Advance Access publication June 6, 2016

Political Analysis (2016) 24:324–338
doi:10.1093/pan/mpw012

Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead

Maximilian Kasy
Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge, MA 02138, USA
e-mail: maximiliankasy@fas.harvard.edu (corresponding author)

Edited by R. Michael Alvarez

Suppose that an experimenter has collected a sample as well as baseline information about the units in the sample. How should she allocate treatments to the units in this sample? We argue that the answer does not involve randomization if we think of experimental design as a statistical decision problem. If, for instance, the experimenter is interested in estimating the average treatment effect and evaluates an estimate in terms of the squared error, then she should minimize the expected mean squared error (MSE) through choice of a treatment assignment. We provide explicit expressions for the expected MSE that lead to easily implementable procedures for experimental design.

1 Introduction

Experiments, and in particular randomized experiments, are the conceptual reference point that gives empirical content to the notion of causality. In recent years, actual randomized experiments have become increasingly popular elements of the methodological toolbox in a wide range of social science disciplines. Examples from the recent political science literature abound. Blattman, Hartman, and Blair (2014), for instance, provided training in "alternative dispute resolution practices" to residents of a random set of towns in Liberia. Kalla and Broockman (2015) analyze whether the possibility of scheduling a meeting with a congressional office changes depending on whether it is revealed that the person seeking a meeting is a political donor. Findley, Nielson, and Sharman (2015) study the effect of varying messages to incorporation services in different locations on the possibility of establishing (illegal) anonymous shell corporations.

Researchers conducting such field experiments in political science (as well as in fields such as economics, medicine, and other social and biomedical sciences) are often confronted with variants of the following situation (cf. Morgan and Rubin 2012). They have selected a random sample from some population and have conducted a baseline survey for the individuals in this sample. Then a discrete treatment is assigned to the individuals in this sample, usually based on some randomization scheme. Finally, outcomes are realized, and the data are used to perform inference on some average treatment effect.

A key question for experimenters is how to use covariates from the baseline survey in the assignment of treatments. Intuition and the literature suggest to use stratified randomization conditional on covariates, also known as blocking. Moore (2012), for instance, makes a convincing argument that blocking on continuous as well as discrete covariates is better than full randomization or blocking only on a small number of discrete covariates. We analyze this question—how to

Author's note: I thank Alberto Abadie, Ivan Canay, Gary Chamberlain, Raj Chetty, Nathaniel Hendren, Guido Imbens, Larry Katz, Gary King, Michael Kremer, and Don Rubin, as well as seminar participants at the Harvard development retreat; the Harvard Labor Economics Workshop; the Harvard Quantitative Issues in Cancer Research Seminar; the Harvard Applied Statistics Seminar; UT Austin, Princeton, Columbia, and Northwestern Econometrics Seminars; at RAND; and at the 2013 CEME Conference at Stanford for helpful discussions. Replication data are available on the Harvard Dataverse at http://dx.doi.org/10.7910/DVN/I5KCWI. See Kasy (2016). Supplementary materials for this article are available on the *Political Analysis* Web site.

© The Author 2016. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

use baseline covariates—as a decision problem. The experimenter has to choose a treatment assignment and an estimator, given knowledge of the covariate distribution in the sample. Her objective is to minimize risk based on a loss function such as the mean squared error (MSE) of a point estimator. The decision criteria considered are Bayesian average risk and (conditional) minimax risk.

We show, first, that experimenters might not want to randomize in general. While surprising at first, the basic intuition for this result is simple and holds for any statistical decision problem. The conditional expected loss of an estimator is a function of the covariates and the treatment assignment. The treatment assignment that minimizes conditional expected loss is in general unique if there are continuous covariates, so that a deterministic assignment strictly dominates all randomized assignments.[1]

We next discuss how to implement optimal and near-optimal designs in practice, where near-optimal designs might involve some randomization. The key problem is to derive tractable expressions for the expected MSE of estimators for the average treatment effect, given a treatment assignment. Once we have such expressions, we can numerically search for the best assignment, or for a set of assignments that are close to optimal. In order to calculate the expected MSE, we need to specify a prior distribution for the conditional expectation of potential outcomes given covariates. We provide simple formulas for the expected MSE for a general class of nonparametric priors.

Our recommendation not to randomize raises the question of identification, cf. the review in Keele (2015). We show that conditional independence of treatment and potential outcomes given covariates still holds for the deterministic assignments considered, under the usual assumption of independent sampling. Conditional independence only requires a controlled trial (CT), not a randomized controlled trial (RCT).

To gain some intuition for our non-randomization result, note that in the absence of covariates the purpose of randomization is to pick treatment and control groups that are similar before they are exposed to different treatments. Formally, we would like to pick groups that have the same (sample) distribution of potential outcomes. Even with covariates observed prior to treatment assignment, it is not possible to make these groups identical in terms of potential outcomes. We can, however, make them as similar as possible in terms of covariates. Allowing for randomness in the treatment assignment to generate imbalanced distributions of covariates can only hurt the balance of the distribution of potential outcomes. Whatever the conditional distribution of unobservables given observables is, having differences in observables implies greater differences in the distribution of unobserables relative to an assignment with no differences in observables. The analogy to estimation might also be useful to understand our non-randomization result. Adding random (mean 0) noise to an estimator does not introduce any bias. But it is never going to reduce the MSE of the estimator.

The purpose of discussing tractable nonparametric priors—and one of the main contributions of this article—is to operationalize the notion of "balance." In general, it will not be possible to obtain exactly identical distributions of covariates in the treatment and control groups. When picking an assignment, we have to trade off balance across various dimensions of the joint distribution of covariates. Picking a prior distribution for the conditional expectation of potential outcomes, as well as a loss function, allows one to calculate an objective function (Bayesian risk) that performs this trade-off in a coherent and principled way.

2 A Motivating Example and Some Intuitions

2.1 *Setup*

Before we present our general results and our proposed procedure, let us discuss a simple motivating example. The example is stylized to allow calculations "by hand," but the intuitions

[1] If experimenters have a preference for randomization for reasons outside the decision problem considered in the present article, a reasonable variant of the procedure suggested here would be to randomize among a set of assignments that are "near-minimizers" of risk. If we are worried about manipulation of covariates, in particular, a final coin flip that possibly switches treatment and control groups might be helpful. I thank Michael Kremer for this suggestion.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

Table 1 Comparing bias, variance, and MSE of different treatment assignments

Assignment				*Designs*					*Model 1*			*Model 2*			*EMSE*
d_1	d_2	d_3	d_4	*1*	*2*	*3*	*4*	*5*	*bias*	*var*	*MSE*	*bias*	*var*	*MSE*	
0	1	1	0	1	1	1	1	1	0.0	1.0	1.0	2.0	1.0	5.0	1.77
1	0	0	1	1	1	1	1	1	0.0	1.0	1.0	−2.0	1.0	5.0	1.77
1	0	1	0	1	1	1	1	0	−1.0	1.0	2.0	3.0	1.0	10.0	2.05
0	1	0	1	1	1	1	1	0	1.0	1.0	2.0	−3.0	1.0	10.0	2.05
1	1	0	0	1	1	1	0	0	−2.0	1.0	5.0	6.0	1.0	37.0	6.51
0	0	1	1	1	1	1	0	0	2.0	1.0	5.0	−6.0	1.0	37.0	6.51
0	1	0	0	1	1	0	0	0	−0.7	1.3	1.8	3.3	1.3	12.4	2.49
0	0	1	0	1	1	0	0	0	0.7	1.3	1.8	−0.7	1.3	1.8	2.49
1	1	0	1	1	1	0	0	0	−0.7	1.3	1.8	0.7	1.3	1.8	2.49
1	0	1	1	1	1	0	0	0	0.7	1.3	1.8	−3.3	1.3	12.4	2.49
1	0	0	0	1	1	0	0	0	−2.0	1.3	5.3	4.7	1.3	23.1	6.77
1	1	1	0	1	1	0	0	0	−2.0	1.3	5.3	7.3	1.3	55.1	6.77
0	0	0	1	1	1	0	0	0	2.0	1.3	5.3	−7.3	1.3	55.1	6.77
0	1	1	1	1	1	0	0	0	2.0	1.3	5.3	−4.7	1.3	23.1	6.77
0	0	0	0	1	0	0	0	0	−	−	∞	−	−	∞	−
1	1	1	1	1	0	0	0	0	−	−	∞	−	−	∞	−
MSE model 1:				∞	3.2	2.7	1.5	1.0							
MSE model 2:				∞	20.6	17.3	7.5	5.0							

Notes: Each row of this table corresponds to one possible treatment assignment $(d_1, \ldots, d_4)$. The columns for "model 1" correspond to the model $Y_i^d = x_i + d + \epsilon_i^d$, and the columns for "model 2" to the model $Y_i^d = -x_i^2 + d + \epsilon_i^d$. Each row shows the bias, variance, and MSE of $\hat{\beta}$, for the given assignment and model. The designs 1–5 correspond to uniform random draws from the assignments with rows marked by an entry of 1. Design 1 randomizes over all rows, design 2 over rows one through fourteen, etc. The last column of the table shows the Bayesian expected MSE for each assignment for the squared exponential prior discussed below. For details, see Section 2.

from this example generalize. Suppose an experimenter has a sample of four experimental units $i = 1, \ldots, 4$, and she observes a continuous covariate X_i for each of them, where it so happens that $(X_1, \ldots, X_4) = (x_1, \ldots, x_4) = (0, 1, 2, 3)$. She assigns every unit to one of two binary treatments, $d_i \in \{0, 1\}$.[2]

Our experimenter wants to estimate the (conditional) average treatment effect of treatment D across these four units,

$$\beta = \frac{1}{n} \sum_i E[Y_i^1 - Y_i^0 | X_i]. \tag{1}$$

The experimenter plans to estimate this treatment effect by calculating the difference in means across treatment and comparison groups in her experiment, that is,

$$\hat{\beta} = \frac{1}{n_1} \sum_i D_i Y_i - \frac{1}{n_0} \sum_i (1 - D_i) Y_i. \tag{2}$$

Since there are four experimental units, there are $2^4 = 16$ possible treatment assignments. The sixteen rows of Table 1 correspond to these assignments.[3] In the first row, we have $(d_1, \ldots, d_4) = (0, 1, 1, 0)$, in the second row $(d_1, \ldots, d_4) = (1, 0, 0, 1)$, etc.

[2]Consider for instance the setting of Nyhan and Reifler (2014), where legislators i received letters about fact-checking ($D_i = 1$) or did not ($D_i = 0$). The outcome of interest Y in this case is the future fact-checking rating of these legislators; an important covariate X_i might be their past rating.

[3]The code producing Table 1 is available online at Kasy (2016). At this address, we also provide Matlab code for implementing our proposed approach in practice.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

Assume now that potential outcomes are determined by the equation

$$Y_i^d = x_i + d + \epsilon_i^d, \tag{3}$$

where the ϵ_i^d are independent given X, and have mean 0 and variance 1. This is labeled as "model 1" in Table 1. The average treatment effect in this model is equal to 1. For any treatment assignment $(d_1, \ldots, d_n)$, we can calculate the corresponding bias, variance, and MSE of $\hat{\beta}$ as an estimator of β, where

$$\text{Var}(\hat{\beta}) = \frac{1}{n_1} + \frac{1}{n_0} \tag{4}$$

with n_1 being the number of units i receiving treatment $d_i = 1$, and n_0 being the same for $d_i = 0$. The bias[4] is given by

$$\text{Bias} = E[\hat{\beta}] - \beta = \frac{1}{n_1}\sum_i d_i x_i - \frac{1}{n_0}\sum_i (1 - d_i)x_i \tag{5}$$

for our model.

For the assignment in the first row of Table 1, we get $\text{Var}(\hat{\beta}) = \frac{1}{2} + \frac{1}{2} = 1$, for the assignment in row 7, $\text{Var}(\hat{\beta}) = \frac{1}{1} + \frac{1}{3} = 1.33$, and similarly for the other rows. The bias for the first row is given by $\text{Bias} = E[\hat{\beta}] - \beta = \frac{1}{2} \cdot (x_2 + x_3 - x_1 - x_4) = 0$, for the third row by $\text{Bias} = \frac{1}{2} \cdot (x_1 + x_3 - x_2 - x_4) = -1$, etc. The MSE for each row is given by

$$\text{MSE}(d_1, \ldots, d_n) = E[(\hat{\beta} - \beta)^2] = \text{Bias}^2 + \text{Var}. \tag{6}$$

2.1.1 Forgetting the covariates

Suppose now for a moment that the information about covariates got lost—somebody deleted the column of X_i in your spreadsheet. Then every i looks the same, before treatment is assigned. The variance of potential outcomes Y^d now includes both the part due to ϵ_i^d and the part due to X_i, and is equal to $\text{Var}(Y_i^d) = \text{Var}(X_i) + \text{Var}(\epsilon_i^d) = \text{Var}(X_i) + 1$. Since units i are indistinguishable in this case, treatment assignments are effectively only distinguished by the number of units treated. Since we observed no covariates and have random sampling, there is no bias (even ex post), and the MSE of any assignment with n_1 treated units is equal to

$$\text{MSE}(d_1, \ldots, d_n) = \text{Var}(\hat{\beta}) = \left(\frac{1}{n_1} + \frac{1}{n_0}\right) \cdot (\text{Var}(X_i) + 1). \tag{7}$$

Any procedure that randomly assigns $n_1 = n/2 = 2$ units to treatment 1 is optimal in this case. This is, of course, the standard recommendation. A similar argument applies when we observe a discrete covariate, with several observations i for each value of the covariate. In this latter case, randomization conditional on X_i is optimal; this is what is known as a blocked design.

2.2 *Randomized designs*

Let us now, and for the rest of our discussion, assume again that we observed the covariates X_i. The calculations we have done so far for this case are for a fixed (deterministic) assignment of treatment. What if we use randomization? Any randomization procedure in our setting can be described by the probabilities $p(d_1, d_2, d_3, d_4)$ it assigns to the different rows of Table 1. The MSE of such a procedure is given by the corresponding weighted average of MSEs for each row:

$$\text{MSE} = \sum_{d_1,\ldots,d_n} p(d_1, \ldots, d_n) \cdot \text{MSE}(d_1, \ldots, d_n). \tag{8}$$

[4]This is the ex post bias, given covariates and treatment assignment. This is the relevant notion of bias for us. It is different from the ex ante bias, which is the expected bias when we do not know the treatment assignment yet.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

We can now compare various randomization procedures (experimental designs) in terms of their expected MSE. Let us first consider the completely random procedure where each D_i is determined by an independent coin flip. Such a procedure assigns equal probability to all sixteen possible treatment assignments. Since there is a chance of 1 in 8 to get all 0s or all 1s, as in the last two assignments in Table 1, the variance of an estimator based on such an assignment is technically infinite. This case corresponds to the first of the designs in Table 1.

Any reasonable experimenter would rerandomize if she got all 0s or all 1s. If we do this, we eliminate the last two rows of Table 1. This procedure still has a sizable chance of picking an assignment with a large MSE, such as for instance $d = (0, 0, 0, 1)$. The MSE for this rerandomization procedure, labeled "design 2," is equal to 3.2, cf. the row labeled "MSE model 1."

Now, again, most experimenters would not use this procedure either. Standard procedure is to reduce randomness further by only allowing for treatment assignments that have treatment and control groups of equal size. This eliminates all but the first six rows and yields the third design considered. The advantage of doing so is to reduce variance relative to designs allowing for unequal group sizes. Correspondingly, the average MSE is reduced to 2.7.

So far we have ignored the information available about the covariate X. Many experimenters use such information by blocking, that is, by forming groups based on covariate values, and randomizing within these groups. A special case is pairwise randomization, where groups are of size 2. In our case, we could group observations 1 and 2, as well as observations 3 and 4, and randomly assign one unit within each group to treatment and one to control. This gives the fourth design and leaves us with random assignment over rows 1 through 4, resulting in a further reduction of average MSE to 1.5.

This procedure, which already has more than halved the expected MSE relative to randomization over all possible assignments, can still be improved upon. The assignment $0, 1, 0, 1$ has systematically higher values of X among the treated relative to the control, resulting in a bias of 1 in our setting, and systematically lower values of X among the treated for the assignment $1, 0, 1, 0$, resulting in a bias of -1. Eliminating these two leaves us with the last design that only allows for the first two treatment assignments. Since these two have the same MSE, we might as well randomize between them. This is what the present article ultimately recommends. This yields an average MSE of 1 in the present setting.

2.3 *Other data-generating processes*

Let us now take a step back and consider what we just did. We first noted that the MSE of any randomized procedure is given by the weighted average of the MSEs of the deterministic assignments it is averaging over. By eliminating assignments with larger MSE from the support of the randomization procedure, we could systematically reduce the MSE. In the limit, we would only randomize between the two assignments with the lowest MSE.

The careful reader will have noted that all our calculations of the MSE depend on the assumed model determining potential outcomes. So let us consider a different model ("model 2") where the covariate affects outcomes in a different way,

$$Y_i^d = -x_i^2 + d + \epsilon_i^d. \tag{9}$$

Relative to model 1, we have flipped the sign of the effect of X, and we have made the relationship nonlinear. Table 1 shows the corresponding values of bias, variance, and MSE for model 2. Again the ranking over the alternative randomized designs is the same as before (cf. the row labeled "MSE model 2"), even though the values of the average MSE are changed. There appears to be a comforting robustness of the optimal procedure.

The careful reader will now again note that we have demonstrated this for two particular models, but one might be able to construct models where the ranking of deterministic assignments is partially reversed. That is indeed true. Experimental design has to proceed without knowledge of the underlying data-generating process. The experimenter, therefore, must trade off

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

performance across alternative data-generating processes, as in any statistical decision problem. This can be done explicitly using a (nonparametric) Bayesian prior. Such a prior allows us to calculate the expected MSE, where the expectation averages over possible data-generating processes. That is the approach we will propose in Section 4 below. The result is an expected MSE for each assignment $(d_1, \ldots, d_n)$. The last column of Table 1 shows the expected MSE for each deterministic treatment assignment; randomized procedures can again be evaluated in terms of weighted averages of the deterministic MSEs.[5] We can then pick an optimal such assignment, or randomize over a set of assignments with low MSE. Before we go there, however, we shall take a step back and briefly review general statistical decision theory. This allows us to demonstrate that the arguments against randomization, or for randomization over a selected set of assignments, and the intuition of our little example, hold under very general circumstances for any statistical decision problem.

3 Decision Theory Tells Us We Should Not Randomize

3.1 *Brief review of decision theory*

Our argument requires that we think of experimental design as a decision problem. Statistical decision theory provides one of the conceptual foundations for modern statistics. Statistical decision theory was pioneered by Abraham Wald in the 1940s. A classic introduction can be found in Berger (1985). We shall provide a brief review of decision theory in this section.

The basic idea of decision theory is to think of statistical problems as involving a choice of some action a (picking an estimate, deciding whether to reject a hypothesis, picking an experimental treatment allocation d,...). The choice is made based on the available evidence (data) X, so that $a = \delta(X)$. The action will ultimately be evaluated by a loss function L, where loss also depends on the unknown state of the world, θ, so $L = L(a, \theta)$. An estimate might be evaluated, for instance, by how much it deviates from the truth, $L = (\hat{\beta} - \beta)^2$. Loss can be thought of as the negative of a utility function, a familiar object in economic and game theoretic models.

Since we do not know what the true state of the world is, we cannot evaluate a procedure based on its actual loss. We can, however, consider various notions of average loss. If we average loss over the randomness of a sampling scheme, and possibly over the randomness of a treatment allocation scheme, holding the state of the world θ fixed, we obtain the frequentist risk function:

$$R(\delta, \theta) = E[L(\delta(X), \theta)|\theta]. \tag{10}$$

We can calculate this frequentist risk function. If loss is the squared error of an estimate $\hat{\beta}$ for some true parameter β, then the frequentist risk function is equal to the MSE we calculated in the example of Section 2 for each row of Table 1. The models 1 and 2 that we considered correspond to two values of θ.

The risk function by itself does not yet provide an answer to the question of which decision procedure to pick, unfortunately. The reason is that one procedure δ might be better in terms of $R(\delta, \theta)$ for one state of the world, while another procedure δ' is better for another state θ' of the world. Returning again to Table 1, we have for instance that the assignment in row 10 is better than row 3 for model 1, but worse for model 2.

We thus have to face a trade-off across states θ. There are different ways to do this. One way is to focus on the worst-case scenario that is on the state of the world with the worst value of the risk

[5]The prior used to calculate this *EMSE* assumes that

$$C((x_1, d_1), (x_2, d_2)) = 10 * \exp\left(-\left(||x_1 - x_2||^2 - (d_1 - d_2)^2\right)/10\right);$$

for details, see Section 4.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

function. This yields the minimax decision criterion. Another way is to assign weights π (prior probabilities) to different states of the world, and to minimize the weighted average. This yields Bayesian procedures.

3.2 *Randomization in general decision problems*

Let us now return to the question of randomization. Why does decision theory tell us we should not randomize? Suppose we allow our decision procedure δ to depend on a randomization device U, so that $a = \delta(U, X)$. U does not tell us anything about the state of the world θ, and U does not directly enter the loss function L. As a consequence, the frequentist risk function for the randomized procedure is simply a weighted average of the risk functions for the nonrandom decisions that our random procedure might pick. An example is the average MSE we calculated in Section 2, cf. the last two rows of Table 1.

Formally, let $\delta^u(x) = \delta(u, x)$ be the decision made when $U = u$ and $X = x$, and let $R(\delta^u, \theta)$ be the risk function of the procedure that picks $\delta^u(x)$ whenever $X = x$. Then, we have

$$R(\delta, \theta) = \sum_u R(\delta^u, \theta) \cdot P(U = u), \tag{11}$$

where we have assumed for simplicity of notation that U is discrete.

To get to a global assessment of δ, we need to trade off risk across values of θ. If we do so using the prior distribution π, we get the Bayes risk

$$\begin{aligned} R^\pi(\delta) &= \int R(\delta, \theta) d\pi(\theta) \\ &= \sum_u \int R(\delta^u, \theta) d\pi(\theta) \cdot P(U = u) \\ &= \sum_u R^\pi(\delta^u) \cdot P(U = u). \end{aligned} \tag{12}$$

In terms of Table 1, Bayes risk averages the MSE for a given assignment δ^u and data-generating process θ, $R(\delta^u, \theta)$, both across the rows, using the randomization device U, and across different models for the data-generating process, using the prior distribution π. If, alternatively, we evaluate each action based on the worst-case scenario for that action, we obtain minimax risk

$$R^{mm}(\delta) = \sum_u \left(\sup_\theta R(\delta^u, \theta) \right) \cdot P(U = u).$$

In terms of Table 1, minimax risk evaluates each row (realized assignment δ^u) in terms of the worst-case model for this assignment, $\sup_\theta R(\delta^u, \theta)$, and evaluates randomized designs in terms of the weighted average across the worst case risk for each assignment.

Letting $R^*(\delta)$ denote either Bayes or minimax risk, we have

$$R^*(\delta) = \sum_u R^*(\delta^u) \cdot P(U = u). \tag{13}$$

The risk of any randomized procedure is given by the weighted average of risk across the deterministic procedures it averages over. Now consider an alternative decision procedure such that

$$\delta^* \in \operatorname*{argmin}_\delta R^*(\delta), \tag{14}$$

where the argmin is taken over non-randomized decision functions $a = \delta(X)$. It follows that

$$R^*(\delta^*) \leq \min_u R^*(\delta^u) \leq R^*(\delta). \tag{15}$$

The second inequality is strict, unless δ only puts positive probability on the minimizers, over u, of

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

$R^*(\delta^u)$. If the optimal deterministic δ^* is unique, $R^*(\delta^*) < R^*(\delta)$ unless δ assigns probability $1–\delta^*$, so that δ does not involve randomization. We have thus shown the following proposition:

Proposition 1 (Optimality of non-randomized decisions). *Consider the general decision problem discussed in Section 3.1. Let R^* be either Bayes risk or minimax risk. Then:*

1. *The optimal risk $R^*(\delta^*)$ when considering only deterministic procedures $\delta(X)$ is no larger than the optimal risk when allowing for randomized procedures $\delta(X, U)$.*
2. *If the optimal deterministic procedure δ^* is unique, then it has strictly lower risk than any non-trivial randomized procedure.*

3.3 *Experimental design*

We can think of an experiment as a decision problem that proceeds in several stages:

1. Sample observations i from the population of interest, and collect baseline covariates X_i.
2. Using the baseline covariates X and possibly a randomization device U, allocate units i to treatments D_i.
3. After running the experiment, collect data on outcomes Y_i of interest. Use these to calculate estimators $\hat{\beta}$ of the objects β that we are interested in. In many settings, β is the average treatment effect.
4. Finally, loss is realized, $L = L(\hat{\beta}, \beta)$. A common choice is $L = (\hat{\beta} - \beta)^2$.[6]

Since this article is about experimental design, we are interested in the second step. Taking everything else (sampling and the matrix of baseline covariates X, the choice of estimator $\hat{\beta}$ and object of interest β, the loss function L) as given, we can ask what is the best way to assign treatment D, using the information in the baseline covariates X, and possibly the randomization device U.

In order to answer this question, we first calculate the risk function of a deterministic choice of treatment allocation $\delta = (d_1, \ldots, d_n)$, which only depends on X but not on U. If loss is the quadratic estimation error $(\hat{\beta} - \beta)^2$, this gives the MSE

$$R(\delta, \theta) = \text{MSE}(d_1, \ldots, d_n) = E[(\hat{\beta} - \beta)^2 | X, \theta]. \tag{16}$$

This is the calculation we did in Section 2 to get the MSE columns for models 1 and 2. The two models can be thought of as two values of θ at which we evaluate the risk function.

As in the general decision problem, we get that the risk function of a randomized procedure is given by the weighted average of the risk functions for deterministic procedures, $R(\delta, \theta) = \sum_u R(\delta^u, \theta) \cdot P(U = u)$. Consequently, the same holds for Bayes risk and minimax risk. This implies that optimal deterministic procedures perform at least as well as any randomized procedures.

But do they perform better? Do we lose anything by randomizing? As stated in Proposition 1, the answer to this question hinges on whether the optimal deterministic assignment is unique. Any assignment procedure, random or not, is optimal if and only if it assigns positive probability only to assignments $(d_1, \ldots, d_n)$ that minimize risk among deterministic assignments.

If we were to observe no covariates, then permutations of treatment assignments would not change risk, since everything "looks the same" after the permutation. In that case, randomization is perfectly fine, as long as the size of treatment and control groups are fixed. Suppose next that we observe a covariate with small support, so that we get several observations for each value of the covariate. In this case, we can switch treatment around within each block defined by a covariate

[6] Squared error loss is the canonical loss function in the literature on estimation. It has a lot of convenient properties allowing for tractable analytic results, and in keeping with the literature we will focus on squared error loss. Other objective functions are conceivable, such as for instance expected welfare of treatment assignments based on experimental evidence; see for instance Kasy (2014).

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

value without changing risk—again, everything "looks the same" after the permutation. So in this case, stratified (blocked) randomization is optimal.

Suppose now, however, that we observe continuous covariates or discrete covariates with large support. In that case, in general, no two observations in our sample will look the same. We cannot just switch treatment around without changing risk. In that case, therefore, the optimal assignment will in general be unique (possibly up to a "sign flip," switching 0s and 1s in treatment assignment). This implies that randomizing comes with a cost of larger MSE relative to the optimal deterministic assignment.

4 Our Proposed Procedure

There is one key question we need to answer in order to implement optimal designs in practice: how to aggregate the risk function $R(\delta, \theta)$ into a global notion of risk R^*. In this section, we take a Bayesian approach, using a nonparametric Bayesian prior over possible data-generating processes.

The key object that we need prior is the conditional expectation of potential outcomes given covariates, since it is this function that determines the bias of any given estimate of the average treatment effect. Denote this function by f, where

$$f(x, d) = E[Y_i^d | X_i = x, \theta]. \tag{17}$$

In model 1 of Section 2, we had $f(x, d) = x + d$; in model 2, we had $f(x, d) = -x^2 + d$. We assume that the prior for f is such that expectations and variances exist. We denote

$$E[f(x, d)] = \mu(x, d), \text{ and}$$

$$\text{Cov}(f(x_1, d_1), f(x_2, d_2)) = C((x_1, d_1), (x_2, d_2)) \tag{18}$$

for a covariance function C. Assume further that we have a prior over the conditional variance of Y_i which satisfies

$$E[\text{Var}(Y_i | X_i, D_i, \theta) | X_i, D_i] = \sigma^2(X_i, D_i) = \sigma^2. \tag{19}$$

Assuming a prior centered on homoskedasticity is not without loss of generally, but is a natural baseline.

How should the prior moments μ and C be chosen? We provide an extensive discussion in the Online Appendix; a popular choice in the machine-learning literature is so-called squared exponential priors, where $\mu \equiv 0$ and

$$C((x_1, d_1), (x_2, d_2)) = exp\left(-\left(||x_1 - x_2||^2 - (d_1 - d_2)^2\right)/l\right). \tag{20}$$

The length scale parameter l determines the assumed smoothness of f for such priors. This is the prior we used to calculate the expected MSE in Table 1, choosing a length scale of $l = 10$. Another popular set of priors builds on linear models. For such priors, we show below that expected MSE corresponds to balance as measured by differences in covariate means.

Recall that we are ultimately interested in estimating the conditional average treatment effect,

$$\beta = \frac{1}{n} \sum_i E[Y_i^1 - Y_i^0 | X_i]. \tag{21}$$

The conditional average treatment effect is the object of interest if we want to learn about treatment effects for units, in the population from which our sample was drawn, which look similar in terms of covariates. We might be interested in this effect both for scientific reasons and for policy reasons (deciding about future treatment allocations). One more question has to be settled before we can give an expression for the Bayes risk of any given treatment assignment: How is β going to be estimated? We will consider two possibilities. The first possibility is that we use an estimator that is optimal in the Bayesian sense, namely the posterior best linear predictor of β. The second possibility is that we estimate β using a simple difference in means, as in the example of Section 2.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

4.1 *Bayes optimal estimation*

It is useful to introduce the following notation for the relevant prior moments:

$$\begin{aligned}\mu_i &= \mu(X_i, d_i)\\ C_{i,j} &= C((X_i, d_i), (X_j, d_j))\\ \mu_\beta &= \frac{1}{n}\sum_i [\mu(X_i, 1) - \mu(X_i, 0)],\\ \overline{C}_i &= \frac{1}{n}\sum_j \big[C((X_i, d_i), (X_j, 1)) - C((X_i, d_i), (X_j, 0))\big].\end{aligned} \tag{22}$$

Let μ and C be the corresponding vector and matrix with entries μ_i and $C_{i,j}$. Note that both μ and C depend on the treatment assignment $(d_1, \ldots, d_n)$. Using this notation, the prior mean and variance of Y are equal to μ and $C + \sigma^2 I$, and the prior mean of β equals μ_β.

Let us now consider the posterior best linear predictor, which is the best estimator (in the Bayesian risk sense) among all estimators linear in Y.[7] The posterior best linear predictor is equal to the posterior expectation if both the prior for f and the distribution of the residuals $Y - f$ are multivariate normal.

Proposition 2 (Posterior best linear predictor and expected loss). *The posterior best linear predictor for the conditional average treatment effect is given by*

$$\hat{\beta} = \mu_\beta + \overline{C}' \cdot (C + \sigma^2 I)^{-1} \cdot (Y - \mu), \tag{23}$$

and the corresponding MSE (Bayes risk) equals

$$\mathrm{MSE}(d_1, \ldots, d_n) = \mathrm{Var}(\beta|X) - \overline{C}' \cdot (C + \sigma^2 I)^{-1} \cdot \overline{C}, \tag{24}$$

where $\mathrm{Var}(\beta|X)$ *is the prior variance of* β.

The proof of this proposition follows from standard characterizations of best linear predictors and can be found in Appendix A. The expression for the MSE provided by equation (24) is easily evaluated for any choice of $(d_1, \ldots, d_n)$. Since our goal is to minimize the MSE, we can in fact ignore the $\mathrm{Var}(\beta|X)$ term, which does not depend on $(d_1, \ldots, d_n)$.

4.2 *Difference in means estimation*

Let us next consider the alternative case where the experimenter uses the simple difference in means estimator. We will need the following notation:

$$\begin{aligned}\mu_i^d &= \mu(X_i, d)\\ C_{i,j}^{d^1,d^2} &= C((X_i, d^1), (X_j, d^2)),\end{aligned} \tag{25}$$

for $d, d^1, d^2 \in \{0, 1\}$. We collect these terms in the vectors μ^d and matrices C^{d^1,d^2}, which are in turn collected as

$$\begin{aligned}\tilde{\mu} &= (\mu^1, \mu^2)\\ \tilde{C} &= \begin{pmatrix} C^{00} & C^{01} \\ C^{10} & C^{11} \end{pmatrix}.\end{aligned} \tag{26}$$

[7] This class of estimators includes all standard estimators of β under unconfoundedness, such as those based on matching, inverse probability weighting, regression with controls, kernel regression, series regression, splines, etc. Linearity of the estimator is unrelated to any assumption of linearity in X; we are considering the posterior BLP of β in Y rather than the BLP of Y_i in X_i.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

$\tilde{\mu}$ and $\tilde{C}$ do *not* depend on $(d_1, \ldots, d_n)$, in contrast to μ and C defined before. The latter are a sub-vector / sub-matrix of the former, selected through a choice of $(d_1, \ldots, d_n)$. Defining $\tilde{f}$ analogously to $\tilde{\mu}$, we get $E[\tilde{f}] = \tilde{\mu}$ and $\mathrm{Var}(\tilde{f}) = \tilde{C}$.

Proposition 3 (Expected MSE for designs, using the difference in means estimator). *Assume that β is estimated using the difference in means estimator*

$$\hat{\beta} = \frac{1}{n_1}\sum_i D_i Y_i - \frac{1}{n_0}\sum_i (1 - D_i) Y_i. \tag{27}$$

Then,

$$\mathrm{MSE}(d_1, \ldots, d_n) = \sigma^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_0}\right] + (w' \cdot \tilde{\mu})^2 + w' \cdot \tilde{C} \cdot w, \tag{28}$$

where

$$\begin{aligned} w &= (w^0, w^1), \\ w_i^1 &= \frac{d_i}{n_1} - \frac{1}{n}, \\ w_i^0 &= -\frac{1 - d_i}{n_0} + \frac{1}{n}. \end{aligned} \tag{29}$$

The expression for the MSE in this proposition has three terms. The first term is the variance of the estimator. The second and third terms together are the expected squared bias. This splits in turn into the square of the prior expected bias, and the prior variance of the bias.

It is interesting to note that we recover the standard notion of balance if, in addition to the assumptions of Proposition 3, we impose a linear, separable model for f, that is,

$$f(x, d) = x' \cdot \gamma + d \cdot \beta, \tag{30}$$

where γ has prior mean 0 and variance Σ. For this model, we get that the bias of $\hat{\beta}$ is equal to

$$(\overline{X}^1 - \overline{X}^2)' \cdot \beta, \tag{31}$$

where $\overline{X}^d$ is the sample mean of covariates in either treatment arm, and thus the prior expected squared bias is equal to

$$(\overline{X}^1 - \overline{X}^2)' \cdot \Sigma \cdot (\overline{X}^1 - \overline{X}^2), \tag{32}$$

and the MSE equals

$$\mathrm{MSE}(d_1, \ldots, d_n) = \sigma^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_0}\right] + (\overline{X}^1 - \overline{X}^2)' \cdot \Sigma \cdot (\overline{X}^1 - \overline{X}^2). \tag{33}$$

Risk is thus minimized by choosing treatment and control arms of equal size, and optimizing balance as measured by the difference in covariate means $(\overline{X}^1 - \overline{X}^2)$.

4.3 *Discrete optimization*

We now have almost all the ingredients for a procedure that can be used by practitioners. One important element is missing: How do we find the optimal assignment d^*? Or how can we at least find a set of assignments that are close to optimal in terms of expected risk? This question arises since solving the problem by brute force is generally not feasible. We could do so in our stylized example in Section 2, since in this example there were only $2^4 = 16$ possible treatment assignments. In general, there are 2^n assignments for a sample of size n, a number that very quickly becomes prohibitive to search by brute force.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

So what to do? One option, which is close to a common practice among experimenters, is to rerandomize a feasible number of times, and to pick the best assignment among those randomly picked. A quick calculation shows that this procedure performs reasonably well. Suppose we rerandomize k times, and pick $\hat{d}$ as the best of the k assignments we got. Then, the probability that $\hat{d}$ is better then 99% of all assignments, say, is equal to $1-0.99^k$. If $k=500$, this chance is already larger than 99%, and significantly larger k are feasible in practice.

Implementation of our design procedure using rerandomization is simple and can be summarized as follows: First, pick a prior. The Online Appendix discusses good defaults. For our example in Section 2, we used $\mu=0$ and $C((x_1,d_1),(x_2,d_2))=10*\exp\left(-\left(||x_1-x_2||^2-(d_1-d_2)^2\right)/10\right)$. Calculate the corresponding matrix $\tilde{C}$, and $\tilde{C}^{-1}$. Then, iterate the following:

1. Draw a random treatment assignment.
2. Calculate the objective function of equation (28).
3. Compare it to the best MSE obtained thus far. If the new MSE is better than the old one, store the new treatment assignment and MSE.
4. Iterate for some prespecified number of times k.

More sophisticated alternative optimization methods are discussed in the Online Appendix.

5 Arguments for Randomization

There are a number of arguments that can be made for randomization and against the framework considered in this article. We shall discuss some of these, and to what extent they appear to be valid.

5.1 *Randomization inference requires randomization*

That is correct. How does this relate to our argument? The arguments of Section 3 show that optimal procedures in a decision theoretic sense do not rely on randomization. Randomization inference cannot be rationalized as an optimal procedure under the conceptual framework of decision theory, however. As a consequence, the arguments of Section 3 do not apply.

One could take the fact that randomization inference is not justified by decision theory as an argument against randomization inference. But one could also consider a compromise approach that is based on randomization among assignments that have a low expected MSE. Such partially random assignments are still close to optimal *and* allow the construction of randomization-based tests.

5.2 *Is identification assuming conditional independence still valid without randomization?*

Yes, it is. Selection on observables holds, as the name suggests, for any treatment assignment that is a function of observables. Put differently, conditional independence is guaranteed by any controlled trial, whether randomized or not, as stated in the following proposition.

Proposition 4 (Conditional independence)
Suppose that (X_i, Y_i^0, Y_i^1) *are i.i.d. draws from the population of interest, which are independent of* U. *Then, any treatment assignment of the form* $D_i = d_i(X_1, \ldots, X_n, U)$ *satisfies conditional independence,*

$$(Y_i^0, Y_i^1) \perp D_i | X_i. \tag{34}$$

This is true, in particular, for deterministic treatment assignments of the form $D_i = d_i(X_1, \ldots, X_n)$.

5.3 *The Bayesian approach to experimental design requires the choice of a prior*

That is correct, as far as the ranking of assignments in terms of MSE is concerned. A treatment assignment might be optimal for a particular prior and objective (such as minimizing the mean squared error (MSE) of an estimator for the average treatment effect (ATE)), but not for another prior or objective.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

This does not imply, however, that the *interpretation* of the data obtained from the experiment has to rely on the same prior as the one used for *designing* the experiment, or on any prior at all. From the perspective of a researcher analyzing data produced by an experiment where treatment was assigned based on observables as well as a randomization device, it is immaterial how treatment was assigned, as long as the conditional independence condition $(Y_i^0, Y_i^1) \perp D_i | X_i$ holds. This is guaranteed by Proposition 4. The design can furthermore be chosen to be optimal for estimators such as the difference in means, which does not use a prior, as shown in Proposition 3.

5.4 *Randomized assignments are perceived as fair, and thus politically easier to implement*

A possible objection to the practical feasibility of optimal designs might be a political perception that randomized assignments are fair, while assignment based on covariates such as sex or race is not fair, in a similar way that "tagging" in taxation is considered unfair.[8] This might sometimes be a constraint on feasible experiments.

Note, however, that optimal designs as discussed in this article seek to balance the covariate distribution across treatments, leading to a *more* equitable distribution of treatment across demographic or other groups, relative to random assignments.

5.5 *Does the proposed approach rely on a parametric model?*

No, it does not. Nowhere have we imposed restrictions on the distribution of potential outcomes or the distribution of covariates, beyond the assumption of i.i.d. sampling. The family of priors considered in Section 4, in particular, is nonparametric, and has support on all sufficiently smooth functions.[9]

6 Conclusion

In this article, we discuss the question of how information from baseline covariates should be used when assigning treatment in an experiment. In order to give a well-grounded answer to this question, we adopt a decision theoretic and nonparametric framework. The nonparametric perspective and the consideration of continuous covariates distinguish this article from much of the previous experimental design literature.

A number of conclusions emerge from our analysis. First, randomization is in general not optimal. Rather, we should pick a risk-minimizing treatment assignment, which is generically unique in the presence of continuous covariates. Second, we can consider nonparametric priors that lead to very tractable estimators and expressions for Bayesian risk (MSE). The general form of the expected MSE for such priors is $\mathrm{Var}(\beta|X) - \overline{C}' \cdot (C + \Sigma)^{-1} \cdot \overline{C}$, where $\overline{C}$ and C are the appropriate covariance vector and matrix from the prior distribution, cf. Section 4. We suggest picking the treatment assignment that minimizes this prior risk. Finally, conditional independence between potential outcomes and treatments given covariates does not require random assignment. It is ensured by conducting *controlled* trials, and does not rely on *randomized* controlled trials. Matlab code to implement our proposed approach is available online at http://dx.doi.org/10.7910/DVN/I5KCWI.

Conflict of interest statement. None declared.

[8] I thank Larry Katz for pointing this out.

[9] To be precise, the support of these priors is the closure of the Reproducing Kernel Hilbert Space corresponding to C. For further discussion, see the Online Appendix.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

Appendix A: Proofs

Proof of Proposition 2. By the usual formula for best linear predictors,

$$\hat{\beta} = E[\beta|X, D] + \mathrm{Cov}(\beta, Y|X, D) \cdot \mathrm{Var}(Y|X, D)^{-1} \cdot (Y - E[Y|X, D]). \tag{A1}$$

Using the notation introduced before Proposition 2, we can denote these terms by

$$\begin{aligned} E[\beta|X, D] &= \mu_\beta \\ \mathrm{Cov}(\beta, Y|X, D) &= \overline{C}' \\ \mathrm{Var}(Y|X, D) &= C + \sigma^2 I \\ E[Y|X, D] &= \mu, \end{aligned} \tag{A2}$$

which yields equation (23). We furthermore have, by the general properties of best linear predictors, that

$$\mathrm{MSE}(d_1, \ldots, d_n) = E[(\beta - \hat{\beta})^2|X, D] = \mathrm{Var}(\beta - \hat{\beta}|X, D), \tag{A3}$$

and $\mathrm{Cov}(\hat{\beta}, \beta - \hat{\beta}|X, D) = 0$, so that

$$\mathrm{Var}(\beta|X, D) = \mathrm{Var}(\hat{\beta}|X, D) + \mathrm{Var}(\beta - \hat{\beta}|X, D). \tag{A4}$$

This immediately implies equation (24). □

Proof of Proposition 3. Let $\epsilon_i^d = Y_i^d - f(X_i, d_i)$. We can write

$$\Delta := \hat{\beta} - \beta = \sum_i \left[\left(\frac{d_i}{n_1} - \frac{1 - d_i}{n_0} \right) \cdot (f(X_i, d_i) + \epsilon_i^d) - \frac{1}{n}(f(X_i, 1) - f(X_i, 0)) \right] \tag{A5}$$

and

$$\begin{aligned} \mathrm{MSE}(d_1, \ldots, d_n) = E[\Delta^2] &= \mathrm{Var}(\Delta) + E[\Delta]^2 \\ &= E[\mathrm{Var}(\Delta|f)] + \mathrm{Var}(E[\Delta|f]) + E[\Delta]^2. \end{aligned} \tag{A6}$$

The first term is equal to

$$\mathrm{Var}(\Delta|f) = \sigma^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_0} \right]. \tag{A7}$$

The second term is equal to the variance of

$$E[\Delta|f] = w' \cdot \tilde{f}. \tag{A8}$$

The third term is equal to the square of

$$\begin{aligned} E[\Delta] &= E[E[\Delta|f]] \\ &= E\left[w' \cdot \tilde{f} \right] \\ &= w' \cdot E[\tilde{f}]. \end{aligned} \tag{A9}$$

The claim follows once we recall $E[\tilde{f}] = \tilde{\mu}$ and $\mathrm{Var}(\tilde{f}) = \tilde{C}$. □

Proof of Proposition 4. The assumption of independent sampling implies that

$$(X_i, Y_i^0, Y_i^1) \perp (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, U), \tag{A10}$$

thus

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press

$$(Y_i^0, Y_i^1) \perp (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, U) | X_i, \tag{A11}$$

and therefore

$$(Y_i^0, Y_i^1) \perp d_i(X_1, \ldots, X_n, U) | X_i. \tag{A12}$$

□

References

Berger, J. 1985. *Statistical decision theory and Bayesian inference*. New York: Springer.

Blattman, C., A. C. Hartman, and R. A. Blair. 2014. How to promote order and property rights under weak rule of law? An experiment in changing dispute resolution behavior through community education. *American Political Science Review* 108:100–120.

Findley, M. G., D. L. Nielson, and J. Sharman. 2015. Causes of noncompliance with international law: A field experiment on anonymous incorporation. *American Journal of Political Science* 59(1):146–61.

Kalla, J. L., and D. E. Broockman. 2015. Campaign contributions facilitate access to congressional officials: A randomized field experiment. *American Journal of Political Science*. "http://sfx.hul.harvard.edu/hvd?__char_set=utf8&id=doi:10.1111/ajps.12180&sid=libx%3Ahul.harvard&genre=article" doi: 10.1111/ajps.12180

Kasy, M. 2014. *Using data to inform policy*. Working Paper.

———. 2016. *Matlab implementation for: Why experimenters might not always want to randomize, and what they could do instead,* Harvard Dataverse. http://dx.doi.org/10.7910/DVN/I5KCWI.

Keele, L. 2015. The statistics of causal inference: A view from political methodology. *Political Analysis*. doi: 10.1093/pan/mpv007.

Moore, R. T. 2012. Multivariate continuous blocking to improve political science experiments. *Political Analysis* 20(4):460–79.

Morgan, K. L., and D. B. Rubin. 2012. Rerandomization to improve covariate balance in experiments. *Annals of Statistics* 40(2):1263–82.

Nyhan, B., and J. Reifler. 2014. The effect of fact-checking on elites: A field experiment on U.S. state legislators. *American Journal of Political Science*.

https://doi.org/10.1093/pan/mpw012 Published online by Cambridge University Press