 103:832–42.

Hyde, Susan. 2010. Experimenting in democracy promotion: International observers and the 2004 Presidential elections in Indonesia. *Perspectives on Politics* 8:511–27.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2011. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association* 106:345–61.

———. 2012. Causal inference without balance checking: Coarsened exact matching. *Political Analysis* 20:1–24.

Ichino, Nahomi, and Matthias Schündeln. 2012. Deterring or displacing electoral irregularities? Spillover effects of observers in a randomized field experiment in Ghana. *Journal of Politics* 74:292–307.

Imai, Kosuke, and David A. van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* 99:854–66.

Imai, Kosuke, Gary King, and Clayton Nall. 2009. 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, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 171:481–502.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105:765–89.

Imbens, Guido W. 2011. Experimental design for unit and cluster randomized trials. Manuscript prepared for the International Initiative for Impact Evaluation.

Keele, Luke, and Jason W. Morgan. 2011. Stronger instruments by design. Poster presented at the 28th Annual Summer Meeting of the Society for Political Methodology.

King, Gary. 1995. Replication, replication. *PS: Political Science and Politics* 28:444–52.

King, Gary, Emmanuela Gakidou, Kosuke Imai, Jason Lakin, Ryan T. Moore, Clayton Nall, Nirmala Ravishankar, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Ávila, and Héctor Hernández Hernández Llamas. 2009. Public policy for the poor? A randomized assessment of the Mexican universal health insurance program. *Lancet* 373:1447–54.

King, Gary, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, and Héctor Hernández Llamas. 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* 26:479–509.

Lachin, John M. 1988. Properties of Simple Randomization in Clinical Trials. *Controlled Clinical Trials* 9:312–26.

Linday, Linda A., J. A. Tsiouris, Ira L. Cohen, Richard Shindledecker, and Robert DeCresce. 2001. Famotidine treatment of children with autistic spectrum disorders: Pilot research using single-subject research design. *Journal of Neural Transmission* 108:593–611.

Loewen, Peter John, and Daniel Rubenson. 2011. For want of a nail: Negative persuasion in a party leadership race. *Party Politics* 17:45–65.

Lu, Bo, Elaine Zanutto, Robert Hornik, and Paul R. Rosenbaum. 2001. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* 96:1245–53.

Mason, Robert L., Richard F. Gunst, and James L. Hess 1989. *Statistical design and analysis of experiments: With applications to engineering and science*. New York: Wiley.

Mebane, Walter R. J., and Jasjeet S. Sekhon. 1998. GENetic Optimization Using Derivatives (GENOUD).

Moore, Ryan T. 2012a. *blockTools: Blocking, assignment, and diagnosing interference in randomized experiments*. R package version 0.5–6, http://www.wustl.edu/software.blockTools.him (accessed August 16, 2012).

———. 2012b. *Replication data for: Multivariate continuous blocking to improve political science experiments*. http://hdl.handle.net/1902.1/18341, IQSS Dataverse Network [Distributor] V1 [Version].

Moore, Ryan T., and Sally A. Moore. 2012. Blocking for sequential political experiments. Manuscript in preparation.

Murray, David M. 1998. *Design and analysis of group-randomized trials*. New York: Oxford University Press.

National Cancer Institute. 1995. NCI Monograph #6: Community-based interventions for smokers: The COMMIT field experience. Technical Report 95-4028, National Institutes of Health.

Paluck, Elizabeth Levy, and Donald P. Green. 2009. Deference, dissent, and dispute resolution: An experimental intervention using mass media to change norms and behavior in Rwanda. *American Political Science Review* 103:622–44.

Panagopoulos, Costas, and Donald P. Green. 2008. Field experiments testing the impact of radio advertisements on electoral competition. *American Journal of Political Science* 52:156–68.

Pechacek, Terry F. 2006. Personal communication, Centers for Disease Control's Associate Director for Science for the Office on Smoking and Health, 5 December.

R Development Core Team. 2012. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rosenbaum, Paul R. 2007. Interference between units in randomized experiments. *Journal of the American Statistical Association* 102:191–200.

———. 2010. *Design of observational studies*. New York: Springer.

Rousseeuw, Peter J. 1985. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications* 8:283–97.

Rousseeuw, Peter J., and Bert C. van Zomeren. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85(411):633–9.

Rubin, Donald B. 1980. Bias reduction using Mahalanobis-metric matching. *Biometrics* 36:293–8.

———. 1990. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 25:279–92.

———. 2001. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2:169–88.

Sekhon, Jasjeet S. 2011. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software* 42:1–52.

Sobel, Michael 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):1398–407.

Tamhane, Ajit C. 2009. *Statistical analysis of designed experiments: Theory and applications*. Hoboken, NJ: John C. Wiley & Sons.

Wantchekon, Leonard. 2003. Clientelism and voting behavior: Evidence from a field experiment in Benin. *World Politics* 55(3):399–422.