Advance Access publication June 17, 2013

Political Analysis (2013) 21:507–523
doi:10.1093/pan/mpt007

Blocking for Sequential Political Experiments

Ryan T. Moore
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 (corresponding author)

Sally A. Moore
VA Puget Sound Health Care System—Seattle Division, University of Washington, Department of Psychiatry and Behavioral Sciences, and Evidence-Based Treatment Centers of Seattle, 1200 Fifth Ave, Suite 800, Seattle, WA 98101

Edited by Jonathan Katz

In typical political experiments, researchers randomize a set of households, precincts, or individuals to treatments all at once, and characteristics of all units are known at the time of randomization. However, in many other experiments, subjects "trickle in" to be randomized to treatment conditions, usually via complete randomization. To take advantage of the rich background data that researchers often have (but underutilize) in these experiments, we develop methods that use continuous covariates to assign treatments sequentially. We build on biased coin and minimization procedures for discrete covariates and demonstrate that our methods outperform complete randomization, producing better covariate balance in simulated data. We then describe how we selected and deployed a sequential blocking method in a clinical trial and demonstrate the advantages of our having done so. Further, we show how that method would have performed in two larger sequential political trials. Finally, we compare causal effect estimates from differences in means, augmented inverse propensity weighted estimators, and randomization test inversion.

1 Introduction to Sequential Randomization

In typical political experiments, researchers randomize a set of households, precincts, or individuals to treatments all at once, and characteristics of all units are known at the time of randomization. However, in a *sequential experiment*, assignment occurs when units enter the study at different, often unpredictable times. Assignment in such experiments is usually done by complete randomization. In political psychology laboratory experiments, for example, subjects usually elect when to participate, and neither their participation nor its order is known to researchers *a priori*. Patients in clinical trials similarly enter studies asynchronously, as their diagnoses are made. In experiments in politics, medicine, public health, education, and elsewhere, new units can be discovered in the midst of ongoing trials.

These common, important research situations differ from nonsequential experiments in which researchers know all participating units and conduct a single randomization. In nonsequential experiments, background data (available for all units simultaneously) can be used to block the experiment, creating homogeneous groups within which treatments can be assigned (Moore 2012). Blocking can improve balance, reduce estimation error, and increase the precision of causal estimates in sequential and nonsequential experiments. However, at a given moment during a sequential experiment, the researcher only has information on the units that have already arrived and the unit that has just appeared to be assigned—the *current* unit—but not future units. Since the

Authors' note: We thank Nicholas Beauchamp, Jens Hainmueller, Kosuke Imai, Rebecca Morton, Kevin Quinn, and the participants in EGAP 6 and ICHPS 9 for helpful suggestions. The replication archive is available at Moore and Moore (2013). Supplementary materials for the article are available on the *Political Analysis* Web site.

sequential experiment has less information about the eventual full sample to use in assigning treatments than does the otherwise equivalent nonsequential experiment, different methods are required to exploit this information.

Sequential randomized experiments encompass a variety of designs,[1] including those where assignment probabilities are fixed throughout the trial, change based on the number of units already in the treatment groups, change based on the covariate profiles of previous units and the current unit, and change as a function of previous units' outcomes. These are *non-adaptive*, *treatment-adaptive covariate-adaptive*, and *response-adaptive* randomizations, respectively (Chow and Chang 2007). Both covariate- and response-adaptive designs vary assignment probabilities based on knowledge about previous units. On the other hand, treatment-adaptive designs may ignore the researcher's detailed data on individual subjects.

We focus on covariate-adaptive designs. In many cases, outcomes are not realized until long after assignment, so previous outcome data are unavailable when subsequent assignments are made. Experiments interested in whether patients survive five years or information effects persist through the end of a campaign provide examples. Even when prior unit outcome information is not available during randomization, covariate data may still be powerful.[2]

We describe covariate-adaptive blocking for sequential randomizations, showing how researchers can incorporate rich background data, even in sequential trials where assignment takes place shortly after a subject arrives. We use simulation to explore advantages of new approaches over complete randomization in a variety of data contexts. To develop foundations for why new approaches are warranted, the next section introduces the canonical biased coin and minimization approaches to blocked sequential randomization using discrete covariates with a few levels each. We describe experiments with two and several treatment conditions to illustrate limitations of common approaches. In particular, researchers often want to incorporate more, finer covariates than is practical in canonical designs.

Section 3 proposes ways to integrate discrete and continuous covariate information into sequential randomizations, and enumerates some decisions experimentalists must make. Section 4 applies the methods to four types of continuous simulated data: uncorrelated and correlated multivariate normal (MVN) data, MVN data with extreme outliers, and extremely bimodal data. In each case, sequentially blocked experiments exhibit more balance and precision than completely randomized experiments. Section 5 tests several methods using simulated data that anticipated a then-upcoming clinical trial which we conducted. Based on the results, we selected and deployed one method in the actual trial. We show how well the chosen method performed when we implemented it. Next, we apply this method to two much larger political studies, again showing improvements. We then compare causal effect estimates from differences in means, augmented inverse propensity weighted (AIPW) estimators, and randomization test inversion.

2 Using Covariate Data in Sequential Randomization

To most meaningfully block a sequential experiment, researchers must collect background data prior to randomization.[3] Since subjects may be assigned to conditions very shortly after initiating contact with an experimenter, these data should be collected quickly and incorporated into randomization in an automated way. Productive data collection from recent political experiments includes a "short background questionnaire" (Chong and Druckman 2007), rich background data from an online panel (Malhotra and Kuo 2008), and a subset of the information from a pre-assignment web survey (Horiuchi, Imai, and Taniguchi 2007).

[1] By "sequential experiment" we denote an experiment in which subjects arrive serially, rather than a series of several experiments or data collection points, as in Morton and Williams (2010).

[2] With high-stakes treatments, early outcome information may influence assignment probabilities. If a drug therapy demonstrates its clear superiority to alternatives, care ethics may prompt researchers to assign the therapy more than would an otherwise optimal design (Zelen 1969; Rosenberger and Sverdlov 2008).

[3] We use "blocking" to refer to our covariate-adaptive designs because, as in nonsequential blocking, covariate information is used to create comparable groups defined by treatment conditions, and the method for doing so biases the randomization against the condition to which similar units are assigned.

Sequential randomization methods have various goals, including creating similar covariate distributions in the treatment groups, creating equal group sizes, and avoiding assignments that researchers can predict. We focus on the first of these. Promoting balance across treatment group covariate distributions benefits from richer covariate information than simply the number of units already assigned to each group. For discussion of benefits from covariate-balancing blocking in general, see Moore (2012), King et al. (2007), and Imai, King, and Nall (2009). On the other hand, permuted block randomizations, for example, seek to balance treatment group sizes and have endured recent criticism (Schulz and Grimes 2002a, 2002b).

Modern covariate-adaptive randomizations begin with "biased coin" designs, which set the current unit's treatment assignment probability using its entire covariate profile at once. Efron (1971) describes the sample-size-balancing properties of biased coin designs and discusses their ability to overcome biases like time trends in clinical trials.[4] An alternative, "minimization," considers covariates one at a time and can limit marginal imbalance across an arbitrarily large set of covariates (Pocock and Simon 1975). However, both biased coins and minimization require discrete, replicated levels of the prognostic factors and focus on the number of units assigned to particular treatments. We develop the intuition of the Efron and Pocock–Simon approaches before exploring some related concerns. We propose alternatives for addressing these concerns in Section 3.

2.1 *Balancing Condition Sizes Using Discrete Covariates*

The biased coin and minimization approaches utilize discrete covariates with few levels (Efron 1971; Pocock and Simon 1975; Whitehead 1997). Consider an experiment with T treatment conditions indexed by t, letting $t \in \{1, 2\}$ correspond to control and treatment conditions, respectively. When a unit enters the experiment, but prior to randomization, J discrete covariates are measured, indexed by j; the jth covariate has l_j levels, indexed by $i \in \{1, \ldots, l_j\}$. When the current unit arrives, all previous units' covariate values and treatment assignments are known.

Consider the example experiment in Table 1. The first covariate takes values of 0, 1, or 2; the second takes values of 0 or 1. When a new unit arrives, 22 units have been assigned treatment. The columns of Table 1 represent the covariate measures for the units; the rows represent treatment assignments. Each cell contains the number of units with the value i on covariate j assigned to treatment t, denoted n_{ijt}. Each unit appears in the table twice—once for each covariate value it has. Let $\mathbf{p}$ represent a covariate profile, such as (0,0), and index the profiles by $p \in \{1, \ldots, \prod_{j=1}^{J} l_j\}$. Table 1 includes $3 \times 2 = 6$ profiles.

To determine the current unit's treatment assignment, a biased coin considers only the units with the same covariate profile $\mathbf{p}$ as the current unit, making the treatment with fewer $\mathbf{p}$ units more likely than the one with more. On the other hand, minimization calculates a score, S_p, for this unit using its $\mathbf{p}$ and some function of the n_{ijt}. The score aggregates information across covariates and summarizes the sample size imbalance between the treatment conditions for the current unit's covariate values (Pocock and Simon 1975). With two treatment conditions, S_p could be $\sum_{j=1}^{J} \frac{n_{ij1}-n_{ij2}}{n_{ij1}+n_{ij2}+1}$ or a weighted version thereof.

Minimization is valid for any $J \geq 1$, and can incorporate information about (1) the direction of the covariate imbalance between treatment and control (e.g., are more women in treatment or control?); (2) the size of the imbalance (how many more?); and (3) the scale of each covariate's imbalance with respect to the sample size (what percentage of women does this imbalance represent?). The univariate biased coin is a rough minimization measure, with $S_p = \text{sgn}(n_{ij1} - n_{ij2})$ ignoring information of types (2) and (3).

Next, the randomization protocol proceeds. Under the biased coin, if $S_p > 0$, assign the current unit to treatment with some probability $\pi > \frac{1}{2}$. Some consensus exists that "the choices $\frac{2}{3}$ or $\frac{3}{4}$ are

[4]Recent work reviews randomization in clinical trials (Harrington 2000), categorizes the goals of allocation methods (Kalish and Begg 1985), introduces a decision-theoretic Bayesian approach (Ball, Smith, and Verdinelli 1993), and summarizes complete, permuted block, and biased-coin randomization procedures, often recommending a design due to sample size and the degree of blinding (Lachin, Matts, and Wei 1988).

Table 1 The minimization approach to balancing a sequential experiment with discrete covariates

	Covariate 1 ($j = 1$)			*Covariate 2* ($j = 2$)	
	0	*1*	*2*	*0*	*1*
Control	2	4	3	7	2
Treatment	5	5	3	2	11

Note. Each unit appears twice, once for each of two covariates.

quite suitable" (Efron 1971; Whitehead 1997). For $S_p = 0$, let $\pi = \frac{1}{2}$; for $S_p < 0$, let $\pi < \frac{1}{2}$. These assignments tend to put units into treatment conditions with fewer units identical to the current unit, thus marginally balancing covariates between conditions.

With more than two treatment conditions, minimization defines an s_{jt} for each condition for the current unit. For example, if the current unit shares attributes with prior units, let $s_{jt} = \frac{n_{ijt}}{n_{ij1}+\ldots+n_{ijT}}$ and $S_{pt} = \sum_{j=1}^{J} s_{jt}$. Next, order the treatments by decreasing levels of S_{pt}, and assign them increasing probabilities of selection, π_t, with $\sum_{t=1}^{T} \pi_t = 1$. Finally, assign the unit according to these probabilities.

2.2 *Limitations and Methodological Opportunities*

Biased coin and minimization procedures, while often useful, have distinct limitations. First, treating every covariate independently may not protect against imbalance in covariate interactions. The experiment in the upper panel of Table 2 appears perfectly balanced, when a Democratic male enters the experiment. The minimization procedure above would produce $S_p = 0$, and the new Democratic male is randomized to each condition with probability $\frac{1}{2}$.

Unfortunately, the covariate-wise balance in the upper panel of Table 2 could mask perfect interaction imbalance, shown in the lower panel. The new Democratic male should have a higher probability of being assigned to the control, but minimization does not tell us this.

To promote balance in the interaction, the two binary covariates could be combined into a single four-category covariate ("Republican Male," "Democratic Male," etc.). Then, a biased coin procedure will protect against assigning the Democratic male to treatment with all those identical to him. Zelen (1974) suggests using permuted blocks within strata. However, these solutions create new problems. Even with only a few covariates with a few levels each, the number of possible profiles, $\prod_{j=1}^{k} l_j$, gets large quickly. Sequential randomization encounters a dimensionality curse similar to that in nonsequential experiments. If a new unit's profile has not yet appeared in an experiment, the biased coin ignores his covariate information.

Despite these drawbacks, researchers may require that a few discrete covariates be more certain to be balanced. Our software described below allows experimentalists to block on some discrete covariates exactly. Our implementation incorporates the minimization procedure described above, taking into consideration the full ranking of treatment group sizes (Pocock and Simon 1975). When only discrete covariates are exact blocked, we assign units after the first to treatment t with probability $\Pr(t^* = t) = \left(1 - \frac{n_{ijt}}{\sum_{t=1}^{T} n_{ijt}}\right) \Big/ \sum_{t=1}^{T} \left(1 - \frac{n_{ijt}}{\sum_{t=1}^{T} n_{ijt}}\right)$.[5]

Older approaches require either categorical covariates or coarsening (Atkinson 2003). Drastically coarsening covariates to a few levels may ignore relevant information and limit an experimental design unnecessarily. On the other hand, mapping continuous covariates onto many categories reintroduces the dimensionality problem described above. Using rich information to foster covariate balance is a key motivation for collecting pre-treatment data. Thus, where

[5]With $T = 2$, this blocks the first two identical units and performs random allocation (Lachin 1988).

Table 2 Consistent marginal and joint distributions of two binary covariates

	Sex		*Party*	
	M	*F*	*Rep*	*Dem*
Control	3	3	3	3
Treatment	3	3	3	3

	Rep M	*Dem M*	*Rep F*	*Dem F*
Control	3	0	0	3
Treatment	0	3	3	0

Note. The upper panel reflects the minimization approach and suggests balance between treatment conditions. However, the lower panel makes clear that results will represent isolated parts of the covariate support.

continuous predictors affect outcomes, we encourage researchers to augment any needed exact blocking with methods for continuous data.

3 Sequential Blocking with Continuous Covariates

The standard approaches sequentially block a few discrete covariates by counting the number of identical units assigned to each condition. The current unit is most likely to be assigned to the condition with the fewest identical units. With a continuous covariate, identical units will not exist. Thus, to avoid losing information through coarsening while still defining a feasible procedure, we propose methods for sequential randomization that exploit relationships between the current unit's covariate profile and those of all previously assigned units. To do so, we must decide how to measure similarity between two units, aggregate pairwise similarity to compare the current unit to each treatment condition, and set a probability of assignment to each condition using the aggregate similarities.

A design may or may not prioritize *exact blocking* variables. A design that prioritizes certain discrete (or coarsened) variables seeks first to balance those, then to balance other covariates. Our software implementation allows the user to specify exact blocking covariates and omit or include any continuous covariates. Without prioritized exact covariates (or within strata defined by an exact covariate subprofile, such as the columns of Table 2's lower panel), we define the similarity between units using the Mahalanobis distance (MD) between units q and r with covariate vectors $\mathbf{x}_q$ and $\mathbf{x}_r$: $\text{MD}_{qr} = \sqrt{(\mathbf{x}_q - \mathbf{x}_r)' \Sigma^{-1} (\mathbf{x}_q - \mathbf{x}_r)}$.

For the aggregation, we implement the mean, median, and trimmed mean of the pairwise MDs between the current unit and the units in each treatment condition. Alternatives include a mean or median weighted by sample size. Index the units with treatment condition t using $r \in \{1, \ldots, R\}$, and define for each condition t an average MD between the current unit q and the units already assigned t. Using the mean as an example, $\overline{\text{MD}}_{qt} = \frac{1}{R} \sum_{\text{r}=1}^{\text{R}} \text{M}D_{qr}$. If the average MD from the current unit is 2 in the control condition and 5 in the treatment condition, then the control condition looks more like the current unit than does the treatment condition.[6]

We investigate several ways to use the pairwise MDs to set the probability of assigning the current unit to each condition. The first set of techniques ignores the covariate and condition-count information and assigns the conditions with fixed, perhaps unequal, probabilities. The second set of strategies, which we call `ktimes` methods, calculates $\overline{\text{MD}}_{qt}$ for all t, then sorts the treatment conditions by these averages. Since this score represents dissimilarity, we bias the randomization toward conditions with high scores, reversing the rankings relative to the scores in Section 2. Next, for values of $k \in \{2, 3, \ldots, 6\}$ we assign the condition with the highest average MD a probability k

[6]One could also consider only a subset of $\{1, \ldots, R\}$ defined by a few nearest-neighbors or within a caliper.

times larger than the other $T-1$ assignment probabilities. For example, using $k=2$ in a two-condition experiment assigns the more dissimilar condition with $\pi_t = \frac{2}{3}$, the more similar condition with $1-\pi_t = \frac{1}{3}$. Using $k=4$ in a three-condition experiment assigns the most dissimilar condition with $\pi_t = \frac{2}{3}$ and the other two conditions with probability $\frac{1-\pi_t}{T-1} = \frac{1}{6}$.

A third approach uses the distributions of the MDs most directly. Instead of fixing an assignment probability for each condition or each condition rank (as in the first two mappings), here the assignment probabilities change with the average MD values and the number of units in each condition. Like the `ktimes` approaches, our first method ensures assignment bias toward conditions with larger average MDs; it assigns the current unit to treatment t with probability $\frac{\overline{\text{MD}}_{qt}}{\sum_{t=1}^{T}\overline{\text{MD}}_{qt}}$. We also consider two methods that use the MD, but are not certain to bias assignment toward conditions with larger MDs. The first assigns conditions proportionally to the sum of the MDs between that condition and the current unit; condition t is assigned probability $\sum_{r=1}^{R}\text{MD}_{qr}$, normalized by the total $\sum_{t=1}^{T}\sum_{r=1}^{R}\text{MD}_{qr}$. An alternative uses the squares of the sums of the MDs. We provide example calculations in the supplementary materials.

4 Improving Balance and Precision in Simulated Data

We employ simulation to assess the balance and precision of our approach. Others advocate simulation for evaluating covariate-adaptive designs (Rosenberger and Lachin 2002).

The software employed below lets researchers implement any of the techniques we describe in R (R Core Team 2013). To facilitate accurate pre-treatment data collection for trickling-in subjects, we include an optional interactive interface to simplify input, perform needed calculations, assign a treatment condition, and store the data transparently (Braucht and Reichardt 1993). Our supplementary materials offer an example of interacting with this query-based interface. We designed the interface to minimize errors in the study described in Section 5, knowing that, as is often true in the social sciences, intake would be done by research assistants. The interface is general, allowing researchers to apply our methods in any sequential experiment.

Anticipating Section 5's application, we simulate a trial with 40 subjects and sequentially block them using the mean MD to compare the conditions, and the $k=2$ method.[7] To examine the advantages of sequential blocking over complete randomization, we compare the covariate balance and efficiency in the blocked experiment to one hundred completely randomized experiments using the same covariates. We repeat this for one hundred different sets of forty units, yielding 10,000 completely randomized assignments for comparison.

As a summary measure of balance, we employ the p-value of the χ^2 statistic from the omnibus d^2 measure of the difference between the treated and control distributions (Hansen and Bowers 2008).[8] Such p-values reflect a fair comparison across simulations (Moore 2012).

4.1 *Uncorrelated MVN Data*

We first draw two MVN covariates using an identity matrix for the covariance. The left panel of Fig. 1 displays the balance results, demonstrating that in the hundred sets of covariates used, sequential blocking performed better on average than complete randomization. After all forty subjects have been assigned in the median experiment, sequential blocking produces more balance than 74% of the complete randomizations. As a benchmark, comparing complete randomizations to other complete randomizations, we would expect each panel of Fig. 1 to be symmetric

[7] The replication archive is available at Moore and Moore (2013).

[8] We also measure imbalance by $|H'Z|$ and obtain similar results (Atkinson 2003). Using the notation of Atkinson (2003, 181), let $i \in \{1, \ldots, N\}$ index units, $j \in \{1, \ldots, J\}$ index treatments, α_j be the treatment effect, $\mathbf{z}_i'$ be the unit's covariate profile, $\mathbf{Z}$ be the $(n+1)\times(q-1)$ matrix of data for the first $n+1$ units, $\mathbf{x}_i' \equiv (1 \quad \mathbf{z}_i')$ be the unit's profile augmented with a 1, and $\mathbf{X}$ be the $(n+1)\times q$ matrix of stacked profiles. Then, the outcome is modeled as $y_i = \alpha_j + \mathbf{z}_i^T\theta + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$. The expectation is $E(Y_n) = H_n\alpha + Z_n\theta$, where H_n is the $n\times t$ matrix with units as rows and treatment condition indicators as columns, and Z_n is the matrix of prognostic factors. If $J=2$, the parameter of interest is $\Delta \equiv \alpha_1 - \alpha_2$. Then, we write $E(Y) = a\Delta + \mathbf{X}\beta$, where a is a vector of 1s and –1s defining the linear contrast of interest.

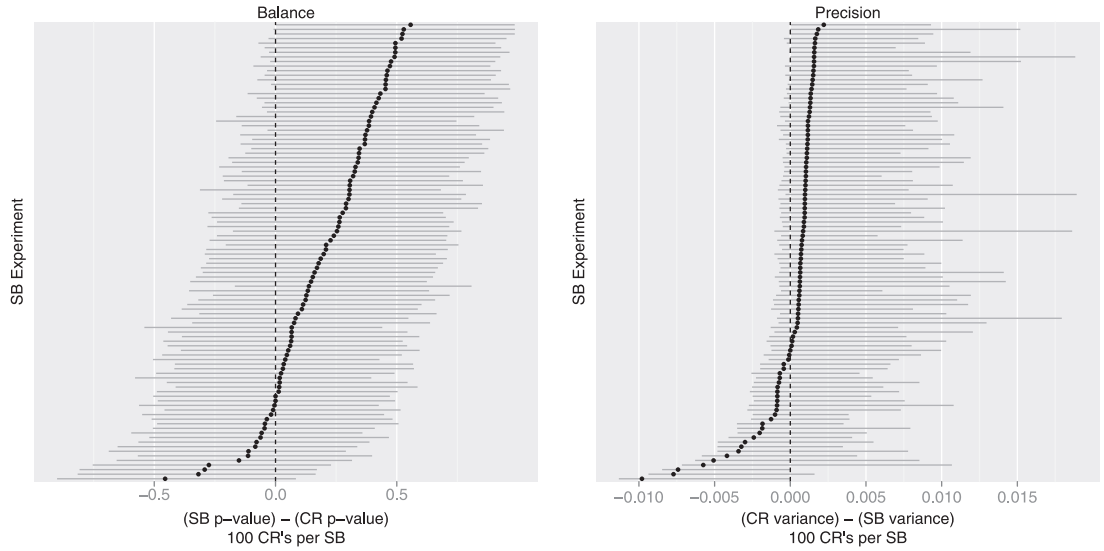

Fig. 1 Sequential blocking more balanced and precise than complete randomization. One hundred blocked experiments, each completely rerandomized one hundred times. Values to the right represent sequential blocking advantage over complete randomization. Segments show range of differences; points are median differences. Bivariate MVN data, $r = 0$; see Section 4.1.

about $x = 0$. The figure's asymmetry, with more differences to the right of zero, represents the advantage of sequential blocking.

To compare the precision of the treatment effect estimates, we employ the variance around a linear estimate of the treatment effect. This variance assumes an error variance σ^2 that scales our measure, but then conditions on each rerandomization's treatment allocation.[9] We find that sequential blocking tends to be more precise than complete randomization. The right panel of Fig. 1 displays the precision across the hundred sequentially blocked experiments and associated complete randomizations, with differences taken so that more precise sequential blockings are displayed to the panel's right. Across each sequentially blocked experiment (i.e., along each segment), the distribution of variances under complete randomization is right-skewed, indicating some very unlucky complete randomizations; each segment's median point is far to its left. The median sequential blocking produces more precision than 75% of its complete rerandomizations. Where either method is extremely imprecise, counts within treatment conditions appear to be most imbalanced.

4.2 *Correlated MVN Data*

We find similar advantages to sequential blocking using two MVN covariates that correlate at $r = 0.8$. Figure 2 organizes the results in two ways. The top row demonstrates improvements on average when comparing one blocked experiment to one hundred completely randomized ones, as in Section 4.1. The bottom row organizes the same data differently: the first segment compares one blocked randomization for all hundred sets of units to one complete randomization for all hundred sets of units; the second segment and subsequent ones represent different blocked and complete randomizations on the same hundred sets of units. Comparing completely randomized experiments

[9]We measure the inverse precision of the estimator by $\text{var}(\hat{\Delta}) = \sigma^2(a^T a - a^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T a)^{-1}$. The data from which $\text{var}(\hat{\Delta})$ are calculated are generated in accord with Atkinson's (2003) derivation. As an alternative, we produced a Monte Carlo estimate of the variance around the treatment effect by rerandomizing an $n = 40$ dataset 10,000 times. In the end, the scale of the differences between the sequentially blocked and completely randomized variances reflects those in Fig. 1, with sequential blocking exhibiting a smaller variance.

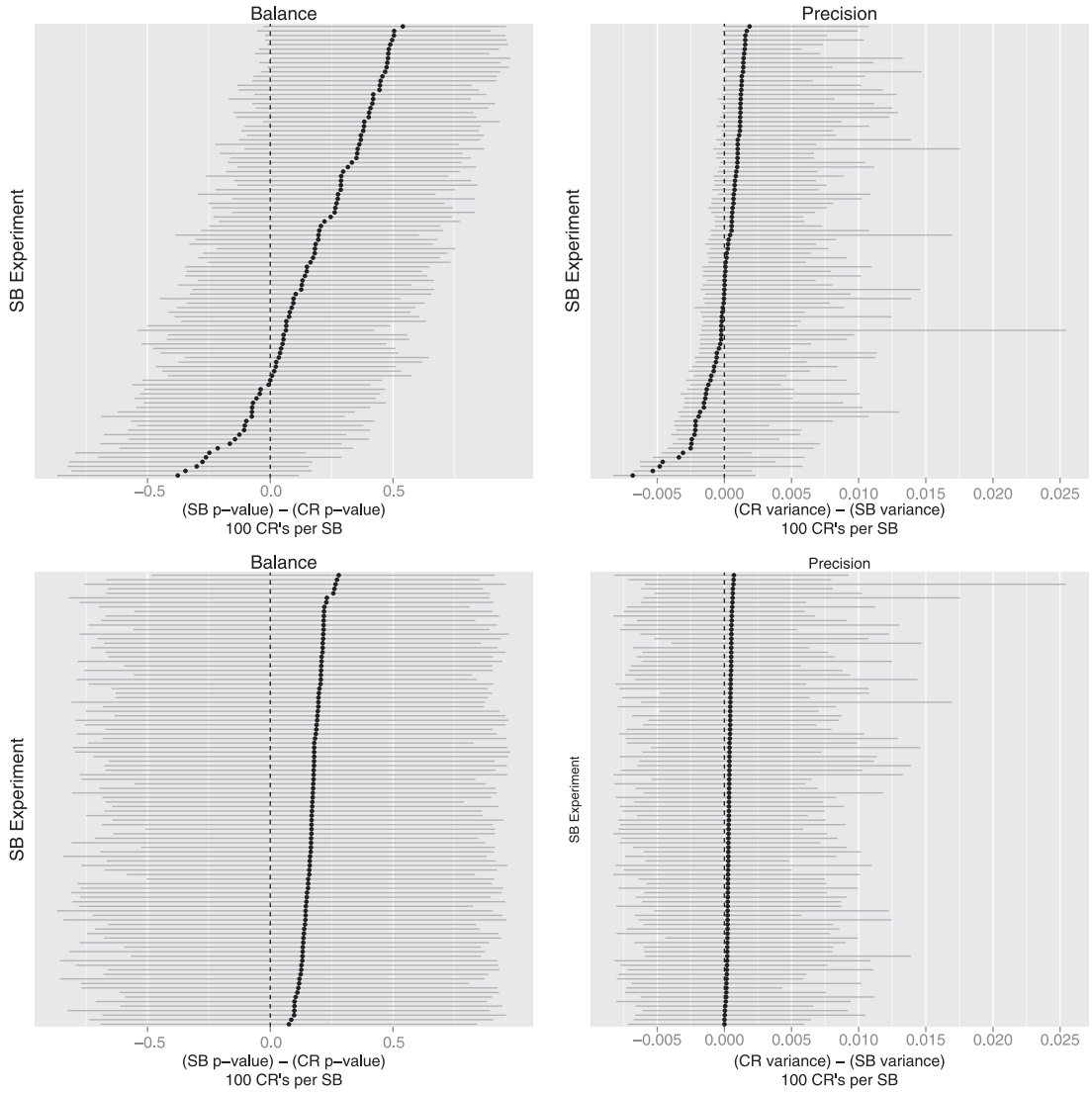

Fig. 2 SB experiments more balanced and precise than CR experiments, aggregated two ways. One hundred blocked experiments, each completely rerandomized one hundred times. Values to right represent sequential blocking advantage. Top segments show range of differences for one SB minus one hundred CRs; bottom segments show range of differences for one hundred SBs minus one hundred CRs; points are median differences. Bivariate MVN data, $r = 0.8$; see Section 4.2.

to other completely randomized experiments would produce distributions that were centered at zero, in contrast to those represented in Fig. 2's bottom row.

The bottom row also demonstrates that across many rerandomizations,[10] the median experiment is always more balanced and precise when sequentially blocked. We display this alternative presentation for Sections 4.1, 4.3, and 4.4 in the supplementary materials.

[10]By "rerandomization," we mean a set of treatment allocations, *all* of which we use to assess balance or precision. Morgan and Rubin's (2012) usage differs: they describe nonsequential experiments where several treatment allocations are proposed, allocations that are too imbalanced are thrown out, one acceptable allocation is implemented, and valid randomization inference considers only the set of acceptable allocations.

4.3 *MVN with Outliers*

One might worry that covariate-adaptive procedures are vulnerable to distortion by even a single outlier. Suppose most units in a trial have a covariate value of 1 or 2, but the first unit has covariate value one hundred and is assigned to Treatment A. Treatment A's covariate mean will be much greater than that of Treatment B, threatening the experiment if too many later units are assigned to Treatment A. Several safeguards can prevent this.

First, even after an outlier, stochastic methods still may place new units in one of the "right" treatment conditions. Second, for the `ktimes` methods, the distribution of MDs only affects a condition's ranking, not how much the assignment of the current unit gets biased.[11] Third, using the median or trimmed mean to compare the conditions invokes a higher breakdown point than the mean. An outlier's covariate value will cease to drive future assignments after a few units have been assigned to its treatment condition.

Since timing dictates how much and what information is available to allocate treatments sequentially, we investigate whether early-, middle-, or late-appearing outliers affect the overall balance of sequentially blocked experiments. We create outliers by taking ten times the maximum covariate values occurring in the sample so far. We simulate one hundred forty-unit experiments with two MVN covariates correlated at $r = 0.6$, and compare the treatment and control group distributions of a covariate using a two-sample Kolmogorov–Smirnov (KS) test. With complete randomization in the absence of outliers, we would expect about 10% of the KS p-values to fall below 0.1, and about 5% to fall below 0.05. However, the sequential blocking more than makes up for the introduction of an outlier that might be expected to generate small p-values by altering the covariate distribution of one condition. Despite the presence of an outlier, few of our simulated experiments produce low p-values. Table 3 shows that the fraction of p-values below the nominal value is lower than expected, whether the outlier occurs early in the trial (as the 2nd unit), in the middle (20th), or at the end (35th). The fraction of p-values representing covariate imbalance is at or below 2% for the 5% nominal test and 6% for the 10% nominal test.

Figure 3 displays balance and precision given early and late outliers. The relative improvements over complete randomization are similar to those in the correlated data, suggesting that sequential blocking is somewhat better at handling an extreme outlier. The relative improvement in precision appears greater when an outlier arrives late rather than early.

4.4 *Extreme Bimodal Data*

Sequential blocking retains its advantages over complete randomization under other extreme conditions. We draw two MVN covariates that correlate at $r = 0.8$ for a sample of 2000 units, then use only the most extreme 2%. Comparing by experiment, the median sequentially blocked randomization usually exhibits more balance and precision than complete randomization. In the median experiment, sequential blocking is more balanced than 69% of the complete randomizations, and more precise than 67%. Comparing by randomization, the median experiment is always more balanced and precise. Echoing the previous figures, Supplementary Fig. 3 shows these results.

5 Application-Specific Design and Balance Assessment

To select an allocation method for the sequential trial we conducted, we constructed realistic data reflecting covariate values we expected in the trial (Schafer and Kang 2008). We expected a small-sample experiment with variable times between individual assignments, making this trial appealing for methodological development. However, the strategies below apply to larger and faster political experiments as well, as we discuss in Section 6. We created one hundred datasets with two binary exact and seven more continuous covariates mirroring those in the trial.[12] Generally, the sequential

[11] This rank of MDs prevents large distances from driving assignments; MDs that themselves rely on ranks rather than levels achieve the same end (Rosenbaum 2010).

[12] Exact variables: gender and depression diagnosis. Continuous variables: post-traumatic stress disorder (PTSD) severity, age, depression severity, verbal fluency, executive control, and immediate and delayed logical memory.

Table 3 Sequential blocking outperforms complete randomization in the presence of extreme outliers

Outlier timing	*Position*	*Proportion* $p < .05$	*Proportion* $p < .1$
Early	2	.01	.05
Middle	20	.02	.06
Late	35	.00	.00

Note. For outliers appearing at three points in the trial, the last two columns give proportion of KS p-values below the nominal p-value. See Section 4.3.

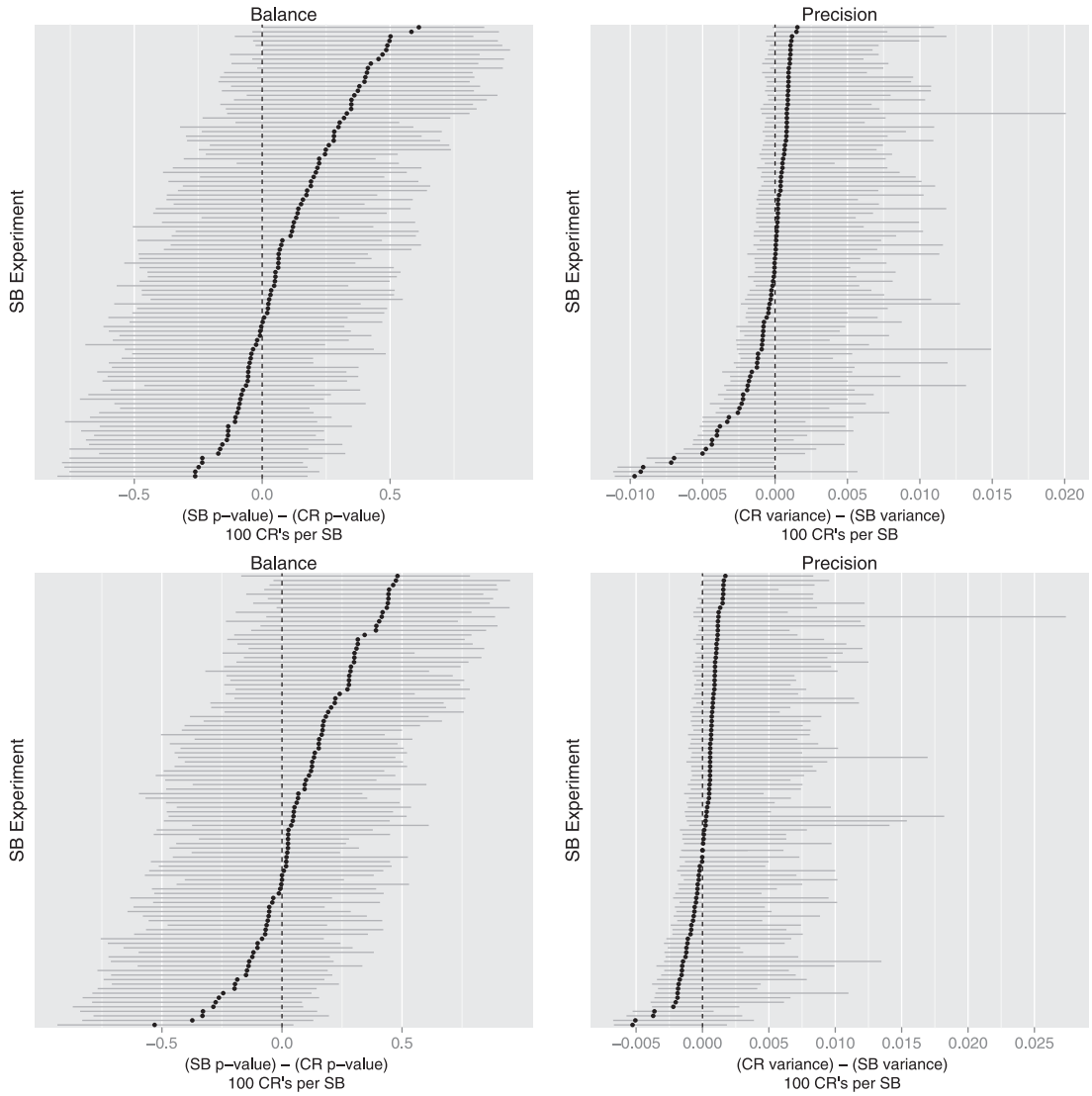

Fig. 3 Improvements in balance and precision in MVN-correlated data ($r = 0.6$) when an extreme outlier arrives early (top row) or late (bottom row) in experiment. See Section 4.3.

blocking methods perform better than complete randomization, though some methods systematically outperform others. For the `ktimes` algorithms, balance generally improves as the k increases. Two mappings that involve probability weights proportional to functions of the mean MDs performed comparably to complete randomization. Figure 4 displays the distribution of d^2 p-values from one hundred datasets sequentially blocked using seven different methods. The left

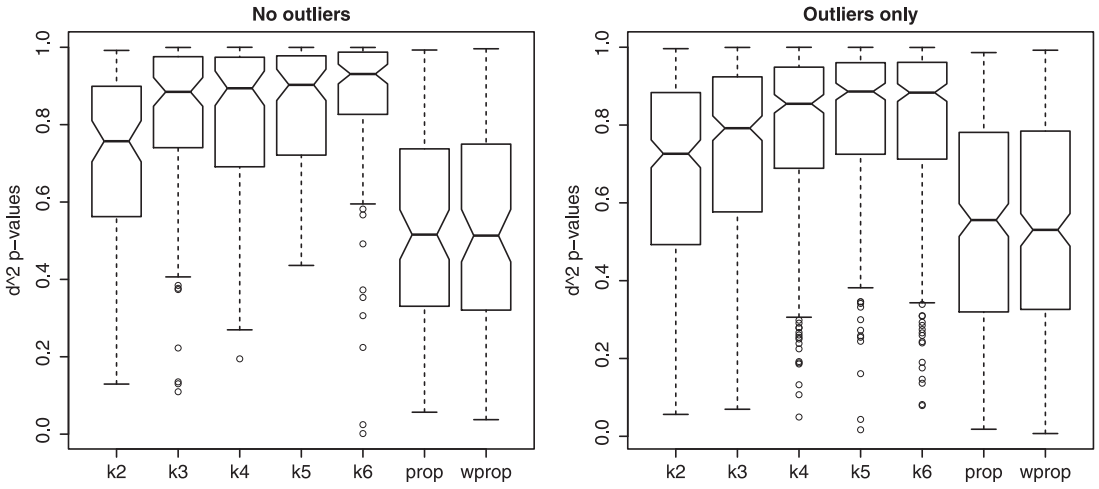

Fig. 4 Selecting a method for the PTSD trial. Boxplots display d^2 p-values from one hundred experiments using seven sequential blocking methods. With no (left panel) or some severe outliers (right panel), method `k5` produces narrow spreads and high minimum p-values.

panel compares the methods in the absence of covariate outliers; the right panel's trials inject outliers at an early, middle, or late stage, some of which are outside the range of plausible empirical data values. Based on these and other balance histograms, densities, and scatterplots, we selected for the actual trial the `ktimes` method overweighting the most-different treatment condition with $k = 5$.

This sequential blocking method was then used to assign one of two experimental conditions to veterans with PTSD in a study conducted at the Veterans Affairs (VA) Puget Sound Health Care System—Seattle Division (Moore, Moore, and Simpson 2013). Starting in November 2009, male and female veterans were recruited through posted flyers and referrals from VA clinicians.[13] Fifty-one eligible subjects were assigned experimental conditions between November 2009 and July 2011.[14] Subjects were randomly assigned to daily practice retrieving specific memories of life events in response to cue words (*memory* condition) or a control task in which they were asked to rearrange the letters of cue words to create new words (*anagrams* condition). Outcomes of interest include measures of PTSD severity and depression two to four weeks after assignment to an experimental condition.

As the experiment unfolds, we can calculate the overall balance after each subject's assignment. Because some units inform a few subsequent assignments, but are eventually lost to follow-up, we examine the balance dynamically with Fig. 5. For example, although subject 19 was in the study during assignment of subjects 20–24, (s)he withdrew before the assignment of subject 25. Thus, Fig. 5 includes subject 19 in the balance calculations of the first gray rectangle, but omits this subject thereafter.

Since analysis will center on the subjects remaining in the study until outcome data are collected, we now focus on this group. Figure 6 displays the covariate data collected for the forty-six subjects with valid follow-up data. In the left panel, the covariate values by treatment condition appear as mosaic plots for the binary exact blocking variables, and quantile–quantile (QQ) plots for the seven continuous measures. Balance assessment was performed without access to the outcome data

[13] An eligibility screening process ensured that subjects met the study criteria. Inclusion criteria included veteran status, a PTSD diagnosis, and age 18–64. The trial focused on those with primary PTSD diagnoses; exclusion criteria included evidence of psychotic disorder, bipolar disorder, substance abuse in the past month or dependence in the past year, recent psychiatric hospitalization, or current suicidal intent.

[14] The study examines the effects of practicing a memory task on PTSD symptoms and related difficulties. Individuals with PTSD exhibit impairment in the retrieval of specific autobiographical memories relative to individuals without PTSD, and this study examines whether practicing the retrieval of specific autobiographical memories leads to short-term changes in symptoms associated with PTSD.

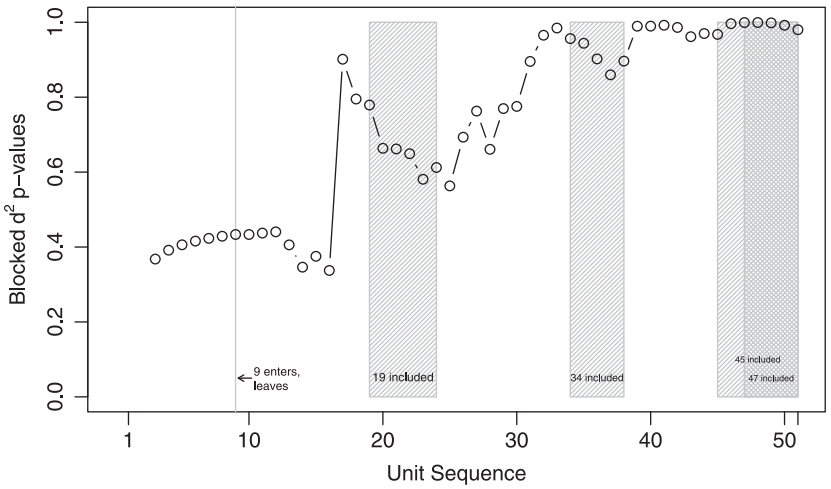

Fig. 5 Dynamic balance during the PTSD experiment. For subjects without full follow-up, gray indicates the period during which covariate information is used to allocate other subjects.

(Rubin 2001; Love et al. 2008). Generally, the balance looks reasonable given the moderate sample size and several covariates considered.

Figure 6's right panel more formally compares the covariates in the two experimental conditions. For each variable, we display p-values from KS, t, and Wilcoxon rank-sum tests, none of which is significant at a 0.05 or 0.1 level. The d^2 p-value suggests no difference in the overall multivariate distributions of the covariates in this experiment. Indeed, as we would expect a uniform distribution of the p-values for both the individual and omnibus statistics under complete randomization, these sequentially blocked results compare quite favorably. Any experimental design has a limited ability to create balance, conditional on the finite sample and the values observed. In this context, we assess the balance created by our chosen design to be an asset to valid inference. We balance the continuous covariates, which biased coin or minimization methods would ignore or severely coarsen. The logical memory covariates take fourteen and fifteen different values, and the other five continuous covariates take on more than twenty. The large number of possible profiles, 8,230,118,400, renders unhelpful methods that rely on unique profiles.

We also compare the d^2 p-value at the top of Fig. 6's right panel to 10,000 complete rerandomizations of the participants. As expected, the sequentially blocked experiment outperforms 99% of the completely randomized experiments (Supplementary Fig. 4).

6 Balancing Larger Political Experiments

We recommend that analysts create realistic data to pre-test several assignment algorithms, as in Section 5. However, we now demonstrate that algorithms not so tailored can still improve balance across important pre-treatment variables, even in larger political experiments.

Horiuchi, Imai, and Taniguchi (2007) conduct a survey experiment during the 2004 Japanese Upper House election; they exact block on sex and turnout intention, and then provide encouragement for policy information seeking.[15] Beyond their blocking variables, we also consider age, education, and previous district-level turnout. Again, the large number of possible unique profiles in these covariates renders unhelpful designs for discrete variables.[16]

We assume that the order of the observations in the replication datafile represents their entry order into a sequential experiment. We then sequentially block the 1397 respondents one hundred

[15] Because our interest is in experimental design, not in replicating prior results or addressing other concerns, we use only the 1397 encouragement-compliers for whom outcome data were observed.

[16] Coarsening age into five-year groups and past turnout into quartiles creates $6 \times 8 \times 2 \times 4 = 384$ profiles; using unique ages and turnout levels creates 42,720 profiles.

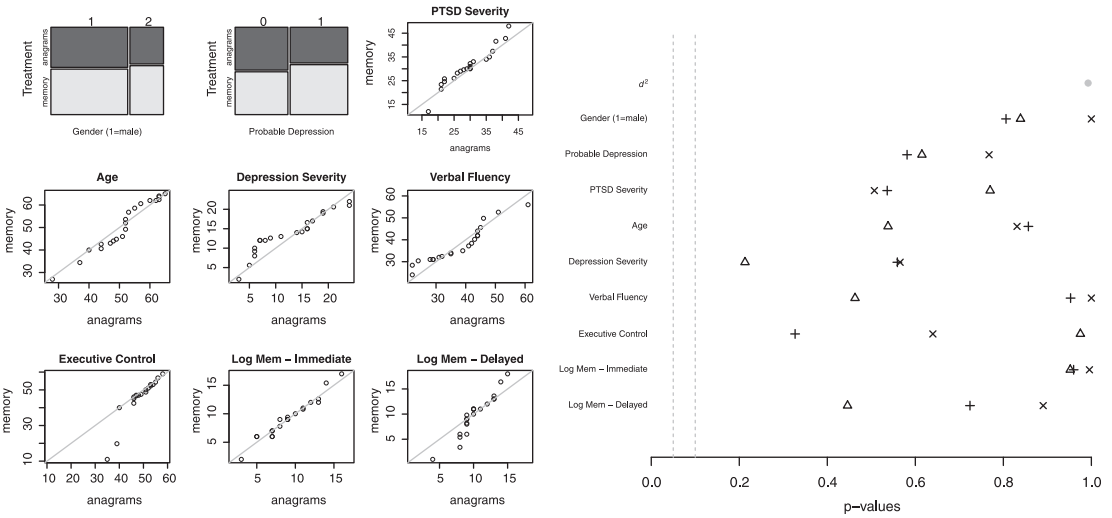

Fig. 6 Balance at conclusion of the PTSD trial. *Left*: Mosaic and QQ plots. *Right*: Bootstrapped KS, t, and Wilcoxon rank-sum test p-values ($\Delta, +$, and $\times$, respectively). Gray disk shows d^2 p-value for overall balance. Dotted lines at $p = .05$, $p = .10$.

times. Figure 7's left panel represents the distribution of the d^2 p-values and compares them to the balance in the actual experiment. The original experiment is well balanced ($p \approx 0.89$), but further variables could be included in the design to improve the balance. With many experimental units, sequential blocking makes the overall balance nearly perfect.

Among several contributions, Cobb, Greiner, and Quinn (2011) train poll workers at randomly selected sites on the appropriate procedure for requesting identification from voters. As it was infeasible to have two sets of poll workers at each polling place and randomly assign individual voters to a trained or untrained set of poll workers, Cobb, Greiner, and Quinn (2011) pair poll locations and implement the training in one site and not the other. We consider the 390 individuals from the blocked pair with the largest number of total respondents with fully observed data on race, language, education, gender, and age. The distributions of individual characteristics in this particular pair differ somewhat in age and education levels; Supplementary Fig. 5 displays these distributions.

Using the time stamps in the original data, we sequentially block these individuals in the order in which they voted, assuming that individuals in either location could have been assigned to treatment or control. Figure 7's right panel summarizes the very good balance across one hundred sequential assignments. While we would expect a uniform distribution of p-values on [0, 1] under complete randomization, blocking confines the distribution to (0.93,1).

7 Comparing Causal Estimates and Confidence Intervals

We consider five outcomes in the forty-six PTSD subjects who completed at least 50% of the practice assigned and on whom follow-up measures were made within four weeks of initial assignment.[17] We examine the difference in means between experimental groups, as well as estimates from AIPW estimators, which model the outcome and use an estimate of the probability of each treatment condition (Tsiatis 2006).

Estimating the propensity score poses a challenge. Though we conditioned on each subject's covariates to assign her to conditions with probability $\frac{5}{6}$ and $\frac{1}{6}$, her propensity score should be marginal over the assignments of the previous subjects (Rosenbaum and Rubin 1983). In a non-adaptive design, the assignments of previous participants would not need to be considered. To

[17]These outcomes are an interview measure of PTSD severity, self-reported PTSD severity, probable diagnosis of major depression (binary), self-reported depression severity, and a measure of ruminative thinking.

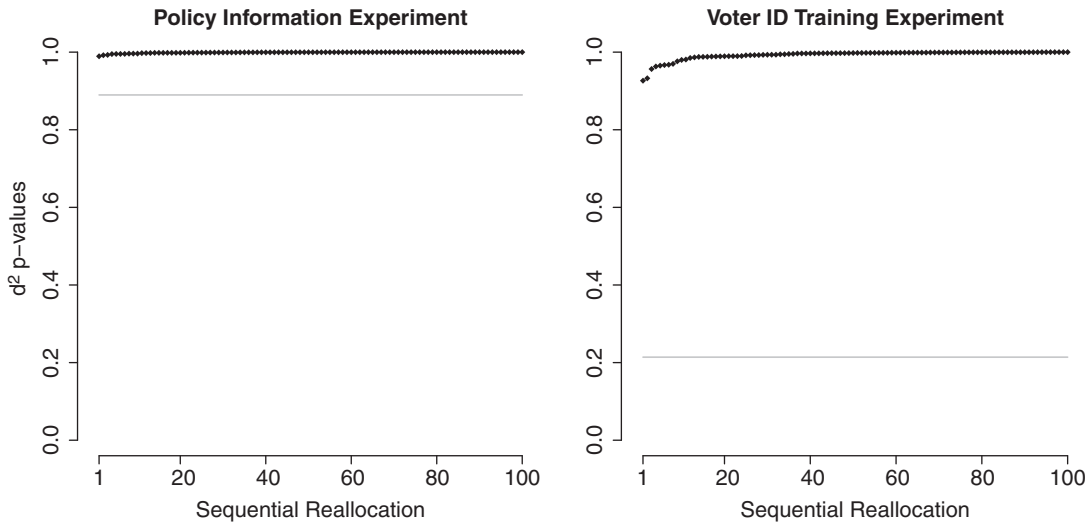

Fig. 7 Balance in two larger political experiments (gray lines), versus one hundred sequential blockings. Horiuchi, Imai, and Taniguchi (2007) exact-block on sex and voting intention; we add age, college, and turnout (left). We assume individual treatment, unlike the original Cobb, Greiner, and Quinn (2011) design (right).

estimate the propensity scores, we take a random sample from the roughly 10^{12} possible assignments of our participants to two treatment conditions by simulating 10,000 sequential assignments, given our participants' actual covariate values. We estimate each participant's probability of being in the treatment group as that participant's mean assignment probability. These estimates are within 2% of 0.5 for all of our subjects.

We use the outcomes observed for our participants under their actual treatment conditions, but simulate 500 sequential assignments and calculate the two estimates of the treatment effect. Thus, in each rerandomization, some outcomes observed under the treatment condition are considered outcomes under control. The estimated propensity scores help us know what to expect: since participants have a roughly equal probability of being assigned treatment and control across trials, our simulated and observed assignments should only correlate by chance, and we expect the effect estimate to be zero. Consistent with our sharp null hypothesis, the effect estimates all center around zero and deviate from it by less than one-fourth of a standard deviation. Figure 8 displays the densities of the estimates for two outcomes, with AIPW estimates summarized by the darker, more precise curve.

To further compare the efficiency of the blocked design to completely randomized designs, we calculate the root mean squared error (RMSE, $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$) for each outcome. Across several rerandomizations, we compare the distribution of the RMSE for each design; we find similar distributions of the RMSE for the blocked and unblocked designs. Figure 9's left panel displays these distributions, with an outcome model that includes baseline levels of PTSD symptoms, age, depression severity, verbal fluency, and executive control.[18]

Finally, applied research usually centers around an estimate from a single experiment. To estimate the uncertainty around such an estimate, we generate distribution-free, nonparametric confidence intervals that utilize the details of our randomization procedure. To do so we invert the randomization inference test, which involves positing a range of constant treatment effect hypotheses, among them the sharp null (Rosenbaum 2002; Ho and Imai 2006). As full results of

[18]To facilitate visual comparison, we multiply the binary outcomes' RMSEs by a constant. Other models yield similar results, including baseline logical memory, depression, gender, or smoothed PTSD symptoms.

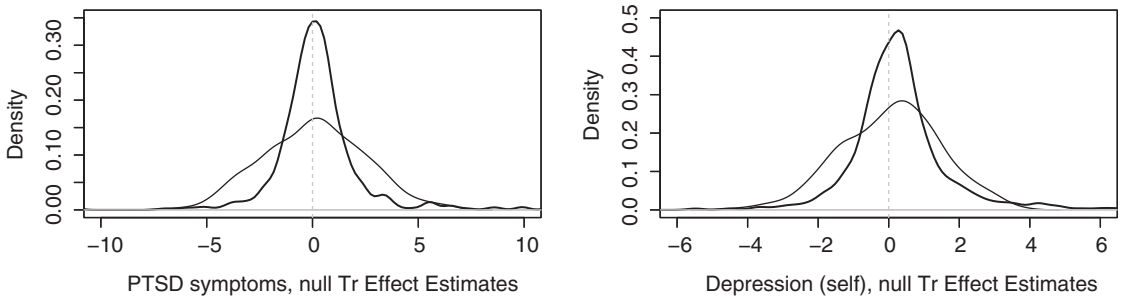

Fig. 8 Estimates of the effect of memory task on PTSD patients, 500 sequential assignment simulations assuming sharp null treatment effect. AIPW estimates are thicker curves.

the trial will appear elsewhere (Moore, Moore, and Simpson 2013), we invert the test using slightly adjusted outcomes, expecting the resulting intervals to center near zero.

We focus on the interview measure of PTSD severity and find that the intervals formed by inverting the randomization test are between 9% and 15% shorter than the confidence intervals represented by the parametric *t*-test for the difference in means. We find 80%, 90%, and 95% confidence intervals that cover the approximate intervals $(-3.0, 3.1)$, $(-4.2, 4.1)$, and $(-4.9, 5.2)$, respectively, shown in the right panel of Fig. 9. Each interval is formed by testing one hundred hypothetical treatment effects using 1000 sequential rerandomizations to generate an acceptance probability for each hypothesized effect. The acceptance regions do not abut the extremes of the tested hypotheses, suggesting that wider testing is not necessary.

8 Discussion

Sequential blocking allows applied researchers to prevent bad luck from sabotaging their trials while maximizing what can be learned from finite samples. Sequential blocking improves covariate balance and the precision of causal estimates in trickle-in experiments across a variety of simulated and real data. We build upon canonical methods for few-category discrete covariates, extending them to incorporate many continuous covariates.

Our three empirical applications contextualize some customary principles of experimental design. First, better balance is easier to obtain with more units. With forty hypothetical units in the PTSD experiment, the *p*-values for our chosen method ranged from about 0.4 to 1; with about 400 units in the voter ID subsample, they range from 0.93 to 1; with about 1400 units in the policy information experiment, they range from 0.99 to 1. Second, individual-level randomizations should produce better individual-level balance than group randomizations, as we see in the voter ID application. Third, a design that focuses on a rich set of background covariates can improve experiments in a variety of contexts.

We think pilot studies can play important roles in sequential randomizations. In particular, the covariate means and covariance matrix from a pilot could be used to approximate the center and spread of the full trial for units arriving early. That is, pilots can provide an estimate of the Σ^{-1} matrix that scales the covariate differences in the MD before the first unit of the full study arrives. By anticipating aspects of the full trial's covariate distribution, researchers allow allocations in their sequential experiments to more closely approximate the advantageous conditions of nonsequential political experiments.

Whether of modest size (such as the trial we conducted), larger scale (such as our political information application), or something in between, randomized social scientific and clinical trials stand to gain from the inclusion of background information (Duflo, Glennerster, and Kremer 2008; Bowers 2011). By incorporating both discrete and continuous covariate data into sequential designs and supporting these designs with appropriate analysis, researchers can strengthen experimental causal inferences.

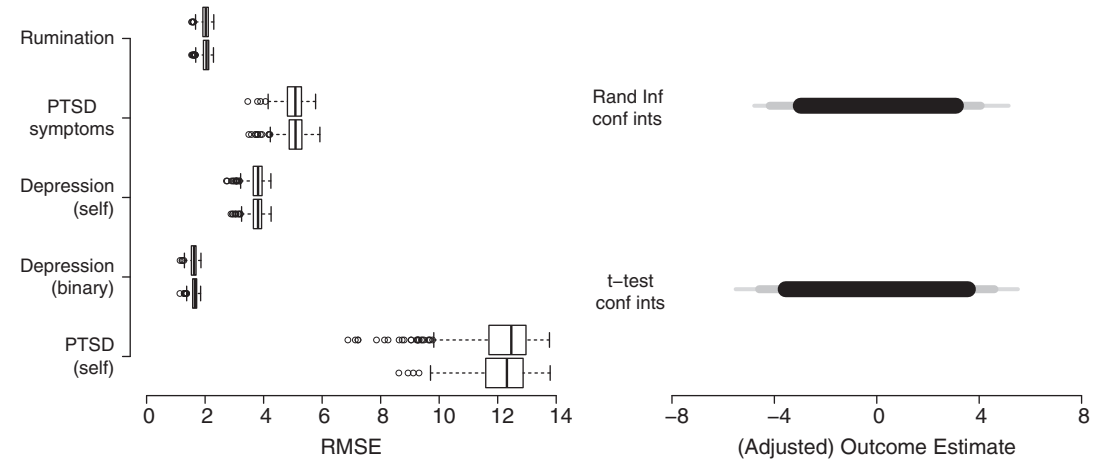

Fig. 9 Causal estimate summaries. *Left*: RMSEs for five outcomes, thirty sequentially blocked (top) and completely randomized designs. *Right*: Nonparametric randomization inference confidence intervals are 9%–15% shorter than *t*-test intervals (80%, 90%, and 95% intervals).

Funding

Pilot Grant (UL1RR025014) from the Institute for Translational Health Sciences (ITHS), which is funded by the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research; Office of Academic Affiliations, Advanced Fellowship Program in Mental Illness Research and Treatment, Department of Veterans Affairs; Robert Wood Johnson Foundation.

References

Atkinson, Anthony C. 2003. The distribution of loss in two-treatment biased-coin designs. *Biometrics* 4:179–93.

Ball, Frank G., Adrian F. M. Smith, and Isabella Verdinelli. 1993. Biased coin designs with a Bayesian bias. *Journal of Statistical Planning and Inference* 34:403–21.

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. New York, NY: Cambridge University Press.

Braucht, G. Nicholas, and Charles S. Reichardt. 1993. A computerized approach to trickle-process, random assignment. *Evaluation Review* 17:79–90.

Chong, Dennis, and James N. Druckman. 2007. Framing public opinion in competitive democracies. *American Political Science Review* 101:637–55.

Chow, Shein-Chung, and Mark Chang. 2007. *Adaptive design methods in clinical trials*. Boca Raton, FL: Chapman & Hall.

Cobb, Rachael V., D. James Greiner, and Kevin M. Quinn. 2011. Can voter ID laws be administered in a race-neutral manner? Evidence from the city of Boston in 2008. *Quarterly Journal of Political Science* 6:1–33.

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.

Efron, Bradley. 1971. Forcing a sequential experiment to be balanced. *Biometrika* 58:403–17.

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

Harrington, David P. 2000. The randomized clinical trial. *Journal of the American Statistical Association* 95:312–15.

Ho, Daniel E., and Kosuke Imai. 2006. Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *Journal of the American Statistical Association* 101:888–900.

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.

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.

Kalish, Leslie A., and Colin B. Begg. 1985. Treatment allocation methods in clinical trials: A review. *Statistics in Medicine* 4:129–44.

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.

Lachin, John M., John P. Matts, and L. J. Wei. 1988. Randomization in clinical trials: Conclusions and recommendations. *Controlled Clinical Trials* 9:365–74.

Love, Thomas E., Randall D. Cebul, Douglas Einstadter, Anil K. Jain, Holly Miller, C. Martin Harris, Peter J. Greco, Scott S. Husak, and Neal V. Dawson. 2008. Electronic medical record-assisted design of a cluster-randomized trial to improve diabetes care and outcomes. *Journal of General Internal Medicine* 23:383–91.

Malhotra, Neil, and Alexander G. Kuo. 2008. Attributing blame: The public's response to hurricane Katrina. *Journal of Politics* 70:120–35.

Moore, Ryan T. 2012. Multivariate continuous blocking to improve political science experiments. *Political Analysis* 20:460–79.

Moore, Ryan T., and Sally A. Moore. 2013. *Replication data for: Blocking for sequential political experiments*. http://hdl.handle.net/1902.1/21042. IQSS Dataverse Network, V1.

Moore, Sally A., Ryan T. Moore, and Tracy L. Simpson. In preparation 2013. The effects of practicing specific autobiographical memory retrieval in veterans with PTSD.

Morgan, Kari Lock, and Donald B. Rubin. 2012. Rerandomization to improve covariate balance in experiments. *Annals of Statistics* 40:1263–82.

Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental political science and the study of causality: From nature to the lab*. New York: Cambridge University Press.

Pocock, Stuart J., and Richard Simon. 1975. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 31:103–15.

R Core Team. 2013. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosenbaum, Paul R. 2002. *Observational studies*. 2nd ed. New York: Springer.

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

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

Rosenberger, William F., and John M. Lachin. 2002. *Randomization in clinical trials: Theory and practice*. Hoboken, NJ: Wiley.

Rosenberger, William F., and Oleksandr Sverdlov. 2008. Handling covariates in the design of clinical trials. *Statistical Science* 23:404–19.

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

Schafer, Joseph L., and Joseph Kang. 2008. Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods* 13:279–313.

Schulz, Kenneth F., and David A. Grimes. 2002a. Generation of allocation sequences in randomized trials: Chance, not choice. *The Lancet* 359:515–19.

———. 2002b. Unequal group sizes in randomised trials: Guarding against guessing. *The Lancet* 359:966–70.

Tsiatis, Anastasios A. 2006. *Semiparametric theory and missing data*. New York: Springer.

Whitehead, John. 1997. *The design and analysis of sequential clinical trials*. New York: Wiley.

Zelen, Marvin. 1969. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* 64:131–46.

———. 1974. The randomization and stratification of patients to clinical trials. *Journal of Chronic Disease* 27:365–75.