The Econometrics Journal

Econometrics Journal (2020), volume **23**, pp. 1–31.
doi: 10.1093/ectj/utz020

Optimal data collection for randomized control trials

PEDRO CARNEIRO[*,†,‡], SOKBAE LEE[§,†,‡] AND DANIEL WILHELM[*,†,‡]

[*]*University College London, Gower Street, London WC1E 6BT, UK.*
E-mail: p.carneiro@ucl.ac.uk, d.wilhelm@ucl.ac.uk

[†]*Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK.*

[‡]*Centre for Microdata Methods and Practice, 7 Ridgmount Street, London WC1E 7AE, U.K.*

[§]*Columbia University, 420 West 118th Street, New York, NY 10027, USA.*
E-mail: sl3841@columbia.edu

First version received: 10 January 2019; final version accepted: 4 May 2019.

Summary: In a randomized control trial, the precision of an average treatment effect estimator and the power of the corresponding t-test can be improved either by collecting data on additional individuals, or by collecting additional covariates that predict the outcome variable. To design the experiment, a researcher needs to solve this trade-off subject to her budget constraint. We show that this optimization problem is equivalent to optimally predicting outcomes by the covariates, which in turn can be solved using existing machine learning techniques using pre-experimental data such as other similar studies, a census, or a household survey. In two empirical applications, we show that our procedure can lead to reductions of up to 58% in the costs of data collection, or improvements of the same magnitude in the precision of the treatment effect estimator.

Keywords: *data collection*, *randomized control trials*, *machine learning*.

JEL codes: *C81*, *C51*, *C52*, *C55*.

1. INTRODUCTION

This paper is motivated by the observation that empirical research in economics increasingly involves the collection of original data through laboratory or field experiments (see, e.g., Duflo et al., 2007; Banerjee and Duflo, 2009; Bandiera et al., 2011; List, 2011; List and Rasul, 2011; Hamermesh, 2013, among others). This observation carries with it a call and an opportunity for research to provide econometrically sound guidelines for data collection.

We analyse the decision problem faced by a researcher designing the survey for a randomized control trial (RCT) in the presence of a budget constraint. We assume that the goal of the researcher is to obtain precise estimates of the average treatment effect and/or a powerful t-test of the hypothesis of no treatment effect, using the experimental data.[1] Data collection is costly, and the research budget limits how much data can be collected. We ask how the researcher optimally trades off the number of individuals included in the RCT against the choice of covariates included in the survey.

[1] Tetenov (2016) provides a decision-theory-based rationale for using hypothesis tests in the RCT, and Banerjee et al. (2017) develop a theory of experimenters, focusing on the motivation of randomization among other things. Kasy (2016) uses the setup of a statistical decision problem to study experimental design.

For example, consider an RCT for studying the impact of an education intervention on students' test scores. This is a typical case where controlling for covariates, namely pre-intervention test scores, leads to improvements in the precision of experimental treatment effects. This is because pre-intervention test scores are highly predictive of post-intervention test scores.

In this context, we can ask whether one achieves more precise estimates of the treatment effect by spending the entire budget collecting the outcome (post-intervention test scores) on a large sample of students ignoring any baseline covariates; or by spending the budget on a smaller sample of students who are tested twice, pre- and post-intervention.

We show below how a rigorous analysis of this problem can potentially lead to first-order improvements in the precision of treatment effect estimates, and to large reductions in the costs of collecting data in these studies. In two realistic empirical applications, it is possible to achieve reductions of 58% in data collection costs, and similar decreases in the variance of the treatment effect estimates.

There are, of course, other factors potentially influencing the choice of covariates to be collected in a survey for an RCT. For example, one may wish to learn about the mechanisms through which the RCT is operating, check whether treatment or control groups are balanced, or measure heterogeneity in the impacts of the intervention being tested. In practice, researchers place implicit weights on each of the objectives they consider when designing surveys, and informally work out the different trade-offs involved in their choices. We show that there is substantial value to making this decision process more rigorous and transparent through the use of data-driven tools that optimize a well-defined objective. Instead of attempting to formalize the whole research design process, we focus on one particular trade-off that is of first-order importance and particularly conducive to data-driven procedures.

We begin by assuming that the researcher has access to pre-experimental data from the population from which the experimental data will be drawn, or at least from a population that shares similar second moments of the variables to be collected. The data set includes the outcome and all potentially relevant covariates that one would consider collecting for the analysis of the experiment. This assumption may be reasonable in many different contexts. Table 1 lists some of these examples.[2] At the top of the table we have several RCTs in education, all taking place in Kenya and examining the impact of particular interventions on test scores and other outcomes. At the bottom we have several RCTs[3] examining the impact of access to microcredit on investment, expenditure, consumption, and other outcomes, conducted across multiple countries.[4] In standard power calculations, researchers informally rely on their knowledge of other existing studies when choosing parameters of the data-generating process. We formalize this link by directly including the data from such prior studies into our process of designing the experiment.

Our procedure might also be useful in the design of (experimental and quasi-experimental) studies based on administrative records, such as the recent study of the Oregon Health Insurance Experiment (Finkelstein et al., 2012). These are cases where administrative records are already available, but need to be assembled, organized and interlinked, at a cost.[5]

[2] See also the Abdul Latif Jamil Poverty Action Lab (JPAL) website (https://www.povertyactionlab.org), which lists several RCTs by topic conducted with multiple datasets.

[3] The entire January issue of American Economic Journal: Applied Economics consists of six randomized Evaluations of Microcredit (Banerjee et al., 2015). See also Meager (2019) for evaluating the external validity of microcredit.

[4] There are several other papers we could add to the table. We chose these either because they focused on a specific area and concerned a similar topic, or because they were on exactly the same issue but took place in different settings.

[5] In the case of Finkelstein et al. (2012), it is plausible that there are multiple important determinants of hospital utilization in the affected population (potential covariates), other than winning the lottery offered in this experiment, such

Table 1. Examples of RCTs by intervention and region.

Topic: education		
Flip charts	Kenya	Duflo et al. (2011); Glewwe et al. (2009); Duflo et al. (2015); Kremer et al. (2009); Miguel and Kremer (2004)
Class size	Kenya	Duflo et al. (2015)
Girls' scholarships	Kenya	Willa et al. (2016)
Teacher incentives	Kenya	Glewwe et al. (2010)
Tracking	Kenya	Duflo et al. (2011)
Deworming	Kenya	Miguel and Kremer (2004)
Textbooks	Kenya	Glewwe et al. (2009)
School meals	Kenya	Vermeersch and Kremer (2004)
Topic: micro-finance		
Access to microcredit	India	Banerjee et al. (2015)
	Ethiopia	Tarozzi et al. (2015)
	Mongolia	Attanasio et al. (2015)
	Morocco	Crépon et al. (2015)
	Mexico	Angelucci et al. (2015)
	Bosnia and Herzegovina	Augsburg et al. (2015)

The researcher faces a fixed budget for implementing the survey for the RCT. Given this budget, she chooses the survey's sample size and set of covariates to optimize the resulting treatment effect estimator's precision and/or the corresponding t-test's power.[6] The trade-offs involved in this choice involve basic economic reasoning. For each possible covariate, one should be comparing the marginal benefit and marginal cost of including it in the survey, which, in turn, depends on all the other covariates included in the survey. As we discuss below, in simple settings it is possible to derive analytic and intuitive solutions to this problem. Although these are insightful, they only apply in unrealistic formulations of the problem.

In general, it is necessary to consider all possible combinations of covariates and sample sizes and then to check which combination optimizes the treatment effect estimator's precision and/or the corresponding t-test's power. This is a computationally difficult combinatorial optimization problem, particularly so when there are a large number of potential covariates to choose from. Our approach is first to show that the optimization problem can be rewritten as the problem of optimally predicting the outcome by the covariates subject to the budget constraint. We assume that the treatment is randomly assigned. Two aspects of the equivalent prediction problem are crucial: first, it does not depend on the treatment allocation, so pre-experimental data on outcomes and covariates suffice to find the optimal combination of covariates and sample size;[7] second,

as education, income, past hospital utilization, or distance to hospitals. It is possible that this additional information exists in other administrative records, which, at a cost, can be assembled and linked to the original records used in Finkelstein et al. (2012). In order to understand which of these records would be most useful to collect for the purposes of this study, one can rely on a large public health literature on the determinants of hospital utilization.

[6] This choice takes place before the implementation of the RCT and could, for example, be part of a pre-analysis plan in which, among other things, the researcher specifies outcomes of interest, covariates to be selected, and econometric techniques to be used.

[7] For this purpose, we rely on the homoskedasticity assumption that requires the residual variance to be the same across the treatment and control group. This assumption allows us to optimally choose covariates and the sample size. The homoskedasticity assumption is not needed to carry out valid inference with collected experimental data.

prediction problems can easily be solved by existing machine learning techniques, making the implementation of our approach practically attractive.

To illustrate the application of our method we examine two recent experiments for which we have detailed knowledge of the process and costs of data collection. We ask two questions. First, if there is a single hypothesis one wants to test in the experiment, concerning the impact of the experimental treatment on one outcome of interest, what is the optimal combination of covariate selection and sample size given by our method, and how much of an improvement in the precision of the impact estimate can we obtain as a result? Second, what would be minimum costs of obtaining the same precision of the treatment effect as in the actual experiment, if one were to select covariates and sample size optimally (what we call the 'equivalent budget')? Analogously, by considering alternative hypothetical cost functions or regression coefficients, we examine how inexpensive or how predictive of the outcome a particular covariate would need to be for it to be worth collecting.

We find from these two applications that by adopting optimal data collection rules, not only can we achieve substantial increases in the precision of the estimates (statistical importance) for a given budget, but we can also accomplish sizeable reductions in the equivalent budget (economic importance). To illustrate the quantitative importance of the latter, we show that the optimal selection of the set of covariates and the sample size leads to a reduction of about 45% (up to 58%) of the original budget in the first (second) example we consider, while maintaining the same level of statistical significance as in the original experiment.

Although this paper focuses on the important case of RCTs with complete randomization, our procedure can be extended to many other data collection efforts and other modes of randomization. One important extension we discuss in Section 6 is that to treatment assignment through re-randomization or stratification.

There is a large and important body of literature on the design of experiments, starting with Fisher (1925, 1935). There also exists an extensive body of literature on sample size (power) calculations; see, for example, McConnell and Vera-Hernandez (2015) for a practical guide. Both bodies of literature are concerned with the precision of treatment effect estimates, but neither addresses the problem that concerns us. For instance, McConnell and Vera-Hernandez (2015) have developed methods to choose the sample size when cost constraints are binding, but they do not consider the issue of collecting covariates, nor its trade-off with selecting the sample size.

In fact, to the best of our knowledge, no paper in the literature directly considers our data collection problem. Some papers address related but very different problems (see Hahn et al., 2011; List et al., 2011; Bhattacharya and Dupas, 2012; McKenzie, 2012; Dominitz and Manski, 2017). They study some issues of data measurement, budget allocation or efficient estimation; however, they do not consider the simultaneous selection of the sample size and covariates for the RCTs as in this paper. Because our problem is distinct from the problems studied in these papers, we give a detailed comparison between our paper and the aforementioned papers in Section 7.

More broadly, this paper is related to a recent emerging literature in economics that emphasizes the importance of micro-level predictions and the usefulness of machine learning for that purpose. For example, Kleinberg et al. (2015) argue that prediction problems are abundant in economic policy analysis, and recent advances in machine learning can be used to tackle those problems. Furthermore, our paper is related to the contemporaneous debates on pre-analysis plans which demand, for example, the selection of sample sizes and covariates before the implementation of an RCT; see, for example, Coffman and Niederle (2015) and Olken (2015) for the advantages and limitations of the pre-analysis plans.

Table 2. Impact of incentives to learn on test scores with and without controls for lagged test scores (from Kremer et al., 2009).

	Avg. school controls		Individual controls	
	Without	With	Without	With
Treatment effect	0.18	0.15^{***}	0.19	0.12
	(0.12)	(0.06)	(0.14)	(0.09)

* Significant at 10%, ** Significant at 5%, *** Significant at 1%

Note: Robust standard errors are in parentheses.

Table 3. Costs and benefits for different alternatives.

	Alternative 1	Alternative 2
Number of surveys	1	2
Number of covariates	0	1
Budget	B	B
Cost per observation	λ	2λ
Sample size	$n_1 = \frac{B}{\lambda}$	$n_2 = \frac{B}{2\lambda}$
$AVar(\hat{\beta})$	$\frac{\sigma_V^2}{n_1 Var(D)}$	$\frac{\sigma_U^2}{n_2 Var(D)}$

Table 4. Gains in precision and cost from optimal covariate choice.

R_{yx}	$\sqrt{\frac{AVar(\hat{\beta}_1)}{AVar(\hat{\beta}_2)}}$	Percentage gain	$\frac{B_1}{B_2}$	Percentage gain
0.45	0.95	5	0.91	9
0.25	0.82	18	0.67	33
0.10	0.75	25	0.56	44
0.05	0.73	27	0.53	47

The remainder of the paper is organized as follows. In Section 2 we present the simplest version of our data collection problem, which illustrates the main issues discussed in this paper. A more general description of the problem is presented in Section 3. In Section 4, we discuss the costs of data collection in experiments. In Section 5, we present two empirical applications; in Section 6, we discuss some of the conceptual and practical properties of our proposed method; in Section 7, we discuss the existing related literature; and in Section 8, we give concluding remarks. In Appendix A, we describe an orthogonal greedy algorithm (OGA) that is used in our procedure; and in Appendix B, we show that this algorithm possesses desirable theoretical properties. The proofs for theoretical results are given in Appendix C. Online appendices provide details that are omitted from the main text.

2. A STYLIZED SPECIAL CASE

Consider the case in which a researcher is designing the survey for an RCT, with a limited budget, B. Her goal is to obtain a precise estimate of the average treatment effect. Take the simplest version of the problem, in which the researcher faces a choice between only two scenarios: (i) collect a single variable, the outcome used to measure the treatment effect, on a sample of size n_1; or (ii) collect two variables, the outcome and a single covariate, on a sample of size n_2, with $n_1 > n_2$. A typically useful covariate is the pre-intervention outcome, because it is often highly predictive of the post-intervention value of the outcome, but we could consider any other type of predictor.

To illustrate, take Table 4 in Kremer et al. (2009), part of which we reproduce in Table 2. This table compares estimates of the effect of a merit scholarship on test scores, in models with and without controls. The first two columns compare estimates with and without school-average lagged test score, while the third and fourth columns compare estimates with and without individual lagged test scores. Introducing the covariate has a large impact on the standard errors, which decline from 0.12 to 0.06 when we use school-level controls, and from 0.14 to 0.09 when we use individual-level controls (the difference between the first and the third column is due to small differences in sample size).[8]

Suppose the researcher is interested in collecting data for an evaluation of a new education intervention in a similar population. She is deciding between collecting only the post-intervention test score on a large sample (Case (i)), or collecting both pre- and post-intervention scores on a smaller sample (Case (ii)). These two cases result in two specifications for estimating the average treatment effect β:

$$Y = \alpha_1 + \beta D + V, \tag{2.1}$$

$$Y = \alpha_2 + \beta D + \gamma_2 X + U, \tag{2.2}$$

where Y and X are post- and pre-intervention test scores, respectively, and D is the treatment indicator, which is randomly assigned to individuals. V and U are mean zero homoskedastic residuals, with variances equal to σ_V^2 and σ_U^2, respectively. Assuming U is uncorrelated with X, we have $\sigma_V^2 = \sigma_U^2 + \gamma_2^2 \sigma_X^2$, where σ_X^2 is the variance of pre-intervention test scores.

Let λ be the cost per wave per student of the survey (so the cost of collecting two waves is 2λ per student). $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the ordinary least squares (OLS) estimators of β in (2.1) and (2.2), and $AVar(\hat{\beta})$ is the asymptotic variance of $\hat{\beta}$ divided by the sample size. Table 3 shows the main quantities influencing the researcher's decision.

Suppose the researcher measures the precision of the treatment effect estimators by their asymptotic variances[9] relative to the sample size and, thus, decides to collect the baseline covariate with a small sample size if it leads to a more precise estimator: $AVar(\hat{\beta}_1) > AVar(\hat{\beta}_2)$, or equivalently, $\frac{n_2}{n_1} > 1 - R_{yx}$, where $R_{yx} = \frac{\gamma_2^2 \sigma_X^2}{\sigma_V^2}$ is the (population) R-squared of a regression of Y on X (using data from only the treatment or only the control group). In this simple case $\frac{n_2}{n_1} = \frac{1}{2}$, so one decides to collect the covariate and to choose a smaller sample size only if $R_{yx} > 0.5$.

Going back to the example of Table 2, suppose we start with columns 3 and 4. The standard errors of the estimated treatment effect decline from 0.14 to 0.09 (roughly a 35% decline) when one controls for the lagged test score of each individual. This means that $\sigma_U^2 = \left(\frac{0.09}{0.14}\right)^2 \sigma_V^2$, so R_{yx}

[8] There are also changes in the point estimates but we abstract from those now. We also abstract from the fact that standard errors account for correlation between students in the same school, although it is possible that the inclusion of lagged test scores also helps absorb part of this correlation.

[9] In Section 3, we consider the finite-sample mean-squared error as a criterion, but since the estimator we consider in that section is unbiased, its variance is equal to the mean-squared error. In Online Appendix S1, we justify this criterion in a decision-theoretic framework.

≈ 0.59, so the decision is to collect the pre-intervention covariate and to choose a smaller sample size. For columns 1 and 2, the implied R_{yx} is even larger.

The decision would be different if R_{yx} were below 50%, or if the cost of collecting the pre-intervention outcome (λ_{pre}) were at least 44% higher than the cost of collecting the post-intervention outcome (λ_{post}).[10] This latter case is not realistic in the example we consider, since there is no reason why collecting a variable before the intervention would cost so much more than collecting the same variable after the intervention. However, it is more plausible in cases where X and Y are not the same variable (whether or not they are collected in the same survey wave).

Within this framework, one can easily evaluate the statistical and economic gains from choosing the optimal combination of covariates and sample size. Imagine, for example, that $R_{yx} \approx 0.45$ instead of $R_{yx} \approx 0.59$ as above, and assume $\lambda_{pre} = \lambda_{post} = \lambda$. In this case, choosing to collect both the pre- and the post-intervention outcome with a small sample size (n_2) is not optimal. We can evaluate the gains of moving from this suboptimal choice to the optimal one (collecting only the post-intervention outcome and choosing the larger sample size n_1) by answering the following two questions:

1. **Statistical gains**: Keeping the budget fixed, how much would $AVar(\hat{\beta})$ fall if we collected only the post-intervention outcome, with a larger sample size (n_1)?
2. **Economic gains**: Keeping $AVar(\hat{\beta})$ fixed, how much would the budget fall (from $B_2 = 2\lambda n_2$ to $B_1 = \lambda n_1$) if we collected only the post-intervention outcome, with a larger sample size?

In response to the first question, notice that $\sigma_U^2 = (1 - 0.45)\sigma_V^2$. In addition, the budget being fixed implies that $n_2 = \frac{n_1}{2}$. Then $\sqrt{\frac{AVar(\hat{\beta}_1)}{AVar(\hat{\beta}_2)}} = \sqrt{\frac{n_2\sigma_V^2}{n_1\sigma_U^2}} \approx 0.95$: the standard error of the estimated treatment effect would fall by 5 percentage points.

In response to the second question, notice that $AVar(\hat{\beta}_1) = AVar(\hat{\beta}_2)$ implies that $\frac{n_1}{n_2} = \frac{\sigma_V^2}{\sigma_U^2} = \frac{1}{1-0.45}$. Then, $\frac{B_1}{B_2} = \frac{n_1}{2n_2} = 0.91$: the cost of the survey would fall by 9 percentage points. Table 4 shows the gains in precision and costs for other typical R_{yx} values.[11] As R_{xy} decreases, both the statistical and the economic gains from choosing the optimal covariates and sample size increase. For example, if $R_{yx} = 0.10$, then the statistical gains in precision of the treatment effect estimator are 25% and the costs of data collection decrease by 44%. In our empirical applications in Section 5, we will find possible gains that are even larger.

Although this is a useful example, reality is more complex. In general, there are many potential covariates that can be collected (which typically are not uncorrelated), and the cost of data collection may be a complicated function of sample size and the set of chosen covariates (e.g., containing fixed costs, heterogeneous prices across covariates, components that depend on the

[10] In order to see this notice that, in this case, $n_1 = \frac{B}{\lambda_{post}}$ and $n_2 = \frac{B}{\lambda_{pre}+\lambda_{post}}$, i.e., $\frac{n_2}{n_1} = \frac{\lambda_{post}}{\lambda_{pre}+\lambda_{post}}$. The researcher chooses to collect post-intervention test scores if n_2 is larger than $\left(\frac{0.09}{0.14}\right)^2 n_1 \approx 0.41 n_1$, i.e., $\frac{\lambda_{post}}{\lambda_{pre}+\lambda_{post}} > 0.41$, or $\lambda_{pre} < \frac{0.59}{0.41}\lambda_{post} \approx 1.44\lambda_{post}$.

[11] If we measured the gains in precision in terms of variances instead of standard deviations we would look at $\frac{AVar(\hat{\beta}_1)}{AVar(\hat{\beta}_2)}$ instead of $\sqrt{\frac{AVar(\hat{\beta}_1)}{AVar(\hat{\beta}_2)}}$. In that case these ratios would be the following in each of the four cases we consider for R_{xy}: 0.91, 0.67, 0.56, and 0.53. Notice that these are exactly the same as the budget gains we obtain in each case, keeping precision constant. Although this exact correspondence is not true in the more general framing of the problem discussed in the next section, it is interesting that it is not far from the truth in the two empirical applications presented below.

size of the survey, etc.). Typically it is not possible to derive analytical decision rules as in the stylized example presented here. The next sections present the general formulation and our proposed solution.

3. GENERAL DATA COLLECTION PROBLEM

Suppose we are planning an RCT in which we randomly assign individuals to either a treatment ($D = 1$) or a control group ($D = 0$) with corresponding potential outcomes Y_1 and Y_0, respectively. After administering the treatment to the treatment group, we collect data on outcomes Y for both groups so that $Y = DY_1 + (1 - D)Y_0$. We also conduct a survey to collect data on a potentially very high-dimensional vector of covariates Z (e.g., from a household survey covering demographics, social background, income etc.) that predicts potential outcomes. These covariates are a subset of the universe of predictors of potential outcomes, denoted by X. Random assignment of D means that D is independent of potential outcomes and of X.

Our goal is to estimate the average treatment effect $\beta_0 := E[Y_1 - Y_0]$ as precisely as possible. In the previous section, for the sake of simplicity, we measured precision by the asymptotic variance of the OLS estimator. In this section, we consider the arguably more relevant measure of the finite-sample mean-squared error (MSE) of the treatment effect estimator. Since our estimator is unbiased, the finite-sample MSE corresponds to the estimator's finite-sample variance.

Instead of simply regressing Y on D, we want to make use of the available covariates Z to improve the precision of the resulting treatment effect estimator. Therefore, we consider estimating β_0 in the regression

$$Y = \alpha_0 + \beta_0 D + \gamma_0' Z + U, \tag{3.1}$$

where $(\alpha_0, \beta_0, \gamma_0')'$ is a vector of parameters to be estimated and U is an error term. The implementation of the RCT requires us to make two decisions that may have a significant impact on the estimation of and inference on the average treatment effect:

1. Which covariates Z should we select from the universe of potential predictors X?
2. From how many individuals (n) should we collect data on (Y, D, Z)?

Obviously, a large experimental sample size n reduces the variance of the treatment effect estimator. Similarly, collecting more covariates, in particular strong predictors of potential outcomes, reduces the variance of the residual U, which, in turn, also improves the variance of the estimator. At the same time, collecting data from more individuals and on more covariates is costly so that, given a finite budget, we want to find a combination of sample size n and covariate selection Z that leads to the most precise treatment effect estimator possible.

In this section, we propose a procedure to make this choice based on a pre-experimental data set on Y and X, such as a pilot study or a census from which we plan to draw the RCT sample. The combined data collection and estimation procedure can be summarized as follows.

1. Obtain pre-experimental data $\mathcal{S}_{\text{pre}}$ on (Y, X).
2. Use data in $\mathcal{S}_{\text{pre}}$ to select the covariates Z and sample size n.
3. Collect the experimental data $\mathcal{S}_{\text{exp}}$ on (Y, D, Z).
4. Estimate the average treatment effect using $\mathcal{S}_{\text{exp}}$. Compute standard errors.

We now describe the five steps listed above in more detail. The main component of our procedure consists of a proposal for the optimal choice of n and Z in Step 2.

3.1. Step 1: Obtain pre-experimental data

We assume the availability of data on outcomes $Y \in \mathbb{R}$ and covariates $X \in \mathbb{R}^M$ for the population from which we plan to draw the experimental data. We denote the pre-experimental sample of size N by $\mathcal{S}_{\text{pre}} := \{Y_i, X_i\}_{i=1}^N$. Our framework allows the number of potential covariates, M, to be very large (possibly much larger than the sample size N). In many contexts, such data could come from similar, existing studies. The Introduction provided several examples and showed that the availability of such datasets is much more common than might appear to be the case at first glance. Other possible candidates for pre-experimental samples are census data, household surveys, or a pilot experiment that was carried out before the larger-scale roll-out of the main experiment.

3.2. Step 2: Optimal selection of covariates and sample size

We want to use the pre-experimental data to choose the sample size, and which covariates should be in our survey. Let $S \in \{0, 1\}^M$ be a vector of ones and zeros of the same dimension as X. We say that the jth covariate ($X^{(j)}$) is selected if $S_j = 1$, and denote by X_S the subvector of X containing elements that are selected by S. For example, consider $X = (X^{(1)}, X^{(2)}, X^{(3)})$ and $S = (1, 0, 1)$. Then $X_S = (X^{(1)}, X^{(3)})$. For any vector of coefficients $\gamma \in \mathbb{R}^M$, let $\mathcal{I}(\gamma) \in \{0, 1\}^M$ denote the nonzero elements of γ such that 1 and 0, respectively, denote nonzero and zero elements. In addition, let $\gamma_{\mathcal{I}(\gamma)}$ denote the sub-vector of γ such that only nonzero elements of γ are included. For instance, if $\gamma = (1, 0.5, 0)$, then $\mathcal{I}(\gamma) = (1, 1, 0)$, $X_{\mathcal{I}(\gamma)} = (X^{(1)}, X^{(2)})$ and $\gamma_{\mathcal{I}(\gamma)} = (1, 0.5)$. Define $Y(\gamma) := Y - \gamma'_{\mathcal{I}(\gamma)} X_{\mathcal{I}(\gamma)}$. We can then rewrite (3.1) as

$$Y(\gamma) = \alpha_0 + \beta_0 D + U(\gamma), \tag{3.2}$$

where $\gamma \in \mathbb{R}^M$ and $U(\gamma) := Y - \alpha_0 - \beta_0 D - \gamma'_{\mathcal{I}(\gamma)} X_{\mathcal{I}(\gamma)}$. For a given γ and sample size n, we denote by $\hat{\beta}(\gamma, n)$ the OLS estimator of β_0 in a regression of $Y(\gamma)$ on a constant and D, using a random sample $\{Y_i, D_i, X_i\}_{i=1}^n$. We also consider the two-sided[12] t-test of

$$H_0 : \ \beta_0 = 0 \qquad \text{vs.} \qquad H_1 : \ \beta_0 \neq 0$$

using the t-statistic

$$\hat{t}(\gamma, n) := \frac{\hat{\beta}(\gamma, n)}{\sigma(\gamma)/\sqrt{n \bar{D}_n (1 - \bar{D}_n)}},$$

where $\sigma^2(\gamma) := Var(U(\gamma))$ is the residual variance and $\bar{D}_n := \sum_{i=1}^n D_i/n$ is the number of individuals in the treatment group divided by the sample size n.

Data collection is costly and therefore constrained by a budget of the form $c(S, n) \leq B$, where $c(S, n)$ are the costs of collecting the variables given by selection S from n individuals, and B is the researcher's budget.

We assume that the researcher is interested in collecting data so as to ensure good statistical properties of the resulting treatment effect estimator and the corresponding t-test. We consider

[12] The same arguments in this paper straightforwardly carry over to a one-sided t-test.

two criteria: the MSE of $\hat{\beta}(\gamma, n)$ and the power of the t-test that employs $\hat{t}(\gamma, n)$. This objective can be rationalized in a decision-theoretic framework as shown in Online Appendix S1.

We now briefly argue that minimizing the MSE of $\hat{\beta}(\gamma, n)$ and maximizing the power of the t-test lead to equivalent optimization problems for selecting the optimal collection of covariates and sample size. The key idea is to find the optimal combination (γ, n) by rewriting the minimization of MSE and the maximization of power equivalently as a prediction problem of predicting outcomes Y by covariates X. These kinds of prediction problems can easily be solved with modern machine learning techniques. Crucially, the prediction problem does not involve D, so pre-experimental data on Y and X suffice to make the optimal choice of (γ, n) for the experiment.

First, consider choosing the experimental sample size n and the covariate selection S so as to minimize the finite-sample MSE of $\hat{\beta}(\gamma, n)$, i.e., we want to choose n and γ to minimize

$$MSE\left(\hat{\beta}(\gamma, n) \,\middle|\, D_1, \ldots, D_n\right) := E\left[\left(\hat{\beta}(\gamma, n) - \beta_0\right)^2 \,\middle|\, D_1, \ldots, D_n\right]$$

subject to the budget constraint.

Assumption 3.1 *(i) $\{(Y_i, X_i, D_i)\}_{i=1}^n$ is an i.i.d. sample from the distribution of (Y, X, D) such that D is completely randomized. (ii) Var(U(γ)|D = 1) = Var(U(γ)|D = 0) for all $\gamma \in \mathbb{R}^M$.*

Part (i) of this assumption states that D is randomly assigned. Part (ii) is a homoskedasticity assumption that requires the residual variance to be the same across the treatment and control group. This assumption is satisfied, for example, when the treatment effect is constant across individuals in the experiment. Similarly to in any type of power calculation for experiments, the random assignment of D together with a homoskedastic variance allows us to express the MSE of the estimator and power of the t-test in ways that do not depend on the actual treatment assignment. This is the key to being able to choose covariates and the sample size *before seeing the experimental data*. In Section 6, we explain possible extensions to other forms of randomization and deviations from the homoskedastic variance.

Denote by c_α and $\Phi(\cdot)$ the α-quantile and cumulative distribution function of the standard normal distribution, respectively. The following lemma characterizes the finite-sample MSE of the estimator and the power of the t-test under the above assumption.

Lemma 3.1 *Suppose Assumption 3.1 holds. Then, for any $\gamma \in \mathbb{R}^M$,*

$$MSE\left(\hat{\beta}(\gamma, n) \,\middle|\, D_1, \ldots, D_n\right) = \frac{\sigma^2(\gamma)}{n\bar{D}_n(1 - \bar{D}_n)}. \tag{3.3}$$

If, in addition, (Y, X) are jointly normal, then, for any $\alpha \in (0, 1)$, $\beta \neq 0$, and $\gamma \in \mathbb{R}^M$,

$$\begin{aligned} &P_\beta\left(\left|\hat{t}(\gamma, n)\right| > c_{1-\alpha/2} \,\middle|\, D_1, \ldots, D_n\right) \\ &\quad = 1 + \Phi\left(\frac{\beta}{\sigma(\gamma)/\sqrt{n\bar{D}_n(1 - \bar{D}_n)}} - c_{1-\alpha/2}\right) - \Phi\left(\frac{\beta}{\sigma(\gamma)/\sqrt{n\bar{D}_n(1 - \bar{D}_n)}} + c_{1-\alpha/2}\right), \end{aligned}$$

where P_β denotes probabilities under the assumption that β is the true coefficient in front of D. Furthermore, $P_\beta(|\hat{t}(\gamma, n)| > c_{1-\alpha/2}|D_1, \ldots, D_n)$ is decreasing in the term $\sigma(\gamma)/\sqrt{n\bar{D}_n(1 - \bar{D}_n)}$, which is the square-root of the MSE.

The proof of this Lemma can be found in Appendix C. Note that for each (γ, n), the MSE is minimized by the equal splitting between the treatment and control groups. Hence, suppose that the treatment and control groups are of exactly the same size (i.e., $\bar{D}_n = 0.5$). By Lemma 3.1, minimizing the MSE of the treatment effect estimator subject to the budget constraint,

$$\min_{n\in\mathbb{N}_+,\, \gamma\in\mathbb{R}^M} MSE\left(\hat{\beta}(\gamma, n)\,\middle|\, D_1, \ldots, D_n\right) \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma), n) \leq B, \tag{3.4}$$

is equivalent to minimizing the residual variance $\sigma^2(\gamma)$, divided by the sample size,

$$\min_{n\in\mathbb{N}_+,\, \gamma\in\mathbb{R}^M} \frac{\sigma^2(\gamma)}{n} \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma), n) \leq B. \tag{3.5}$$

Here, for a given γ, $c(\mathcal{I}(\gamma), n)$ are the costs of collecting the variables whose regression coefficients (γ) are nonzero from n individuals, and B is the researcher's budget. Note that by the homoskedastic error assumption,

$$\begin{aligned}\sigma^2(\gamma) = Var(U(\gamma)) &= \text{Var}\left(Y - \alpha_0 - \beta_0 D - \gamma'_{\mathcal{I}(\gamma)} X_{\mathcal{I}(\gamma)}\,\middle|\, D = 0\right)\\ &= \text{Var}\left(Y - \gamma'_{\mathcal{I}(\gamma)} X_{\mathcal{I}(\gamma)}\,\middle|\, D = 0\right) = \text{Var}\left(Y - \gamma' X\,\middle|\, D = 0\right),\end{aligned}$$

which equals the residual variance in a regression of Y on X.

Now, consider choosing the experimental sample size n and the covariate selection S so as to maximize power of the two-sided t-test based on $\hat{t}(\gamma, n)$. Lemma 3.1 shows that, under the normality assumption and for any alternative $\beta \neq 0$ and size α, the power of the two-sided t-test is a decreasing transformation of $\frac{\sigma^2(\gamma)}{n\bar{D}_n(1-\bar{D}_n)}$. Therefore, assigning as many individuals to the treatment as to the control group, besides minimizing the MSE above also maximizes power. Therefore, assuming again $\bar{D}_n = 0.5$, maximizing power subject to the budget constraint,

$$\max_{n\in\mathbb{N}_+,\, \gamma\in\mathbb{R}^M} P_\beta\left(\left|\hat{t}(\gamma, n)\right| > c_{1-\alpha/2}\,\middle|\, D_1, \ldots, D_n\right) \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma), n) \leq B,$$

is also equivalent to minimizing the residual variance in a regression of Y on X, divided by the sample size, as in (3.7). Notice that even when (Y, X) are not jointly normal, the power expression in Lemma 3.1 may be approximately correct because the Berry–Esseen bound guarantees that the t-statistic's distribution is close to normal as long as n is not too small.

Having motivated the optimization problem in (3.7) in terms of minimization of the MSE of the treatment effect estimator as well as in terms of maximization of power of the corresponding t-test, we now discuss how to approximate the solution to (3.7) in a given finite sample.

Importantly, notice that the optimization problem (3.7) depends on the data only through the residual variance $\sigma^2(\gamma)$, which, under Assumption 3.1, can be estimated before the randomization takes place, i.e., using the pre-experimental sample $\mathcal{S}_{\text{pre}}$. Therefore, employing the standard sample variance estimator of $\sigma^2(\gamma)$, the sample counterpart of our population optimization problem (3.7) is

$$\min_{n\in\mathbb{N}_+,\, \gamma\in\mathbb{R}^M} \frac{1}{nN}\sum_{i=1}^{N}(Y_i - \gamma' X_i)^2 \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma), n) \leq B. \tag{3.6}$$

The problem (3.8), which is based on the pre-experimental sample, approximates the population problem (3.7) for the experiment if the second moments in the pre-experimental sample are close to the second moments in the experiment (which holds, for example, if the population in the pre-experimental sample is the same as the population in the experiment).

The important feature of (3.8) is that it is a prediction problem of Y by X, subject to the budget constraint. Since the budget constraint depends on the nonzero elements of γ and the cost function can be highly nonlinear, it is computationally challenging to obtain an exact solution to (3.8). Instead, we obtain an approximate solution using modern machine learning techniques (e.g., LASSO—least absolute shrinkage and selection operator, or OGA—orthogonal greedy algorithm) that are readily available. The LASSO is well known in the economics literature, whereas the OGA is much less known. In Appendix A, we therefore describe in more detail an OGA that is adapted to the data collection problem. In Appendix B, we show that it possesses desirable theoretical properties in the following sense. We derive the finite-sample bound on the MSE of the average treatment effect estimator resulting from the OGA method. The natural target for this MSE is an infeasible MSE when γ_0 is known a priori. Theorem C.2 establishes conditions under which the difference between the MSE resulting from our method and the infeasible MSE decreases at a rate of $1/k$ as k increases, where k is the number of the steps in the OGA. It is known in a simpler setting than ours that this rate $1/k$ cannot generally be improved (see, e.g., Barron et al., 2008). In Online Appendix S4, we provide simulations to show that the OGA works well in finite samples.

The two machine learning methods, LASSO and OGA, are complementary to each other; no one method dominates the other in simulations, and both are well motivated computationally and theoretically in the machine learning literature. One practical advantage of using the OGA relative to the LASSO is that the former can select among overlapping groups of covariates, unlike the latter. Denote by $(\hat{n}, \hat{\gamma})$ the machine learning algorithm's solution to (3.8) and let $\hat{\mathcal{I}} := \mathcal{I}(\hat{\gamma})$ denote the selected covariates.[13]

3.3. Step 3: Experiment and data collection

Given the optimal selection of covariates $\hat{\mathcal{I}}$ and sample size $\hat{n}$, we collect the covariates $Z := X_{\hat{\mathcal{I}}}$ from individuals in the experimental sample, randomly assign $\hat{n}$ individuals to either the treatment or the control group (with equal probability), and then collect the outcome Y from them. This yields the experimental data $\mathcal{S}_{\text{exp}} := \{Y_i, D_i, Z_i\}_{i=1}^{\hat{n}}$ from $(Y, D, X_{\hat{\mathcal{I}}})$.

3.4. Step 4: Estimation of the average treatment effect

We regress $Y_i - \hat{\gamma} Z_i$ on $(1, D_i)$ using the experimental sample $\mathcal{S}_{\text{exp}}$, where $\hat{\gamma}$ is obtained in Step 2 with the pre-experimental data $\mathcal{S}_{\text{pre}}$. The OLS estimator of the coefficient on D_i is the average treatment effect estimator $\hat{\beta}$. It is important to note that this estimator is different from that of running an OLS regression of Y onto a constant, D, and Z.

3.5. Step 5: Computation of standard errors

Assuming the two samples $\mathcal{S}_{\text{pre}}$ and $\mathcal{S}_{\text{exp}}$ are independent, and that treatment is randomly assigned, the presence of the covariate selection in Step 2 does not affect the asymptotic validity of the standard errors that one would use in the absence of Step 2. Therefore, asymptotically valid standard errors of $\hat{\beta}$ can be computed in the usual fashion (see, e.g., Imbens and Rubin, 2015).

[13] In the paper, we are mainly concerned with the situation where the sample size N in the pre-experimental sample is large and the budget B is small, so that the possibility of overfitting is of secondary concern. One may consider a penalized version of (3.8) if N is relatively small compared with B. See Appendix B for a detailed discussion on this issue.

4. THE COSTS OF DATA COLLECTION

Our proposed data collection procedure could employ any researcher-chosen cost function $c(S, n)$ that defines the budget constraint of the researcher. However, in this section, we propose a particular specification that we believe captures the typical costs incurred in RCTs in economics and is the one we use in the empirical part of Section 5.

In principle, it is possible to construct a matrix containing the value of the costs of data collection for every possible combination of S and n without assuming any particular form of relationship between the individual entries. However, determination of the costs for every possible combination of S and n is a cumbersome and, in practice, probably infeasible exercise. Therefore, we consider the specification of cost functions that capture the costs of all stages of the data collection process in a more parsimonious fashion.

We propose to decompose the overall costs of data collection into three components: administration costs $c_{\text{admin}}(S)$, training costs $c_{\text{train}}(S, n)$, and interview costs $c_{\text{interv}}(S, n)$, so that

$$c(S, n) = c_{\text{admin}}(S) + c_{\text{train}}(S, n) + c_{\text{interv}}(S, n). \tag{4.1}$$

In the remainder of this section, we discuss possible specifications of the three types of costs by considering fixed and variable cost components corresponding to the different stages of the data collection process. The exact functional form assumptions are based on the researcher's knowledge about the operational details of the survey process. Even though this section's general discussion is driven by our experience in the empirical applications of Section 5, the operational details are likely to be similar for many surveys, so we expect the following discussion to provide a useful starting point for other data collection projects.

We start by specifying survey time costs. Let τ_j, $j = 1, \ldots, M$, be the costs of collecting variable j for one individual, measured in units of survey time. Similarly, let τ_0 denote the costs of collecting the outcome variable, measured in units of survey time. Then, the total time costs of surveying one individual to elicit the variables indicated by S are

$$T(S) := \tau_0 + \sum_{j=1}^{M} \tau_j S_j.$$

4.1. Administration and training costs

A data collection process typically incurs costs due to administrative work and training prior to the start of the actual survey. Examples of such tasks are developing the questionnaire and the program for data entry, piloting the questionnaire, developing the manual for administration of the survey, and organizing the training required for the enumerators.

Fixed costs, which depend neither on the size of the survey nor on the sample size of survey participants, can simply be subtracted from the budget. We assume that B is already net of such fixed costs.

Most administrative and training costs tend to vary with the size of the questionnaire and the number of survey participants. Administrative tasks such as development of the questionnaire, data entry, and training protocols are independent of the number of survey participants, but depend on the size of the questionnaire (measured by the number of positive entries in S), as smaller questionnaires are less expensive to prepare than larger ones. We model those costs by

$$c_{\text{admin}}(S) := \phi T(S)^{\alpha}, \tag{4.2}$$

where ϕ and α are scalars to be chosen by the researcher. We assume $0 < \alpha < 1$, which means that marginal costs are positive but decline with survey size.

Training of the enumerators depends on the survey size, because a longer survey requires more training, and on the number of survey participants, because surveying more individuals usually requires more enumerators (which, in turn, may raise the costs of training), especially when there are limits on the duration of the fieldwork. We therefore specify training costs as

$$c_{\text{train}}(S, n) := \kappa(n)\, T(S), \tag{4.3}$$

where $\kappa(n)$ is some function of the number of survey participants.[14]Training costs are typically lumpy because, for example, there exists only a limited set of room sizes one can rent for the training, so we model $\kappa(n)$ as a step function:

$$\kappa\,(n) = \begin{cases} \overline{\kappa}_1 \text{ if } \; 0 < n \leq \overline{n}_1 \\ \overline{\kappa}_2 \text{ if } \overline{n}_1 \; < n \leq \overline{n}_2 \\ \quad\;\; \vdots \end{cases} .$$

Here, $\overline{\kappa}_1, \overline{\kappa}_2, \ldots$ is a sequence of scalars describing the costs of sample sizes in the ranges defined by the cut-off sequence $\overline{n}_1, \overline{n}_2, \ldots$.

4.2. *Interview costs*

Enumerators are often paid by the number of interviews conducted, and the payment increases with the size of the questionnaire. Let η denote fixed costs per interview that are independent of the size of the questionnaire and of the number of participants. These are often due to travel costs and can account for a substantive fraction of the total interview costs. Suppose the variable component of the interview costs is linear so that total interview costs can be written as

$$c_{\text{interv}}(S, n) := n\eta + np\, T(S), \tag{4.4}$$

where $T(S)$ should now be interpreted as the average time spent per interview, and p is the average price of one unit of survey time. We employ the specification (4.9) with (4.10)–(4.12) when studying the impact of free day-care on child development in Section 5.1.

Remark 4.1 Because we always collect the outcome variable, we incur the fixed costs $n\eta$ and the variable costs $np\tau_0$ even when no covariates are collected.

Remark 4.2 Non-financial costs are difficult to model, but could in principle be added. They are primarily related to the impact of sample and survey size on data quality. For example, if we design a survey that takes more than four hours to complete, the quality of the resulting data is likely to be affected by interviewer and interviewee fatigue. Similarly, conducting the training of enumerators becomes more difficult as the survey size grows. Hiring high-quality enumerators may be particularly important in that case, which could result in even higher costs (although this latter observation could be explicitly considered in our framework).

[14] It is of course possible that κ depends not only on n but also on $T(S)$. We model it this way for simplicity, and because it is a sensible choice in the applications we discuss below.

4.3. Clusters

In many experiments, randomization is carried out at a cluster level (e.g., school level), rather than at an individual level (e.g., student level). In this case, training costs may depend not only on the ultimate sample size $n = c\, n_c$, where c and n_c denote the number of clusters and the number of participants per cluster, respectively, but also on a particular combination (c, n_c), because the number of required enumerators may be different for different (c, n_c) combinations. Therefore, training costs (which now also depend on c and n_c) may be modelled as

$$c_{\text{train}}(S, n_c, c) := \kappa(c, n_c)\, T(S). \tag{4.5}$$

The interaction of cluster and sample size in determining the number of required enumerators and, thus, the quantity $\kappa(c, n_c)$, complicates the modelling of this quantity relative to the case without clustering. Let $\mu(c, n_c)$ denote the number of required survey enumerators for c clusters of size n_c. As in the case without clustering, we assume that the training costs is lumpy in the number of enumerators used:

$$\kappa(c, n_c) := \begin{cases} \overline{\kappa}_1 \text{ if } \ 0 < \mu(c, n_c) \leq \overline{\mu}_1 \\ \overline{\kappa}_2 \text{ if } \overline{\mu}_1 < \mu(c, n_c) \leq \overline{\mu}_2 \\ \quad \vdots \end{cases}.$$

The number of enumerators required, $\mu(c, n_c)$, may also be lumpy in the number of interviewees per cluster, n_c, because there are bounds to how many interviews each enumerator can carry out. Also, the number of enumerators needed for the survey typically increases with in the number of clusters in the experiment. Therefore, we model $\mu(c, n_c)$ as

$$\mu(c, n_c) := \lfloor \mu_c(c) \cdot \mu_n(n_c) \rfloor,$$

where $\lfloor \cdot \rfloor$ denotes the integer part, $\mu_c(u) := \lambda u$ for some constant λ (i.e., $u \mapsto \mu_c(u)$ is assumed to be linear in the argument), and

$$\mu_n(n_c) := \begin{cases} \overline{\mu}_{n,1} \text{ if } \ 0 < n_c \leq \overline{n}_1 \\ \overline{\mu}_{n,2} \text{ if } \overline{n}_1 < n_c \leq \overline{n}_2 \\ \quad \vdots \end{cases}.$$

In addition, while the variable interview costs component continues to depend on the overall sample size n as in (4.12), the fixed part of the interview costs is determined by the number of clusters c rather than by n. Therefore, the total costs per interview become

$$c_{\text{interv}}(S, n_c, c) := \psi(c)\eta + c n_c p\, T(S), \tag{4.6}$$

where $\psi(c)$ is some function of the number of clusters c.[15]

4.4. Covariates with heterogeneous prices

In randomized experiments, the data collection process often differs across blocks of covariates. For example, the researcher may want to collect outcomes of psychological tests for the members

[15] One issue we have not yet explicitly addressed concerns the implications for inference of a clustered randomized design. It is well known that intra-cluster correlation in residuals has large effects on the standard errors of treatment effect estimates. It is possible that covariates contribute to changes in the MSE of treatment effect estimators not only by absorbing part of the residual variance, but also by absorbing part of the intra-cluster correlation.

of the household that is visited. These tests may need to be administered by trained psychologists, whereas administering a questionnaire about background variables such as household income, number of children, or parental education, may not require any particular set of skills or qualifications other than the training provided as part of the data collection project.

Partition the covariates into two blocks, a high-cost block (e.g., outcomes of psychological tests) and a low-cost block (e.g., standard questionnaire). Order the covariates such that the first M_{low} covariates belong to the low-cost block, and the remaining $M_{\text{high}} := M - M_{\text{low}}$ together with the outcome variable belong to the high-cost block. Let

$$T_{\text{low}}(S) := \sum_{j=1}^{M_{\text{low}}} \tau_j S_j \qquad \text{and} \qquad T_{\text{high}}(S) := \tau_0 + \sum_{j=M_{\text{low}}+1}^{M} \tau_j S_j$$

be the total time costs per individual of surveying all low-cost and high-cost covariates, respectively. Then, the total time costs for all variables can be written as $T(S) = T_{\text{low}}(S) + T_{\text{high}}(S)$.

Because we require two types of enumerators, one for the high-cost covariates and one for the low-cost covariates, the financial costs of each interview (fixed and variable) may be different for the two blocks of covariates. Denote these by $\psi_{\text{low}}(c, n_c)\eta_{\text{low}} + cn_c p_{\text{low}} T_{\text{low}}(S)$ and $\psi_{\text{high}}(c, n_c)\eta_{\text{high}} + cn_c p_{\text{high}} T_{\text{high}}(S)$, respectively.

The fixed costs for the high-cost block are incurred regardless of whether high-cost covariates are selected or not, because we always collect the outcome variable, which here is assumed to belong to this block. The fixed costs for the low-cost block, however, are incurred only when at least one low-cost covariate is selected (i.e., when $\sum_{j=1}^{M_{\text{low}}} S_j > 0$). Therefore, the total interview costs for all covariates can be written as

$$c_{\text{interv}}(S, n) := \mathbb{1}\left\{\sum_{j=1}^{M_{\text{low}}} S_j > 0\right\}(\psi_{\text{low}}(c, n_c)\eta_{\text{low}} + cn_c p_{\text{low}} T_{\text{low}}(S)) \tag{4.7}$$

$$+ \psi_{\text{high}}(c, n_c)\eta_{\text{high}} + cn_c p_{\text{high}} T_{\text{high}}(S). \tag{4.8}$$

The administration and training costs can also be assumed to differ for the two types of enumerators. In that case,

$$c_{\text{admin}}(S) := \phi_{\text{low}} T_{\text{low}}(S)^{\alpha_{\text{low}}} + \phi_{\text{high}} T_{\text{high}}(S)^{\alpha_{\text{high}}}, \tag{4.9}$$

$$c_{\text{train}}(S, n) := \kappa_{\text{low}}(c, n_c)\, T_{\text{low}}(S) + \kappa_{\text{high}}(c, n_c)\, T_{\text{high}}(S). \tag{4.10}$$

We employ specification (4.9) with (4.14)–(4.18) when, in Section 5.2, we study the impact on student learning of cash grants which are provided to schools.

5. EMPIRICAL APPLICATIONS

5.1. Access to free day-care in Rio de Janeiro

In this section, we re-examine the experimental design of Attanasio et al. (2014), who evaluate the impact of access to free day-care on child development and household resources in Rio de Janeiro. In their dataset, access to care in public day-care centres, most of which are located in slums, is allocated through a lottery, administered to children in the waiting lists for each day-care centre.

Just before the 2008 school year, children applying for a slot at a public day-care centre were put on a waiting list. At this time, children were between the ages of 0 and 3. For each centre, when the demand for day-care slots in a given age range exceeded the supply, the slots were allocated using a lottery (for that particular age range). The use of such an allocation mechanism means that we can analyse this intervention as if it were an RCT, where the offer of free day-care slots is randomly allocated across potentially eligible recipients. Attanasio et al. (2014) compare the outcomes of children and their families who were awarded a day-care slot through the lottery, with the outcomes of those not awarded a slot.

The data for the study were collected mainly during the second half of 2012, four and a half years after the randomization took place. Most children were between the ages of 5 and 8.[16] A survey was conducted, which had two components: a household questionnaire, administered to the mother or guardian of the child; and a battery of health and child development assessments, administered to children. Each household was visited by a team of two field workers, one for each component of the survey.

The child assessments took a little less than one hour to administer, and included five tests per child, plus the measurement of height and weight. The household survey took between one and a half and two hours, and included about 190 items, in addition to a long module asking about day-care history, and the administration of a vocabulary test to the main carer of each child.

As we explain below, we use the original sample, with the full set of items collected in the survey, to calibrate the cost function for this example. However, when solving the survey design problem described in this paper we consider only a subset of items of these data, with the original budget being scaled down properly. This is done for simplicity, so that we can essentially ignore the fact that some variables are missing for part of the sample, either because some items are not applicable to everyone in the sample, or because of item non-response. We organize the child assessments into three indices: cognitive tests, executive function tests, and anthropometrics (height and weight). These three indices are the main outcome variables in the analysis. However, we use only the cognitive tests and anthropometrics indices in our analysis, as we have fewer observations for executive function tests.

We consider only 40 covariates out of the total set of items on the questionnaire. The variables not included can be arranged into four groups: (a) variables that can be seen as final outcomes, such as questions about the development and the behaviour of the children in the household; (b) variables that can be seen as intermediate outcomes, such as labour supply, income, expenditure, and investments in children; (c) variables for which there is an unusually large number of missing values; and (d) variables that are either part of the day-care history module, or the vocabulary test for the child's carer (because these could have been affected by the lottery assigning children to day-care vacancies). We then drop four of the 40 covariates chosen, because their variance is zero in the sample. The remaining $M = 36$ covariates are related to the respondent's age, literacy, educational attainment, household size, safety, burglary at home, day-care, neighbourhood, characteristics of the respondent's home and its surroundings (the number of rooms, garbage collection service, water filter, stove, refrigerator, freezer, washer, TV, computer, Internet, phone, car, type of roof, public light in the street, pavement, etc.). We drop individuals for whom at least one value in each of these covariates is missing, which leads us to use a subsample with 1,330 individuals from the original experimental sample, which included 1,466 individuals.

[16] An additional wave on an expanded sample was collected in 2015, but we abstract from it here.

Calibration of the cost function

We specify the cost function (4.9) with components (4.10)–(4.12) to model the data collection procedure as implemented in Attanasio et al. (2014). We calibrate the parameters using the actual budgets for training, administrative, and interview costs in the authors' implementation. The contracted total budget of the data collection process was R$665,000.[17]

For the calibration of the cost function, we use the originally planned budget of R$665,000, and the original sample size of 1,466. As mentioned above, there were 190 variables collected in the household survey, together with a day-care module and a vocabulary test. In total, this translates into a total of roughly 240 variables.[18] Online Appendix S2 provides a detailed description of all components of the calibrated cost function.

Implementation

We studentized all covariates to have variance one. We use the OGA to solve the optimization problem in (3.8). The basic idea of the OGA (in its simplest form) is straightforward. Fix a sample size n. Start by finding the covariate that has the highest correlation with the outcome. Regress the outcome on that variable, and keep the residual. Then, among the remaining covariates, find the one that has the highest correlation with the residual. Regress the outcome onto both selected covariates, and keep the residual. Again, among the remaining covariates, find the one that has the highest correlation with the new residual, and proceed as before. We iteratively select additional covariates up to the point when the budget constraint is no longer satisfied. Finally, we repeat this search process for alternative sample sizes, and search for the combination of sample size and covariate selection that minimizes the residual variance. See Appendix A for more details.

To compare the OGA with alternative approaches, we also consider LASSO and POST-LASSO for the inner optimization problem in Step 2 of our procedure. The LASSO solves

$$\min_{\gamma} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \gamma' X_i\right)^2 + \lambda \sum_{j} |\gamma_j| \tag{5.1}$$

with a tuning parameter $\lambda > 0$ that ensures that the budget constraint is satisfied (more details below). The POST-LASSO procedure runs an OLS regression of Y_i on the selected covariates (nonzero entries of γ) in (5.19). Belloni and Chernozhukov (2013), for example, provide a detailed description of the two algorithms. It is known that LASSO yields biased regression coefficient estimates and that POST-LASSO can mitigate this bias problem. Together with the outer optimization over the sample size, using the LASSO or POST-LASSO solutions in the inner loop may lead to different selections of covariate-sample size combinations. This is because POST-LASSO re-estimates the regression equation, which may lead to more precise estimates of γ and thus result in a different estimate for the MSE of the treatment effect estimator.

In both LASSO implementations, the penalization parameter λ is chosen so as to satisfy the budget constraint as close to equality as possible. We start with a large value for λ, which leads to a large penalty for nonzero entries in γ, so that few or no covariates are selected and the budget constraint holds. Similarly, we consider a very small value for λ, which leads to the selection

[17] There were some adjustments to the budget during the period of fieldwork.

[18] The budget is for the 240 variables (or so) actually collected. In spite of that, we only use 36 of these as covariates in this paper, as the remaining variables in the survey were not so much covariates as they were measuring other intermediate and final outcomes of the experiment, as we have explained before. The actual budget used in solving the survey design problem is scaled down to match the use of only 36 covariates.

Table 5. Day-care (outcome: cognitive test).

Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	RMSE	EQB	Relative EQB
Experiment	1,330	36	1	0.025285	R$562,323	1
OGA	2,677	1	0.9939	0.018776	R$312,363	0.555
LASSO	2,762	0	0.99475	0.018789	R$313,853	0.558
POST-LASSO	2,677	1	0.9939	0.018719	R$312,363	0.555

Table 6. Day-care (outcome: health assessment).

Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	RMSE	EQB	Relative EQB
Experiment	1,330	36	1	0.025442	R$562,323	1
OGA	2,762	0	0.99475	0.018799	R$308,201	0.548
LASSO	2,762	0	0.99475	0.018799	R$308,201	0.548
POST-LASSO	2,677	1	0.9939	0.018735	R$306,557	0.545

of many covariates and violation of the budget. Then, we use a bisection algorithm to find the λ-value in this interval for which the budget is satisfied within some pre-specified tolerance.

OGA, LASSO and POST-LASSO are three alternative procedures, and all of them provide approximate solutions to the problem we study in this paper. Of the three, OGA is easier to implement and is computationally more attractive, but in theory we could use any of the three. We show below that, in the applications we present, these three methods deliver very similar solutions.

Results

Tables 5 and 6 summarize the results of the covariate selection procedures. For the cognitive test outcome, OGA and POST-LASSO select one covariate ('$\lvert\hat{I}\rvert$'),[19] whereas LASSO does not select any covariate. The selected sample sizes ('$\hat{n}$') are 2,677 for OGA and POST-LASSO, and 2,762 for LASSO, which are almost twice as large as the actual sample size in the experiment. The performance of the three covariate selection methods in terms of the precision of the resulting treatment effect estimator is measured by the square-root value of the minimized MSE criterion function ('RMSE') from Step 2 of our procedure. We focus on the MSE, but notice that gains in MSE translate into gains in the power of the corresponding t-test, as discussed in Section 3. The three methods perform similarly well and improve precision by about 25% relative to the experiment. Also, all three methods manage to exhaust the budget, as indicated by the cost-to-budget ratios ('Cost/B') close to one. We do not put any strong emphasis on the selected covariates, as the improvement of the criterion function is minimal relative to the case that no covariate is selected (i.e., the selection with LASSO). The results for the health assessment outcome are very similar to those of the cognitive test with POST-LASSO selecting one variable (the number of rooms in the house), whereas OGA and LASSO do not select any covariate.

[19] For OGA, it is an indicator variable whether the respondent has finished secondary education, which is an important predictor of outcomes; for POST-LASSO, it is the number of rooms in the house, which can be considered as a proxy for wealth of the household, and again, an important predictor of outcomes.

To assess the economic gain of having performed the covariate selection procedure after the first wave, we include the column ’EQB’ (abbreviation of ’equivalent budget’) in Tables 5 and 6. The first entry of this column in Table 5 reports the budget necessary for the selection of $\hat{n} =$ 1,330 and all covariates, as was carried out in the experiment. For the three covariate selection procedures, the column shows the budget that would have sufficed to achieve the same precision as the actual experiment in terms of the minimum value of the MSE criterion function in Step 2. For example, for the cognitive test outcome, using the OGA to select the sample size and the covariates, a budget of R$312,363 would have sufficed to achieve the experimental RMSE of 0.025285. This is a huge reduction of costs by about 45%, as shown in the last column called ’Relative EQB’. Similar reductions in costs are possible when using the LASSO procedures and also when considering the health assessment outcome.

In Online Appendix S6, we perform an out-of-sample evaluation by splitting the dataset into training samples for the covariate selection step and evaluation samples for the computation of the performance measures RMSE and EQB. The results are very similar to those in Tables 5 and 6.

Online Appendix S4 presents the results of Monte Carlo simulations that mimic this dataset, and shows that all three methods select more covariates and smaller sample sizes as we increase the predictive power of some covariates. This finding suggests that the covariates collected in the survey were not predicting the outcome very well, and, therefore, in the next wave the researcher should spend more of the available budget to collect data on more individuals, with no (or only a minimal) household survey. Alternatively, the researcher may want to redesign the household survey to include questions whose answers are likely better predictors of the outcome.

5.2. *Provision of school grants in Senegal*

In this subsection, we consider the study by Carneiro et al. (2015), who evaluate, using an RCT, the impact of school grants on student learning in Senegal. The authors collect original data not only on the treatment status of schools (treatment and control) and on student learning, but also on a variety of household, principal, and teacher characteristics that could potentially affect learning.

The dataset contains two waves, a baseline and a follow-up, which we use for the study of two different hypothetical scenarios.[20] In the first scenario, the researcher has access to a pre-experimental dataset consisting of all outcomes and covariates collected in the baseline survey of this experiment, but not the follow-up data. The researcher applies the covariate selection procedure to this pre-experimental dataset to find the optimal sample size and set of covariates for the randomized control trial to be carried out after the first wave. In the second scenario, in addition to the pre-experimental sample from the first wave the researcher now also has access to the post-experimental outcomes collected in the follow-up (second wave). In this second scenario, we treat the follow-up outcomes as the outcomes of interest and include baseline outcomes in the pool of covariates that predict follow-up outcomes.

As in the previous subsection, we calibrate the cost function based on the full dataset from the experiment, but for solving the survey design problem we focus on a subset of individuals and variables from the original questionnaire. For simplicity, we exclude all household variables from the analysis, because they were only collected for 4 out of the 12 students tested in each school, and we remove covariates whose sample variance is equal to zero. Again, for simplicity, of the four outcomes (maths test, French test, oral test, and receptive vocabulary) in the original experiment,

[20] There is also a third wave of data from which we abstract in this paper.

we only consider the first one (maths test) as our outcome variable. We drop individuals for whom at least one answer in the survey or the outcome variable is missing. This sample selection procedure leads to sample size of $N = 2{,}280$ for the baseline maths test outcome. For the second scenario discussed above, where we use also the follow-up outcome, the sample size is smaller ($N = 762$) because of non-response in the follow-up outcome and because we restrict the sample to the control group of the follow-up. In the first scenario in which we predict the baseline outcome, dropping household variables reduces the original number of covariates in the survey from 255 to $M = 142$. The remaining covariates are school- and teacher-level variables. In the second scenario in which we predict follow-up outcomes, we add the three baseline outcomes to the covariate pool, but at the same time remove two covariates because they have sample variance zero when restricted to the control group. Therefore, there are $M = 143$ covariates in the second scenario.

Calibration of the cost function

We specify the cost function (4.9) with components (4.15)–(4.18) to model the data collection procedure as implemented in Carneiro et al. (2015). Each school forms a cluster. We calibrate the parameters using the costs faced by the researchers and their actual budgets for training, administrative, and interview costs. The total budget for one wave of data collection in this experiment, excluding the costs of the household survey, was approximately $192,200.

For the calibration of the cost function, we use the original sample size, the original number of covariates in the survey (except those in the household survey), and the original number of outcomes collected at baseline. The three baseline outcomes were much more expensive to collect than the remaining covariates. In the second scenario, we therefore group the former together as high-cost variables, and all remaining covariates as low-cost variables. Online Appendix S2 provides a detailed description of all components of the calibrated cost function.

Implementation

The implementation of the covariate selection procedures is identical to the one in the previous subsection, except that we consider here two different specifications of the pre-experimental sample $\mathcal{S}_{\text{pre}}$, depending on whether the outcome of interest is the baseline or follow-up outcome.

Results

Table 7 summarizes the results of the covariate selection procedures. Panel (a) shows the results of the first scenario in which the baseline maths test is used as the outcome variable to be predicted. Panel (b) shows the corresponding results for the second scenario in which the baseline outcomes are treated as high-cost covariates and the follow-up maths test is used as the outcome to be predicted.

For the case where baseline maths is the outcome of interest in panel (a), the OGA selects only $|\hat{I}| = 14$ out of the 145 covariates with a selected sample size of $\hat{n} = 3,018$, which is about 32% larger than the actual sample size in the experiment. The results for the LASSO and POST-LASSO methods are similar. These two methods, as mentioned above, also provide good approximations to the solution of the problem we are studying, but are computationally less attractive than OGA.

As before, we measure the performance of the three covariate selection methods by the estimated precision of the resulting treatment effect estimator ('RMSE'). Our focus is on the MSE, but notice that gains in MSE translate into gains in the power of the corresponding t-test, as

Table 7. School grants (outcome: maths test).

Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	RMSE	EQB	Relative EQB
			(a) Baseline outcome			
Experiment	2,280	142	1	0.0042272	\$30,767	1
OGA	3,018	14	0.99966	0.003916	\$28,141	0.91
LASSO	2,985	18	0.99968	0.0039727	\$28,669	0.93
POST-LASSO	2,985	18	0.99968	0.0038931	\$27,990	0.91
			(b) Follow-up outcome			
Experiment	762	143	1	0.0051298	\$52,604	1
OGA	6,755	0	0.99961	0.0027047	\$22,761	0.43
LASSO	6,755	0	0.99961	0.0027047	\$22,761	0.43
POST-LASSO	6,755	0	0.99961	0.0027047	\$22,761	0.43
			(c) Follow-up outcome, no high-cost covariates			
Experiment	762	143	1	0.0051298	\$52,604	1
OGA	5,411	140	0.99879	0.0024969	\$21,740	0.41
LASSO	5,444	136	0.99908	0.00249	\$22,082	0.42
POST-LASSO	6,197	43	0.99933	0.0024624	\$21,636	0.41
			(d) Follow-up outcome, force baseline outcome			
Experiment	762	143	1	0.0051298	\$52,604	1
OGA	1,314	133	0.99963	0.0040293	\$41,256	0.78
LASSO	2,789	1	0.9929	0.0043604	\$42,815	0.81
POST-LASSO	2,789	1	0.9929	0.0032823	\$32,190	0.61

discussed in Section 3. The three methods improve the precision by about 7% relative to the experiment. Also, all three methods manage to essentially exhaust the budget, as indicated by cost-to-budget ratios ('Cost/B') close to one. As in the previous subsection, we measure the economic gains from using the covariate selection procedures by the equivalent budget ('EQB') that each of the methods requires to achieve the precision of the experiment. All three methods require equivalent budgets that are 7–9% lower than that of the experiment.

All variables that the OGA selects as strong predictors of baseline outcome are plausibly related to student performance on a maths test.[21] They are related to important aspects of the community surrounding the school (e.g., distance to the nearest city), school equipment (e.g., number of computers), school infrastructure (e.g., number of temporary structures), human resources (e.g., teacher–student ratio, teacher training), and teacher and principal perceptions about which factors are central for success in the school and about which factors are the most important obstacles to school success.

For the case where the follow-up maths score is the outcome to be predicted in panel (b), the budget used in the experiment increases owing to the addition of the three expensive baseline outcomes to the pool of covariates. All three methods select no covariates and exhaust the budget by using the maximum feasible sample size of 6,755, which is almost nine times larger than the sample size in the experiment. The implied precision of the treatment effect estimator improves

[21] Online Appendix S5 shows the full list and definitions of selected covariates for the baseline outcome.

by about 47% relative to the experiment, which translates into the covariate selection methods requiring less than half of the experimental budget to achieve the same precision as in the experiment. These are striking statistical and economic gains from using our proposed procedure to choose covariates (in the case where covariates are mainly useful to improve the precision of the treatment effect estimator).

In the remainder of this section, we present sensitivity checks and counterfactual experiments that provide insights into why the covariate selection procedures do not recommend the inclusion of any covariates, not even baseline outcomes.

Sensitivity checks

In RCT's, baseline outcomes tend to be strong predictors of the follow-up outcome. One may therefore be concerned that, because the OGA first selects the most predictive covariates, which in this application are also much more expensive than the remaining low-cost covariates, the algorithm never examines what would happen to the estimator's MSE if it first selected the most predictive low-cost covariates instead. In principle, such selection could lead to a lower MSE than any selection that includes the very expensive baseline outcomes. As a sensitivity check, we therefore perform the covariate selection procedures on the pool of covariates that excludes the three baseline outcomes. Panel (c) shows the corresponding results. In this case, all methods indeed select more covariates and smaller sample sizes than in panel (b), and achieve a slightly smaller MSE. The budget reductions relative to the experiment as measured by EQB are also almost identical to those in panel (b). Therefore, selecting no covariates and large sample size (panel (b)) or many low-cost covariates with somewhat smaller sample size (panel (c)) yield very similar and significant improvements in precision or significant reductions in the experimental budget, respectively.[22]

One may want to ensure balance of the control and the treatment group, especially in terms of strong predictors such as baseline outcomes. Checking balance requires collection of the relevant covariates. Therefore, we also perform the three covariate selection procedures when we force each of them to include the baseline maths outcome as a covariate. In the OGA, we can force the selection of a covariate by performing group OGA as described in Appendix A, where each group contains a low-cost covariate together with the baseline maths outcome. For the LASSO procedures, we simply perform the LASSO algorithms after partialling out the baseline maths outcome from the follow-up outcome. The corresponding results are reported in panel (d). Since baseline outcomes are very expensive covariates, the selected sample sizes relative to those in panels (b) and (c) are much smaller. OGA selects a sample size of 1,314, which is almost twice as large as the experimental sample size, but about 4–5 times smaller than the OGA selections in panels (b) and (c). In contrast to OGA, the two LASSO procedures do not select any other covariates beyond the baseline maths outcome. As a result of forcing the selection of the baseline outcome, all three methods achieve an improvement in precision, or reduction of budgets respectively, of around 20% relative to the experiment. These are still substantial gains, but the requirement of checking balance on the expensive baseline outcome comes at the cost of smaller improvements in precision owing to our procedure.

[22] Note that there is no sense in which we need to be concerned about identification of the optimal set of covariates. There may indeed exist several combinations of covariates that yield similar precision of the resulting treatment effect estimator. Our objective is the highest possible precision without any direct interest in the identities of the covariates that achieve this optimum.

Table 8. School grants (outcome: maths test): varying the costs or predictive power of expensive covariates.

Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	RMSE	EQB	Relative EQB
			(a) Reduce price of high-cost covariates by 50%			
Experiment	762	143	1	0.00513	$31,948	1
OGA	14,569	0	0.99666	0.00184	$12,816	0.40
LASSO	14,569	0	0.99666	0.00184	$12,816	0.40
POST-LASSO	14,569	0	0.99666	0.00184	$12,816	0.40
			(b) Reduce price of high-cost covariates by 60%			
Experiment	762	143	1	0.00513	$27,461	1
OGA	8,623	1	0.99864	0.00187	$10,632	0.39
LASSO	8,623	1	0.99864	0.00214	$10,632	0.39
POST-LASSO	8,623	1	0.99864	0.00187	$10,632	0.39
			(c) Increase predictive power of baseline outcome by 20%			
Experiment	762	143	1	0.00434	$52,604	1
OGA	6,755	0	0.99961	0.00270	$27,304	0.52
LASSO	6,755	0	0.99961	0.00270	$27,304	0.52
POST-LASSO	6,755	0	0.99961	0.00270	$27,304	0.52
			(d) Increase predictive power of baseline outcome by 30%			
Experiment	762	143	1	0.00382	$52,604	1
OGA	2,789	1	0.99290	0.00245	$32,058	0.61
LASSO	6,755	0	0.99961	0.00270	$32,058	0.61
POST-LASSO	2,789	1	0.99290	0.00245	$32,058	0.61

In Online Appendix S6, we also perform an out-of-sample evaluation for this application by splitting the dataset into training samples for the covariate selection step and evaluation samples for the computation of the performance measures RMSE and EQB. The results are qualitatively similar to those in Table 7.

Counterfactual costs and predictive power of baseline outcomes

It is well known that in education interventions such as the one we study, pre-intervention test scores are expensive covariates but strong predictors of post-intervention test scores. In our data, about 25% of the variance of the follow-up maths score can be accounted for by the variance in the baseline maths score. The main reason why our procedure does not select it is because of its high cost. Therefore, it is worth considering the following two questions. First, by how much would we need to reduce the cost of the high-cost covariate for it to be worth collecting? Second, keeping costs unaltered, by how much would we need to improve the predictive power of the high-cost covariate in order for it to be worth collecting.

To answer the first question, we compute solutions to the covariate choice problem under different counterfactual cost functions. In particular, we examine what happens to the results when we consider hypothetical values of prices τ_j for the baseline test scores. There are three high-cost covariates, namely the three baseline outcomes. We reduce their prices τ_j simultaneously, all by the same factor. In panel (a) of Table 8, they all have price 0.5 times the actual price τ_j, and in panel (b) of Table 8, the factor is 0.4. All other aspects of the problem are kept identical to those in panel (b) of Table 7. In panel (b) of Table 8, the selected covariate is the same for all three selection procedures: the baseline outcome for the maths test.

Panel (a) of Table 8 shows that even if we reduce the cost of baseline outcomes by 50%, our procedure still provides a solution in which no covariates are chosen, just as in panel (b) of Table 7. However, if we reduce the cost by 60% (panel (b)) then the baseline maths test score is chosen. Therefore, in order to make it worthwhile to ever collect the baseline outcome as a covariate its cost would have to be substantially low.

To answer the second question, we examine what happens to our results when we consider hypothetical values of the predictive power of the baseline maths score. To this end we increase the correlation of the baseline score with the follow-up score as described in Online Appendix S8.

In panels (c) and (d) of Table 8, we increase the predictive power by 20% and 30%, respectively. All other aspects of the problem are identical to those in panel (b) of Table 7. In panel (d), the covariate selected by OGA and POST-LASSO is the same: the baseline outcome for the maths test. Panel (c) of Table 8 shows that even if we increase the predictive power of this variable by 20%, our procedure opts for not collecting it. In fact, it will only start choosing this covariate when its predictive power on the outcome is 30% higher than what we currently observe in the data. Therefore, in order to make it worthwhile to ever collect the baseline outcome as a covariate at its current cost, its predictive power on the outcome would have to be considerably high.

6. DISCUSSION

In this section, we discuss some of conceptual and practical properties of our proposed data collection procedure.

Availability of pre-experimental data

As in standard power calculations, pre-experimental data provide essential information for our procedure. The availability of such data is commonly available very high, ranging from census datasets and other household surveys to studies that were conducted in a similar context to the RCT we are planning to implement. In addition, if no such dataset is available, one may consider running a pilot project that collects pre-experimental data. We recognize that in some cases it might be difficult to have the required information readily available. However, this is a problem that affects any attempt at a data-driven design of surveys, including standard power calculations. Even when pre-experimental data are imperfect, such calculations provide a valuable guide to survey design, as long as the available pre-experimental data are not very different from the ideal data. In particular, our procedure only requires second moments of the pre-experimental variables to be similar to those in the population of interest.

The optimization problem in a simplified setup

In general, the problem in (3.8) does not have a simple solution and requires a joint optimization problem over the sample size n and the coefficient γ. To gain some intuition about the trade-offs in this problem, in Online Appendix S3 we consider a simplified setup in which all covariates are orthogonal to each other, and the budget constraint has a very simple form. In this case, the constraint can be substituted into the objective, and the optimization becomes univariate and unconstrained. We show that if all covariates have the same price, then one wants to choose covariates up to the point where the percentage increase in survey costs equals the percentage reduction in the residual variance from the last covariate. Furthermore, the elasticity of the residual variance with respect to changes in sample size should equal the elasticity of the residual variance with respect to an additional covariate. If the costs of data collection vary with covariates, then this conclusion is slightly modified. If we organize variables by type according to their contribution to the residual variance, then we want to choose variables of each type up to the point where the percentage marginal contribution of each variable to the residual variance equals its percentage marginal contribution to survey costs.

Imbalance, re-randomization, stratification

In RCTs, covariates typically do not serve only as a means to improving the precision of treatment effect estimators, but also for checking whether the control and treatment groups are balanced. See, for example, Bruhn and McKenzie (2009) for practical issues concerning randomization and balance. To rule out large biases due to imbalance, it is important to carry out balance checks for strong predictors of potential outcomes. Our procedure selects the strongest predictors as long as they are not too expensive (e.g., household survey questions such as gender, race, number of children etc.) and we can check balance for these covariates.

An alternative approach to avoiding imbalance considers re-randomization until some criterion capturing the degree of balance is met (e.g., Bruhn and McKenzie, 2009; Morgan and Rubin, 2012, 2015; Li et al., 2018). Our criterion for the covariate selection procedure in Step 2 can readily be adapted to this case; however, the details are not worked out here. It is an interesting future research topic to fully develop a data collection method for re-randomization based on the modified variance formulae in Morgan and Rubin (2012) and Li et al. (2018), which account for the effect of re-randomization on the treatment effect estimator. Similarly, we could incorporate stratified random sampling through re-randomization with preset clusters (Morgan and Rubin, 2012). It would also be interesting to extend our data collection procedure to include stratification trees, as proposed in Tabord-Mehan (2018).

Expensive, strong predictors

When some covariates have similar predictive power, but respective prices that are substantially different, our covariate selection procedure may produce a suboptimal choice. For example, if the covariate with the highest price is also the most predictive, OGA selects it first even when there are other covariates that are much cheaper but only slightly less predictive. In Section 5.2, we encounter an example of such a situation and propose a simple robustness check for whether removing an expensive, strong predictor may be beneficial.

Heteroskedasticity and robust standard errors

Recall that, in Step 2, we assumed that errors are homoskedastic with respect to treatment. We needed this assumption in order to derive the finite-sample MSE of the OLS estimator and provide a tractable solution to our data collection problem when D is not observed in the pre-experimental sample. Should one be concerned about common unobserved shocks among individuals in the experimental sample, robust (e.g., clustered or Eicker–Huber–White) standard errors can be employed in Step 5 and are valid.

Heteroskedasticity in the pre-experimental sample, however, may cause our data collection procedure in Step 2 to produce a suboptimal selection of covariates. Without observing the treatment indicator in the pre-experimental sample, a strong homogeneity assumption like our homoskedasticity condition allows us to express the MSE of the treatment effect estimator in a way that can be estimated solely based on Y and X. This is similar to standard power calculations which require much stronger assumptions, including the assumption that potential outcomes have the same variance. Should the treatment indicator be observed in the pre-experimental sample, then one could exploit the observed dependence structure between D and X to estimate the treatment effect estimator's MSE even in the presence of heteroskedasticity. For example, one may extend our framework by introducing an additional sampling step as in Hahn et al. (2011).

Multivariate outcomes

It is straightforward to extend our data collection method to the case when there are multivariate outcomes. Online Appendix S7 provides details regarding how to deal with a vector of outcomes when we select the common set of regressors for all outcomes.

7. RELATION TO THE EXISTING LITERATURE

In this section, we discuss related papers in the literature. We emphasize that the research question in our paper is different from those studied in the literature and that our paper is a complement to the existing work.

In the context of experimental economics, List et al. (2011) suggest several simple rules of thumb that researchers can apply to improve the efficiency of their experimental designs. They discuss the issue of experimental costs and estimation efficiency but do not consider the problem of selecting covariates.

Hahn et al. (2011) consider the design of a two-stage experiment for estimating an average treatment effect, and propose to select the propensity score that minimizes the asymptotic variance bound for estimating the average treatment effect. Their recommendation is to assign individuals randomly between the treatment and control groups in the second stage, according to the optimized propensity score. They use the covariate information collected in the first stage to compute the optimized propensity score.

Bhattacharya and Dupas (2012) consider the problem of allocating a binary treatment under a budget constraint. Their budget constraint limits what fraction of the population can be treated, and hence is different from our budget constraint. They discuss the costs of using a large number of covariates in the context of treatment assignment.

McKenzie (2012) demonstrates that taking multiple measurements of the outcomes after an experiment can improve power under the budget constraint. His choice problem is how to allocate a fixed budget over multiple surveys between a baseline and follow-ups. The main source of the improvement in his case comes from taking repeated measures of outcomes; see Frison and Pocock (1992) for this point in the context of clinical trials. In the setup of McKenzie (2012), a baseline survey measuring the outcome is especially useful when there is high autocorrelation in outcomes. This would be analogous in our paper to devoting part of the budget to the collection of a baseline covariate, which is highly correlated with the outcome (in this case, the baseline value of the outcome), instead of just selecting a post-treatment sample size that is as large as the budget allows for. In this way, McKenzie (2012) is perhaps closest to our paper in spirit.

In a recent paper, Dominitz and Manski (2017) proposed the use of statistical decision theory to study allocation of a predetermined budget between two sampling processes of outcomes: a high-cost process of good data quality and a low-cost process with non-response or low-resolution interval measurement of outcomes. Their main concern is data quality between two sampling processes and is distinct from our main focus, namely the simultaneous selection of the set of covariates and the sample size.

8. CONCLUDING REMARKS

We have developed data-driven methods for designing a survey in a randomized experiment using information from a pre-existing dataset. Our procedure is optimal in the sense that it minimizes the mean squared error of the average treatment effect estimator and maximizes the power of the corresponding t-test, and can handle a large number of potential covariates as well as complex budget constraints faced by the researcher. We have illustrated the usefulness of our approach by showing substantial improvements in precision of the resulting estimator or substantial reductions in the researcher's budget in two empirical applications.

We recognize that there are many potential reasons guiding the choice of covariates in a survey. These may be as important as the one we focus on, which is the precision of the treatment effect estimator. We show that it is possible and important to develop practical tools to help researchers make such decisions. We regard our paper as part of the broader task of making the research design process more rigorous and transparent.

Some important issues remain as interesting future research topics. For example, we have assumed that the pre-experimental sample $\mathcal{S}_{\text{pre}}$ is large, and therefore the difference between the minimization of the sample average and that of the population expectation is negligible. However, if the sample size of $\mathcal{S}_{\text{pre}}$ is small (e.g., in a pilot study), one may be concerned about over-fitting, in the sense of selecting too many covariates. A straightforward solution would be to add a term to the objective function that penalizes a large number of covariates via some information criteria (e.g., the Akaike information criterion or the Bayesian information criterion). Another possibility is to consider data collection for the RCTs that are likely to suffer from partial compliance. One may focus on the local average treatment effect in this setup and investigate the problem of optimal data collection by combining the insights from this paper with those from using Neyman-orthogonal moment conditions (e.g., Chernozhukov et al., 2018).

ACKNOWLEDGEMENTS

We thank the Co-Editor, Victor Chernozhukov, two anonymous referees, Frank Diebold, Kirill Evdokimov, Michal Kolesar, David McKenzie, Ulrich Müller, Andriy Norets, Imran Rasul, and participants at various seminars for helpful comments and discussions. An early version of this paper was presented at Columbia University and Princeton University in September 2014, and at New York University and the University of Pennsylvania in December 2014. This work was supported in part by the European Research Council (ERC-2014-CoG-646917-ROMIA and ERC-2015-CoG-682349) and by the UK Economic and Social Research Council (ESRC) through a grant (ES/P008909/1) to the ESRC CeMMAP.

REFERENCES

Angelucci, M., D. Karlan and J. Zinman (2015). Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics 7*, 151–82.

Attanasio, O., B. Augsburg, R. De Haas, E. Fitzsimons and H. Harmgart (2015). The impacts of microfinance: Evidence from joint-liability lending in Mongolia. *American Economic Journal: Applied Economics 7*, 90–122.

Attanasio, O., R. P. de Barros, P. Carneiro, D. Evans, L. Lima, R. Mendonca, P. Olinto and N. Schady (2014). Free access to child care, labor supply, and child development, University College London. Technical report.

Augsburg, B., R. De Haas, H. Harmgart and C. Meghir (2015). The impacts of microcredit: Evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics 7*, 183–203.

Bandiera, O., I. Barankay and I. Rasul (2011). Field experiments with firms. *Journal of Economic Perspectives 25*(3), 63–82.

Banerjee, A., S. Chassang, S. Montero and E. Snowberg (2017). A theory of experimenters. Working Paper 23867, NBER.

Banerjee, A., E. Duflo, R. Glennerster and C. Kinnan (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics 7*, 22–53.

Banerjee, A., D. Karlan and J. Zinman (2015). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics 7*, 1–21.

Banerjee, A. V. and E. Duflo (2009). The experimental approach to development economics. *Annual Review of Economics 1*, 151–78.

Barron, A. R., A. Cohen, W. Dahmen and R. A. DeVore (2008). Approximation and learning by greedy algorithms. *Annals of Statistics 36*, 64–94.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*, 521–47.

Bhattacharya, D. and P. Dupas (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics 167*, 168–96.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*, 200–32.

Carneiro, P., O. Koussihouèdé, N. Lahire, C. Meghir and C. Mommaerts (2015). Decentralizing education resources: School grants in Senegal. Working Paper 21063, NBER.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters, *Econometrics Journal 21*, C1–C68.

Coffman, L. C. and M. Niederle (2015). Pre-analysis plans have limited upside, especially where replications are feasible, *Journal of Economic Perspectives 29*(3), 81–98.

Crépon, B., F. Devoto, E. Duflo and W. Parienté (2015). Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco. *American Economic Journal: Applied Economics 7*, 123–50.

Davis, G., S. Mallat and M. Avellaneda (1997). Adaptive greedy approximations. *Constructive Approximation 13*, 57–98.

Dominitz, J. and C. F. Manski (2017). More data or better data? A statistical decision problem, *Review of Economic Studies 84*, 1583–605.

Duflo, E., P. Dupas and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya, *American Economic Review 101*(5), 1739–74.

Duflo, E., P. Dupas and M. Kremer (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics 123*, 92–110.

Duflo, E., R. Glennerster and M. Kremer (2007). Using randomization in development economics research: A toolkit. In Schultz, T. P. and J. A. Strauss (Eds.), *Handbook of Development Economics*, Volume *4*, Chapter 61, pp. 3895–3962. Elsevier.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen and K. A. Baicker (2012). The Oregon health insurance experiment: Evidence from the first year. *Quarterly Journal of Economics 127*, 1057–1106.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*(1st ed.). New York, NY: Hafner.

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

Frison, L. and S. J. Pocock (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design, *Statistics in Medicine 11* 1685–704.

Glewwe, P., N. Ilias and M. Kremer (2010). Teacher incentives. *American Economic Journal: Applied Economics 2*, 205–27.

Glewwe, P., M. Kremer and S. Moulin (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics 1*, 112–35.

Hahn, J., K. Hirano and D. Karlan (2011). Adaptive experimental design using the propensity score, *Journal of Business and Economic Statistics 29*, 96–108.

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature 51*, 162–72.

Huang, J., T. Zhang and D. Metaxas (2011). Learning with structured sparsity. *Journal of Machine Learning Research 12*, 3371–3412.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Ing, C.-K. and T. L. Lai (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica 21*, 1473–513.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis 24*, 324–38.

Kleinberg, J., J. Ludwig, S. Mullainathan and Z. Obermeyer (2015). Prediction policy problems. *American Economic Review 105*(5), 491–95.

Kremer, M., E. Miguel and R. Thornton (2009). Incentives to learn. *Review of Economics and Statistics 91*, 437–56.

Li, X., P. Ding and D. B. Rubin (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences 115*, 9157–62.

List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives 25*(3), 3–15.

List, J. A. and I. Rasul (2011). Field experiments in labor economics. In Ashenfelter, O. and D. Card (Eds.), *Handbook of Labor Economics*, Volume *4*, Part A, Chapter 2, pp. 103–228. Elsevier.

List, J. A., S. Sadoff and M. Wagner (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics 14*, 439–57.

McConnell, B. and M. Vera-Hernandez (2015). Going beyond simple sample size calculations: A practioner's guide. IFS Working Paper W15/17, Institute for Fiscal Studies, London.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics 99*, 210–21.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics 11*, 57–91.

Miguel, E. and M. Kremer (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*, 159–217.

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

Morgan, K. L. and D. B. Rubin (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association 110*, 1412–21.

Natarajan, B. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing 24*, 227–34.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives 29*, 61–80.

Sancetta, A. (2016). Greedy algorithms for prediction. *Bernoulli 22*, 1227–77.

Tabord-Mehan, M. (2018). Stratification trees for adaptive randomization in randomized controlled trials. Technical report, Northwestern University.

Tarozzi, A., J. Desai and K. Johnson (2015). The impacts of microcredit: Evidence from Ethiopia. *American Economic Journal: Applied Economics 7*, 54–89.

Temlyakov, V. N. (2011). *Greedy Approximation*. Cambridge: Cambridge University Press.

Tetenov, A. (2016). An economic theory of statistical testing. Technical report, University of Bristol.

Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory 50*, 2231–42.

Tropp, J. A. and A. C. Gilbert (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory 53*, 4655–66.

Vermeersch, C. and M. Kremer (2004). Schools meals, educational achievement and school competition: Evidence from a randomized evaluation. Policy Research Working Paper Series 3523, The World Bank.

Willa, F., K. Michael, M. Edward and T. Rebecca (2016). Education as liberation? *Economica 83*, 1–30.

Zhang, T. (2009). On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research 10*, 555–68.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Co-editor Victor Chernozhukov handled this manuscript.