cemmap

centre for microdata methods and practice

Optimal Data Collection for Randomized Control Trials

Pedro Carneiro
Sokbae Lee
Daniel Wilhelm

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP45/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Optimal Data Collection for Randomized Control Trials

By Pedro Carneiro, Sokbae Lee, and Daniel Wilhelm*

October 12, 2017

Abstract

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. We propose the use of pre-experimental data such as other similar studies, a census, or a household survey, to inform the choice of both the sample size and the covariates to be collected. Our procedure seeks to minimize the resulting average treatment effect estimator's mean squared error and/or maximize the corresponding t-test's power, subject to the researcher's budget constraint. We rely on a modification of an orthogonal greedy algorithm that is conceptually simple and easy to implement in the presence of a large number of potential covariates, and does not require any tuning parameters. 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.

JEL codes: C55, C81.

Key words: randomized control trials, big data, data collection, optimal survey design, orthogonal greedy algorithm, survey costs.

*Carneiro: University College London, Institute for Fiscal Studies (IFS), and Centre for Microdata Methods and Practice (CeMMAP); Lee: Columbia University, IFS and CeMMAP; Wilhelm: University College London and CeMMAP. We thank Frank Diebold, Kirill Evdokimov, Michal Kolesar, David McKenzie, Ulrich Müller, Andriy Norets, Imran Rasul, and participants at various seminars for helpful discussions. An early version of this paper was presented at Columbia University and Princeton University in September 2014, and at New York University and 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.

I 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, Glennerster, and Kremer, 2007; Banerjee and Duflo, 2009; Bandiera, Barankay, and Rasul, 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 analyze 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.

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 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

[1] Tetenov (2015) provides a decision-theory-based rationale for using hypothesis tests in the RCT and Banerjee, Chassang, Montero, and Snowberg (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.

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 all the potentially relevant variables that one would consider collecting for the analysis of the experiment. This assumption may be reasonable in many different contexts. For example, the Abdul Latif Jamil Poverty Action Lab (JPAL) website lists more than 10 papers written since 2000 analyzing different education interventions taking place in Kenya, almost all of them involving a single district (Busia).[2] When designing the data collection procedures, researchers for the later studies could have relied on information from the earlier studies.

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.[3]

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.[4] The trade-offs involved in this choice involve basic economic reasoning. For

[2]See, e.g., Duflo, Dupas, and Kremer (2011, 2015); Kremer, Miguel, and Thornton (2009); Miguel and Kremer (2004), and papers listed at `https://www.povertyactionlab.org`.

[3]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 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.

[4]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.

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, for each covariate, there is a discrete choice of whether to include it or not, and for each possible sample size, one needs to consider all possible combinations of covariates within the budget. This requires a solution to a computationally difficult combinatorial optimization problem. This problem is especially challenging when the set of potential variables to choose from is large, a case that is increasingly encountered in today's big data environment. Fortunately, with the increased availability of high-dimensional data, methods for the analysis of such data sets have received growing attention in several fields, including economics (Belloni, Chernozhukov, and Hansen, 2014). This literature makes available a rich set of new tools, which can be adapted to our study of optimal survey design.

In this paper, we propose the use of a computationally attractive procedure based on the orthogonal greedy algorithm (OGA) – also known as the orthogonal matching pursuit; see, for example, Tropp (2004) and Tropp and Gilbert (2007) among many others. In the standard implementation of the OGA, it is necessary to specify a stopping rule, which in turn requires a tuning parameter. In our modified version of the algorithm, the budget constraint, an economically meaningful constraint, replaces this tuning parameter and thus does not require the researcher to choose any other statistical tuning parameters. This new, automated procedure is easy to implement as it requires only running linear regressions (a Matlab implementation is available at `http://github.com/danielwilhelm/Matlab-data-coll`) and it performs well even when there are a large number of potential covariates in the pre-experimental data set.

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 are the minimum costs of obtaining the same precision of the treatment effect as in the actual experiment, if one was 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 percent (up to 58 percent) of the original budget in the first (second) example we consider, while maintaining the same level of the 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 III.C 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 (1935). There also exists an extensive body of literature on sample size (power) calculations; see, for example, McConnell and Vera-Hernández (2015) for a practical guide. Both bodies of this literature are concerned with the precision of treatment effect estimates, but neither addresses the problem that concerns us. For instance, McConnell and Vera-Hernández (2015) have developed methods to choose the sample size when cost constraints are binding, but they neither 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, Hirano, and Karlan, 2011; List, Sadoff, and Wagner, 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 VII.

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, Ludwig, Mullainathan, and Obermeyer (2015) argue that prediction problems are abundant in economic policy analysis, and recent advances in machine learning can be used to tackle those problems. Fur-

thermore, 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.

The remainder of the paper is organized as follows. In Section II 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 III. In Section IV, we propose the use of a simple algorithm based on the OGA. In Section V, we discuss the costs of data collection in experiments. In Section VI, we present two empirical applications, in Section VII, we discuss the existing related literature, and in Section VIII, we give concluding remarks. Online appendices provide details that are omitted from the main text.

II A Stylized Special Case

Consider the case in which a researcher is designing the survey for a 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, Miguel, and Thornton (2009), part of which we reproduce in Table 1. 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 column 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 third column is due to small differences in sample size).[5]

Suppose the researcher is interested in collecting data for an evaluation of a new

[5]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.

Table 1: Impact of incentives to learn on test scores with and without controls for lagged test scores (from Kremer, Miguel, and Thornton (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%				

education intervention in a similar population. She is deciding between collecting only the post-intervention test score on a large sample (Case (i)), or collect 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:

$$Y = \alpha_1 + \beta D + V, \tag{1}$$

$$Y = \alpha_2 + \beta D + \gamma_2 X + U, \tag{2}$$

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. If U is uncorrelated with X, then $\sigma_V^2 = \sigma_U^2 + \gamma_2^2 \sigma_X^2$, 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 OLS estimators of β in (1) and (2), and $AVar(\hat{\beta})$ is the asymptotic variance of $\hat{\beta}$ divided by the sample size. Table 2 shows the main quantities influencing the researcher's decision.

Suppose the researcher measures precision of the treatment effect estimators by their asymptotic variances 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: $AVar(\hat{\beta}_1) > AVar(\hat{\beta}_2)$,[6] 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 1, suppose we start with columns 3 and 4. The

[6]Since the OLS estimators in an RCT are unbiased, the variance of the estimator is equal to its mean-squared error. In Section III.A, we justify this criterion in a decision-theoretic framework.

Table 2: 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	$n_1 = \frac{B}{\lambda}$	$n_2 = \frac{B}{2\lambda}$
$AVar(\hat{\beta})$	$\frac{\sigma_V^2}{n_1 Var(D)}$	$\frac{\sigma_U^2}{n_2 Var(D)}$

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 $R_{yx} \approx 0.59$, 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} was below 50%, or if the cost of collecting the pre-intervention outcome (λ_{pre}) was at least 44% higher than the cost of collecting the post-intervention outcome (λ_{post}).[7] 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)?

[7] 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 $\lambda_{pre} < \frac{0.59}{0.41}\lambda_{post} \approx 1.44\lambda_{post}$.

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 3 shows the gains in precision and costs for other typical R_{yx} values.[8] As R_{xy} decreases both the statistical and economic gains from choosing the optimal covariates and sample size grow. For example, if $R_{yx} = 0.10$, then the statistical gains in precision of the treatment effect estimator are 25% and the the costs of data collection decrease by 44%. In our empirical applications in Section VI, we will find possible gains that are even larger.

Table 3: Gains in precision and cost from optimal covariate choice

R_{yx}	$\sqrt{\frac{AVar(\hat{\beta}_1)}{AVar(\hat{\beta}_2)}}$	% gain	$\frac{B_1}{B_2}$	% 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

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 size of the survey etc.). Typically it is not possible

[8] 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 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.

to derive analytical decision rules as in the stylized example presented here. The next sections present the general formulation and our proposed solution.

III 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, where we measure precision by the finite sample mean-squared error (MSE) of the treatment effect estimator, and/or produce a powerful t-test of the hypothesis $H_0 : \beta_0 = 0$. 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

$$Y = \alpha_0 + \beta_0 D + \gamma_0' Z + U, \tag{3}$$

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 the remainder of this section, we first apply decision theory to our data collection problem, then present our proposed procedure that implements the decision-theoretic solution, and finally discuss properties of our procedure.

III.A Decision Theory Applied to Data Collection

To describe our decision theoretic problem, we first introduce some notation. Let $\Theta := \Theta_1 \times \Gamma$, where $\Theta_1 \subseteq \mathbb{R}$, $\Gamma \subseteq \mathbb{R}^M$, be a parameter space and decompose a typical element θ as $\theta = (\beta, \gamma')'$ where $\beta \in \Theta_1$ corresponds to possible average treatment effect parameters and $\gamma \in \Gamma$ to possible coefficients for the additional covariates. Let $\mathcal{S}_{pre} := \{Y_i, X_i\}_{i=1}^N$ and $\mathcal{S}_{exp} := \{Y_i, D_i, X_i\}_{i=1}^n$ denote the pre-experimental and experimental random samples that take values in the sample spaces $\mathbb{R}^{N(M+1)}$ and $\mathbb{R}^{n(M+2)}$, respectively. Let $\mathbb{A} \subseteq \mathbb{R}$ be the action space. The goal is to make a decision, i.e. choose an estimator, $\hat{\beta} : \mathbb{S}_{exp} \to \mathbb{A}$, which maps the experimental sample into an action, i.e. into an estimate of the average treatment effect β_0. We evaluate the performance of the decision rule $\hat{\beta}$ by average risk $R_n : \mathbb{D}_n \to \mathbb{R}$ defined as

$$R_n(\hat{\beta}) := \int_\Theta L_n(\theta, \hat{\beta}(\mathcal{S}_{exp}))d\mu(\theta),$$

where $\mu : \Theta \to \mathbb{R}$ is a weight function that integrates to one, $L_n : \Theta \times \mathbb{A} \to \mathbb{R}$ is a loss function, and $\mathbb{D}_n$ a set of feasible decision rules. To make a decision we minimize average risk:

$$\min_{\hat{\beta} \in \mathbb{D}_n} R_n(\hat{\beta}). \tag{4}$$

In the context of our data collection problem, we can implement our proposed data collection procedure only after making some specific choices about the loss function, the set of feasible decision rules and the weighting function μ. First, we pick average squared loss (or mean-squared error):

$$L_n(\theta, \hat{\beta}(\mathcal{S}_{exp})) := E_{exp,\theta}\left[\left(\beta - \hat{\beta}(\mathcal{S}_{exp})\right)^2\right],$$

where $E_{exp,\theta}$ denotes the expectation with respect to the distribution of the experimental random sample $\mathcal{S}_{exp}$, which might depend on the unknown state θ. This is a common loss function for the evaluation of estimators and leads to a particularly simple formulation of the resulting data collection problem that can easily be implemented in practice.

For any random sample $\mathcal{S} \in \mathbb{S}_{exp}$ and any $\gamma \in \Gamma$, let $b(\gamma, \mathcal{S})$ be the OLS estimator from a regression of $Y - \gamma' X$ on a constant and D. We choose the set of feasible decision rules to contain all $\hat{\beta}$ that can be written as the OLS estimator $b(\gamma, \cdot)$ for some value of $\gamma \in \Gamma$ and also satisfy the budget constraint:

$$\mathbb{D}_n := \Big\{ b(\gamma, \cdot) : \mathbb{S}_{exp} \to \mathbb{A} \text{ s.t. } \gamma \in \Gamma \text{ and } c(\mathcal{I}(\gamma), n) \leq B \Big\}.$$

This definition means that the decision rules in $\mathbb{D}_n$ can be indexed by $\gamma \in \Gamma$ that satisfy the budget constraint. Therefore, choosing a decision rule in $\mathbb{D}_n$ is equivalent to choosing a feasible value of γ.

The weighting function μ is chosen as a product of two densities, $\mu := \mu_\beta \times \mu_\gamma$, where μ_β is a density on Θ_1 and μ_γ a density on Γ.

Since prior to the experiment we do not observe the experimental sample and the distribution of the experimental data is unknown, in practice it is not possible to minimize $R_n(\hat{\beta})$ over $\hat{\beta} \in \mathbb{D}_n$. However, we now show that, under a homoskedasticity assumption, this minimization problem is equivalent to a minimization problem that only involves the pre-experimental sample and thus can be solved prior to the experiment. To this end define the variance $Var_{pre,\gamma}$ based on $E_{pre,\gamma}$, the expectation with respect to the distribution of $\mathcal{S}_{pre}$, which may depend on γ, but not on β. Similarly, let $Var_{exp,\theta}$ be the variance based on $E_{exp,\theta}$. Define $\Delta(n) := (E[(\bar{D}_n(1 - \bar{D}_n))^{-1}])^{-1}$ with $\bar{D}_n := n^{-1} \sum_{i=1}^n D_i$. We will assume that the treatment assignment D_i is completely randomized, so that in particular the distribution of D_i does not depend on θ and thus the expectation in the definition of $\Delta(n)$ is independent of θ. Given a distribution for treatment assignment (e.g. Bernoulli with success probability $p = 1/2$), the quantity $\Delta(n)$ is known.

Assumption 1. **(i)** *$\mathcal{S}_{exp}$ is an i.i.d. sample and D_i is completely randomized, i.e. $P(D_i = 1|X_i) = p$ a.s. for some $p \in (0, 1)$.* **(ii)** *$Var_{exp,\theta}(Y - \hat{\gamma}' X|D = 1) = Var_{exp,\theta}(Y - \hat{\gamma}' X|D = 0)$, where $\theta = (\beta, \gamma')'$, for all $\beta \in \Theta_1$ and $\gamma, \hat{\gamma} \in \Gamma$.* **(iii)** *$E_{exp,\theta}[X_i X_i'] = E_{pre,\gamma}[X_i X_i']$, where $\theta = (\beta, \gamma')'$, for all $\beta \in \Theta_1$ and $\gamma, \hat{\gamma} \in \Gamma$.*

Theorem 1. *Under Assumption 1, the minimizer of* (4) *is equal to $b(\hat{\gamma}^*, \cdot)$, where $\hat{\gamma}^*$ is the minimizer of*

$$\min_{\hat{\gamma} \in \Gamma :\, c(\mathcal{I}(\hat{\gamma}), n) \leq B} \frac{1}{n\Delta(n)} \int_\Gamma (\gamma - \hat{\gamma})' E_{pre,\gamma}[X_i X_i'](\gamma - \hat{\gamma}) d\mu_\gamma(\gamma).$$

The proof of the theorem is given in Section IX. This theorem implies that we can

minimize risk of the treatment effect estimator by finding the optimal combination of covariates in the pre-experimental sample subject to the budget constraint. This optimization problem can be solved in practice for any user-chosen weight function μ_γ. We make it operational in the next subsection.[9]

III.B Our Proposed Procedure

In this section, we propose our data collection procedure whose main step consists of an implementation of the decision-theoretic solution from the previous section, based on the availability of a pre-experimental 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. Implement the RCT and collect the experimental data $\mathcal{S}_{\text{exp}}$ on (Y, D, Z).
4. Estimate the average treatment effect using $\mathcal{S}_{\text{exp}}$.
5. 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, which is described more formally in Section IV.

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.[10] This pre-experimental sample $\mathcal{S}_{\text{pre}} = \{Y_i, X_i\}_{i=1}^N$ is of size N. Our framework allows the number of potential covariates, M, to be very large (possibly much larger than the sample size N). Typical examples would be census data, household surveys, or data from other, similar experiments. Another possible candidate is a pilot experiment that was carried out before the larger-scale role-out of the main experiment, provided that the sample size N of the pilot study is large enough for our econometric analysis in Step 2.

[9] Although we do not need this assumption for the theorem above, it is reasonable to consider the case where $E_{exp,\theta}[Y_i X_i'] = E_{pre,\gamma}[Y_i X_i']$. With this assumption it is reasonable to choose a weight function μ_γ that puts all weight on the available information from the pre-experimental sample.

[10] In fact, we do not need the populations to be identical, but only require second moments to be the same.

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. For each candidate sample size n, we want to solve the decision-theoretic problem in Theorem 1. To this end we choose a weight function μ_γ that puts all weight on the available information from the pre-experimental sample: a point mass at the constraint least-squares estimator for γ in the pre-experimental sample given the budget constraint. Then, assuming that $E_{pre,\gamma}[X_i X_i']$ is of full rank, the risk minimization problem (4) for a candidate sample size n becomes

$$\min_{\hat{\gamma}\in\Gamma:\, c(\mathcal{I}(\hat{\gamma}),n)\leq B} \frac{1}{n\Delta(n)N}\sum_{i=1}^{N}(Y_i - \hat{\gamma}' X_i)^2,$$

which is simply constraint least-squares in the pre-experimental sample. This optimization problem is now in the form of a prediction problem that can be solved either directly or, in the case of high-dimensional covariate vector X_i, using machine learning techniques.

In Section IV, we describe a particular machine learning algorithm, OGA, that we have found to be computationally attractive and to work well in our simulations (Appendix D) and empirical applications (Section VI). The OGA has been studied extensively in the signal extraction literature and is implemented in most statistical software packages. Appendices A and D show that this algorithm possesses desirable theoretical and practical properties.

The basic idea of the algorithm (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 n in a grid $\mathcal{N} := \{n_0, n_1, \ldots, n_K\}$, and search for the combination of sample size and covariate selection that minimizes the residual variance. The joint optimization

problem over n and γ is thus

$$\min_{n\in\mathcal{N}} \min_{\hat{\gamma}\in\Gamma:\, c(\mathcal{I}(\hat{\gamma}),n)\leq B} \frac{1}{n\Delta(n)N} \sum_{i=1}^{N} (Y_i - \hat{\gamma}' X_i)^2. \tag{5}$$

Denote the OGA solution by $(\hat{n}, \hat{\gamma})$ and let $\hat{\mathcal{I}} := \mathcal{I}(\hat{\gamma})$ denote the selected covariates. See Section IV for more details.

Note that, in general applications of the algorithm, the OGA requires the researcher to specify a stopping rule for the iterative procedure. One attractive feature of the use of OGA in our data collection context is that the budget constraint provides a natural, economically motivated stopping rule, avoiding the need for any other tuning parameters.

Step 3. **Experiment and data collection.** Given the optimal selection of covariates $\hat{\mathcal{I}}$ and sample size $\hat{n}$, we randomly assign $\hat{n}$ individuals to either the treatment or the control group (with equal probability), and collect the covariates $Z := X_{\hat{\mathcal{I}}}$ from each of them. This yields the experimental sample $\mathcal{S}_{\text{exp}} = \{Y_i, D_i, Z_i\}_{i=1}^{\hat{n}}$ from $(Y, D, X_{\hat{\mathcal{I}}})$.

Step 4. **Estimation of the average treatment effect.** We regress Y_i on $(1, D_i, Z_i)$ using the experimental sample $\mathcal{S}_{\text{exp}}$. The OLS estimator of the coefficient on D_i is the average treatment effect estimator $\hat{\beta}$.

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 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).

III.C Discussion

In this subsection, 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 very common, ranging from census data sets and other household surveys to studies that were conducted in a similar context as the RCT we are planning to implement. In addition, if no such data set 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 to 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 (5) does not have a simple solution and requires joint optimization problem over the sample size n and the coefficient γ. To gain some intuition about the trade-offs in this problem, in Appendix C 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 percent marginal contribution of each variable to the residual variance equals its percent marginal contribution to survey costs.

Imbalance and Re-randomization. In RCTs, covariates typically do not only serve 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. However, in principle, it is possible that our procedure does not select a strong predictor that is very expensive (e.g. baseline test scores). Such a situation occurs in our second empirical application (Section VI.B). In this case, in Step 2, we recommend running the OGA a second time, forcing the inclusion of such expensive predictors. If the MSE of the resulting estimate is not much larger than that from the selection without the expensive predictor, then we may prefer the former selection to the latter so as to reduce the potential for bias due to imbalance at the expense of slightly larger variance of the treatment effect estimator.

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) and Li, Ding, and Rubin (2016)). 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, Ding, and Rubin (2016), which account for the effect of re-randomization on the treatment effect estimator.

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 VI.B, we encounter an example of such a situation and propose a simple robustness check for whether removing an expensive, strong predictor may be beneficial.

Properties of the Treatment Effect Estimator. Since the treatment indicator is assumed independent of X, standard asymptotic theory of the treatment effect estimator continues to hold for our estimator (despite the addition of a covariate selection step). For example, it is unbiased, consistent, asymptotically normal, and adding the covariates X in the regression in (3) cannot increase the asymptotic variance of the estimator. In fact, inclusion of a covariate strictly reduces the estimator's asymptotic variance as long as the corresponding true regression coefficient is not zero. All these results hold regardless of whether the true conditional expectation of Y given D and X is in fact linear and additive separable as in (3) or not. In particular, in some applications one may want to include interaction terms of D and X (see, e.g., Imbens and

Rubin, 2015). Finally, the treatment effect can be allowed to be heterogeneous (i.e. vary across individuals i) in which case our procedure estimates the average of those treatment effects.

An Alternative to Regression. Step 4 consists of running the regression in (3). There are instances when it is desirable to modify this step. For example, if the selected sample size $\hat{n}$ is smaller than the number of selected covariates, then the regression in (3) is not feasible. However, if the pre-experimental sample $\mathcal{S}_{\text{pre}}$ is large enough, we can instead compute the OLS estimator $\hat{\gamma}$ from the regression of Y on $X_{\hat{\mathcal{I}}}$ in $\mathcal{S}_{\text{pre}}$. Then use Y and Z from the experimental sample $\mathcal{S}_{\text{exp}}$ to construct the new outcome variable $\hat{Y}_i^* := Y_i - \hat{\gamma}' Z_i$ and compute the treatment effect estimator $\hat{\beta}$ from the regression of $\hat{Y}_i^*$ on $(1, D_i)$. This approach avoids fitting too many parameters when the experimental sample is small and has the additional desirable property that the resulting estimator is free from bias due to imbalance in the selected covariates.

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

IV A Simple Greedy Algorithm

In practice, the vector X of potential covariates is typically high-dimensional, which makes it challenging to solve the optimization problem (5). In this section, we propose a computationally feasible algorithm that is both conceptually simple and performs well in our simulations. In particular, it requires only running many univariate, linear regressions and can therefore easily be implemented in popular statistical packages.

We split the joint optimization problem in (5) over n and γ into two nested problems. The outer problem searches over the optimal sample size n, while the inner problem determines the optimal selection of covariates for each sample size n:

$$\min_{n \in \mathcal{N}} \frac{1}{n\Delta(n)} \min_{\gamma \in \mathbb{R}^M} \frac{1}{N} \sum_{i=1}^{N} (Y_i - \gamma' X_i)^2 \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma), n) \leq B. \tag{6}$$

To convey our ideas in a simple form, suppose for the moment that the budget con-

straint has the following linear form,

$$c(\mathcal{I}(\gamma), n) = n \cdot |\mathcal{I}(\gamma)| \leq B,$$

where $|\mathcal{I}(\gamma)|$ denotes the number of non-zero elements of γ. Note that the budget constraint puts the restriction on the number of selected covariates, that is, $|\mathcal{I}(\gamma)| \leq B/n$.

It is known to be NP-hard (non-deterministic polynomial time hard) to find a solution to the inner optimization problem in (6) subject to the constraint that γ has m non-zero components, also called an m-term approximation, where m is the integer part of B/n in our problem. In other words, solving (6) directly is not feasible unless the dimension of covariates, M, is small (Natarajan, 1995; Davis, Mallat, and Avellaneda, 1997).

There exists a class of computationally attractive procedures called greedy algorithms that are able to approximate the infeasible solution. See Temlyakov (2011) for a detailed discussion of greedy algorithms in the context of approximation theory. Tropp (2004), Tropp and Gilbert (2007), Barron, Cohen, Dahmen, and DeVore (2008), Zhang (2009), Huang, Zhang, and Metaxas (2011), Ing and Lai (2011), and Sancetta (2016), among many others, demonstrate the usefulness of greedy algorithms for signal recovery in information theory, and for the regression problem in statistical learning. We use a variant of OGA that can allow for selection of groups of variables (see, for example, Huang, Zhang, and Metaxas (2011)).

To formally define our proposed algorithm, we introduce some notation. For a vector v of N observations $v_1, \ldots, v_N$, let $\|v\|_N := (1/N \sum_{i=1}^{N} v_i^2)^{1/2}$ denote the empirical L^2-norm and let $\mathbf{Y} := (Y_1, \ldots, Y_N)'$.

Suppose that the covariates $X^{(j)}$, $j = 1, \ldots, M$, are organized into p pre-determined groups $X_{G_1}, \ldots, X_{G_p}$, where $G_k \subseteq \{1, \ldots, p\}$ indicates the covariates of group k. We denote the corresponding matrices of observations by bold letters (i.e., $\mathbf{X}_{G_k}$ is the $N \times |G_k|$ matrix of observations on X_{G_k}, where $|G_k|$ denotes the number of elements of the index set G_k). By a slight abuse of notation, we let $\mathbf{X}_k := \mathbf{X}_{\{k\}}$ be the column vector of observations on X_k when k is a scalar. One important special case is that in which each group consists of a single regressor. Furthermore, we allow for overlapping groups; in other words, some elements can be included in multiple or even all groups. The group structure occurs naturally in experiments where data collection is carried out through surveys whose questions can be grouped in those concerning income, those concerning education, and so on. This can also occur naturally when we consider multivariate

outcomes. See Appendix G for details.

Suppose that the largest group size $J_{\max} := \max_{k=1,\ldots,p} |G_k|$ is small, so that we can implement orthogonal transformations *within each group* such that $(\mathbf{X}'_{G_j}\mathbf{X}_{G_j})/N = \mathbf{I}_{|G_j|}$, where $\mathbf{I}_d$ is the d-dimensional identity matrix. In what follows, assume that $(\mathbf{X}'_{G_j}\mathbf{X}_{G_j})/N = \mathbf{I}_{|G_j|}$ without loss of generality. Let $|\cdot|_2$ denote the ℓ_2 norm. The following procedure describes our algorithm for a general cost function c.

Step 1. Set the initial sample size $n = n_0$.

Step 2. Group OGA for a given sample size n:

(a) initialize the inner loop at $k = 0$ and set the initial residual $\hat{\mathbf{r}}_{n,0} = \mathbf{Y}$, the initial covariate indices $\hat{\mathcal{I}}_{n,0} = \emptyset$ and the initial group indices $\hat{\mathcal{G}}_{n,0} = \emptyset$;

(b) separately regress $\hat{\mathbf{r}}_{n,k}$ on each group of regressors in $\{1,\ldots,p\}\backslash\hat{\mathcal{G}}_{n,k}$; call $\hat{j}_{n,k}$ the group of regressors with the largest ℓ_2 regression coefficients,

$$\hat{j}_{n,k} := \arg \max_{j\in\{1,\ldots,p\}\backslash\hat{\mathcal{G}}_{n,k}} \left|\mathbf{X}'_{G_j}\hat{\mathbf{r}}_{n,k}\right|_2;$$

add $\hat{j}_{n,k}$ to the set of selected groups, $\hat{\mathcal{G}}_{n,k+1} = \hat{\mathcal{G}}_{n,k} \cup \{\hat{j}_{n,k}\}$;

(c) regress $\mathbf{Y}$ on the covariates $\mathbf{X}_{\hat{\mathcal{I}}_{n,k+1}}$ where $\hat{\mathcal{I}}_{n,k+1} := \hat{\mathcal{I}}_{n,k} \cup G_{\hat{j}_{n,k}}$; call the regression coefficient $\hat{\gamma}_{n,k+1} := (\mathbf{X}'_{\hat{\mathcal{I}}_{n,k+1}}\mathbf{X}_{\hat{\mathcal{I}}_{n,k+1}})^{-1}\mathbf{X}'_{\hat{\mathcal{I}}_{n,k+1}}\mathbf{Y}$ and the residual $\hat{\mathbf{r}}_{n,k+1} := \mathbf{Y} - \mathbf{X}_{\hat{\mathcal{I}}_{n,k+1}}\hat{\gamma}_{n,k+1}$;

(d) increase k by one and continue with (b) as long as $c(\hat{\mathcal{I}}_{n,k}, n) \leq B$ is satisfied;

(e) let k_n be the number of selected groups; call the resulting submatrix of selected regressors $\mathbf{Z} := \mathbf{X}_{\hat{\mathcal{I}}_{n,k_n}}$ and $\hat{\gamma}_n := \hat{\gamma}_{n,k_n}$, respectively.

Step 3. Set n to the next sample size in $\mathcal{N}$, and go to Step 2 until (and including) $n = n_K$.

Step 4. Set $\hat{n}$ as the sample size that minimizes the residual variance:

$$\hat{n} := \arg\min_{n\in\mathcal{N}} \frac{1}{nN}\sum_{i=1}^{N}\left(Y_i - \mathbf{Z}_i\hat{\gamma}_n\right)^2.$$

The algorithm above produces the selected sample size $\hat{n}$, the selection of covariates $\hat{\mathcal{I}} := \hat{\mathcal{I}}_{\hat{n},k_{\hat{n}}}$ with $k_{\hat{n}}$ selected groups and $\hat{m} := m(\hat{n}) := |\hat{\mathcal{I}}_{\hat{n},k_{\hat{n}}}|$ selected regressors. Here, $\hat{\gamma} := \hat{\gamma}_{\hat{n}}$ is the corresponding coefficient vector on the selected regressors Z.

Remark 1. The minimal sample size n_0 in $\mathcal{N}$ could, for example, be determined by power calculations (see, e.g. Duflo, Glennerster, and Kremer, 2007; McConnell and Vera-Hernández, 2015) that guarantee a certain power level for an hypothesis test of $\beta = 0$.

Remark 2. In Appendix A, we provide theoretical properties of the OGA approximation. Theorem A.1 in Appendix A gives a finite-sample bound on the difference between the best possible MSE for the estimator of the average treatment effect and the MSE for the OGA approximation. If we assume that the pre-experimental and experimental samples are from the same population, the pre-experimental sample size N is large, and the budget B is relatively small, then the difference between two MSEs 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 similar setting that this rate $1/k$ cannot generally be improved (see, e.g., Barron, Cohen, Dahmen, and DeVore, 2008). In this sense, we show that our proposed method has a desirable property. See Appendix A for further details.

Remark 3. There are many important reasons for collecting covariates, such as checking whether randomization was carried out properly and identifying heterogeneous treatment effects, among others. If a few covariates are essential for the analysis, we can guarantee their selection by including them in every group G_k, $k = 1, \ldots, p$.

Remark 4. In a simple model such as the one in Appendix C, the optimal combination of covariates equalizes the percent marginal contribution of an additional variable to the residual variance with the percent marginal contribution of the additional variable to the costs per interview. Step 2 of the OGA selects the next covariate as the one that has the highest predictive power independent of its cost. Outside a class of very simple models as in Appendix C, it is difficult to determine an OGA approximation to the optimum that jointly takes into account both predictive power as it requires comparison of all possible covariate combinations. In our empirical application of Section V.B, we study a case with heterogeneous costs and propose a sensitivity analysis that assesses whether the OGA solution significantly changes with perturbations of the set of potential covariates.

V The Costs of Data Collection

In this section, we discuss the specification of the cost function $c(S, n)$ that defines the budget constraint of the researcher. 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

$$c(S, n) = c_{\text{admin}}(S) + c_{\text{train}}(S, n) + c_{\text{interv}}(S, n). \tag{7}$$

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 VI, 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

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

V.A 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

$$c_{\text{admin}}(S) := \phi T(S)^{\alpha}, \tag{8}$$

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

$$c_{\text{train}}(S, n) := \kappa(n)\, T(S), \tag{9}$$

where $\kappa(n)$ is some function of the number of survey participants.[11] 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:

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

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 $\overline{n}_1, \overline{n}_2, \ldots$.

V.B 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

[11] 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.

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

$$c_{\text{interv}}(S, n) := n\eta + np\,T(S), \tag{10}$$

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 (7) with (8)–(10) when studying the impact of free day-care on child development in Section VI.A.

Remark 5. 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 6. 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).

V.C 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 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 modeled as

$$c_{\text{train}}(S, n_c, c) := \kappa(c, n_c)\,T(S). \tag{11}$$

The interaction of cluster and sample size in determining the number of required enumerators and, thus, the quantity $\kappa(c, n_c)$, complicates the modeling of this quantity relative to the case without clustering. Let $\mu(c, n_c)$ denote the number of required sur-

vey 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:

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

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 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, $\mu_c(c) := \lambda c$ for some constant λ (i.e., $\mu_c(c)$ is assumed to be linear in c), and

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

In addition, while the variable interview costs component continues to depend on the overall sample size n as in (10), 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

$$c_{\text{interv}}(S, n_c, c) := \psi(c)\eta + c n_c p\, T(S), \tag{12}$$

where $\psi(c)$ is some function of the number of clusters c.[12]

V.D 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 of the household that is visited. These tests may need to be

[12] One issue we have not yet explicitly address 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.

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

$$c_{\text{interv}}(S, n) := \mathbb{1}\Big\{ \sum_{j=1}^{M_{\text{low}}} S_j > 0 \Big\} (\psi_{\text{low}}(c, n_c)\eta_{\text{low}} + cn_c p_{\text{low}} T_{\text{low}}(S)) + \psi_{\text{high}}(c, n_c)\eta_{\text{high}} + cn_c p_{\text{high}} T_{\text{high}}(S). \quad (13)$$

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

$$c_{\text{admin}}(S) := \phi_{\text{low}} T_{\text{low}}(S)^{\alpha_{\text{low}}} + \phi_{\text{high}} T_{\text{high}}(S)^{\alpha_{\text{high}}}, \quad (14)$$

$$c_{\text{train}}(S, n) := \kappa_{\text{low}}(c, n_c)\, T_{\text{low}}(S) + \kappa_{\text{high}}(c, n_c)\, T_{\text{high}}(S). \quad (15)$$

We employ specification (7) with (12)–(15) when, in Section VI.B, we study the impact

on student learning of cash grants which are provided to schools.

VI Empirical Applications

VI.A 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 centers, most of which are located in slums, is allocated through a lottery, administered to children in the waiting lists for each day-care center.

Just before the 2008 school year, children applying for a slot at a public day-care center were put on a waiting list. At this time, children were between the ages of 0 and 3. For each center, 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 analyze this intervention as if it was 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.[13] 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

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

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: (i) variables that can be seen as final outcomes, such as questions about the development and the behavior of the children in the household; (ii) variables that can be seen as intermediate outcomes, such as labor supply, income, expenditure, and investments in children; (iii) variables for which there is an unusually large number of missing values; and (iv) 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, neighborhood, 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.

Calibration of the cost function. We specify the cost function (7) with components (8)–(10) 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.[14]

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

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

a vocabulary test. In total, this translates into a total of roughly 240 variables.[15] Appendix B provides a detailed description of all components of the calibrated cost function.

Implementation. In implementing the OGA, we take each single variable as a possible group (i.e., each group consists of a singleton set). As $\Delta(n)$ is almost constant for any reasonable sample size n that we consider, we ignore it in the optimization problem. We studentized all covariates to have variance one. 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

$$\min_{\gamma} \frac{1}{N} \sum_{i=1}^{N} (Y_i - \gamma' X_i)^2 + \lambda \sum_{j} |\gamma_j| \tag{16}$$

with a tuning parameter $\lambda > 0$. The POST-LASSO procedure runs an OLS regression of Y_i on the selected covariates (non-zero entries of γ) in (16). 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 non-zero 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 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

[15] 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.

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.

Table 4: 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 5: 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

Results. Tables 4 and 5 summarize the results of the covariate selection procedures. For the cognitive test outcome, OGA and POST-LASSO select one covariate ("$|\hat{I}|$"),[16] 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 III. The three methods perform similarly well and improve precision by about

[16]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.

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.

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 4 and 5. The first entry of this column in Table 4 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 percent, 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 Appendix F, 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 4 and 5.

Appendix D 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.

VI.B 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.[17] 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 (math test, French test, oral test, and receptive vocabulary) in the original experiment, we only consider the first one (math 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 sizes of $N = 2,280$ for the baseline math 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 variance zero when restricted to the control group. Therefore, there are $M = 143$ covariates in the second scenario.

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

Calibration of the cost function. We specify the cost function (7) with components (13)–(15) 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. Appendix B 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 6 summarizes the results of the covariate selection procedures. Panel (a) shows the results of the first scenario in which the baseline math 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 math test is used as the outcome to be predicted.

For the case where baseline math 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 discussed in Section III. 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")

Table 6: School grants (outcome: math 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

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 method 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 math test.[18] 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),

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

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 math score is the outcome to be predicted in panel (b), the budget used in the experiment increases due 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 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 selects 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, both selections of either no covariates and large sample size (panel (b)) and 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.[19]

[19]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 highest possible precision without any

As discussed in Section III.C, one may want to ensure balance of the control and 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 math outcome as a covariate. In the OGA, we can force the selection of a covariate by performing group OGA as described in Section IV, where each group contains a low-cost covariate together with the baseline math outcome. For the LASSO procedures, we simply perform the LASSO algorithms after partialing out the baseline math 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 math 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 due to our procedure.

In Appendix F, 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 6.

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 math score can be accounted for by the variance in the baseline math 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.

direct interest in the identities of the covariates that achieve this optimum.

Table 7: School grants (outcome: math 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

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, i.e. the three baseline outcomes. We reduce their prices τ_j simultaneously, all by the same factor. In panel (a) of Table 7, they all have price 0.5 times the actual price τ_j, and in panel (b) of Table 7, the factor is 0.4. All other aspects of the problem are kept identical to those in panel (b) of Table 6. In panel (b) of Table 7, the selected covariate is the same for all three selection procedures: the baseline outcome for the math test.

Panel (a) of Table 7 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 6. However, if we reduce the cost by 60% (panel (b)) then the baseline math 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 extremely 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 math score. To this end we increase the correlation of the baseline score with the follow-up score as described in Section H of the appendix.

In panels (c) and (d) of Table 7, 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 6. In panel (d), the covariate selected by OGA and POST-LASSO is the same: the baseline outcome for the math test. Panel (c) of Table 7 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 extremely high.

VII 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, Sadoff, and Wagner (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 did not consider the problem of selecting covariates.

Hahn, Hirano, and Karlan (2011) consider the design of a two-stage experiment for estimating an average treatment effect, and proposed 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 treat-

ment 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 set-up 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 very 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.

VIII Concluding Remarks

We develop data-driven methods for designing a survey in a randomized experiment using information from a pre-existing dataset. Our procedure is optimal in a 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 several other 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 (AIC) or the Bayesian information criterion (BIC)).

IX Proof of Theorem 1

Proof of Theorem 1. Consider:

$$\begin{aligned}
\min_{\hat{\beta}\in\mathbb{D}_n} R_n(\hat{\beta}) &= \min_{\hat{\beta}\in\mathbb{D}_n} \int_\Theta L_n(\theta, \hat{\beta}(\mathcal{S}_{exp}))d\mu(\theta) \\
&= \min_{\hat{\beta}\in\mathbb{D}_n} \int_\Theta E_{exp,\theta}\left[\left(\beta - \hat{\beta}(\mathcal{S}_{exp})\right)^2\right] d\mu(\theta) \\
&= \min_{\hat{\gamma}\in\Gamma\colon c(\mathcal{I}(\hat{\gamma}),n)\leq B} \int_\Theta E_{exp,\theta}\left[(\beta - b(\hat{\gamma}, \mathcal{S}_{exp}))^2\right] d\mu(\theta)
\end{aligned}$$

The homoskedastic error assumption, Assumption 1(ii), implies that conditional on $D_1, \ldots, D_n$, the estimator $b(\hat{\gamma}, \mathcal{S}_{exp})$ is unbiased and thus a straight-forward calculation shows that its finite-sample MSE (conditional on $D_1, \ldots, D_n$) is

$$\begin{aligned}
E_{exp,\theta}\left[(\beta - b(\hat{\gamma}, \mathcal{S}_{exp}))^2 \mid D_1, \ldots, D_n\right] &= Var_{exp,\theta}\left(b(\hat{\gamma}, \mathcal{S}_{exp}) \mid D_1, \ldots, D_n\right) \\
&= \frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i \mid D_i = 0\right)}{n\bar{D}_n(1-\bar{D}_n)} \\
&= \frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i\right)}{n\bar{D}_n(1-\bar{D}_n)}.
\end{aligned}$$

Therefore,

$$\int_\Theta E_{exp,\theta}\left[(\beta - b(\hat{\gamma}, \mathcal{S}_{exp}))^2\right] d\mu(\theta)$$

$$
\begin{aligned}
&= \int_{\Theta} E_{exp,\theta}\left\{E_{exp,\theta}\left[(\beta - b(\hat{\gamma}, \mathcal{S}_{exp}))^2 \mid D_1, \ldots, D_n\right]\right\} d\mu(\theta) \\
&= \int_{\Theta} E_{exp,\theta}\left\{\frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i\right)}{n\bar{D}_n(1-\bar{D}_n)}\right\} d\mu(\theta) \\
&= \int_{\Theta} \frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i\right)}{n\kappa(n)} d\mu(\theta)
\end{aligned}
$$

The usual least-squares calculation then shows that minimizing the residual variance is the same as minimizing the quadratic distance of $\hat{\gamma}$ from γ:

$$
\begin{aligned}
&\text{argmin}_{\hat{\gamma}\in\Gamma:\, c(\mathcal{I}(\hat{\gamma}),n)\leq B} \int_{\Theta} \frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i\right)}{n\kappa(n)} d\mu(\theta) \\
&\quad = \text{argmin}_{\hat{\gamma}\in\Gamma:\, c(\mathcal{I}(\hat{\gamma}),n)\leq B} \int_{\Theta} \frac{E_{exp,\theta}[((\gamma - \hat{\gamma})' X_i)^2]}{n\kappa(n)} d\mu(\theta) \\
&\quad = \text{argmin}_{\hat{\gamma}\in\Gamma:\, c(\mathcal{I}(\hat{\gamma}),n)\leq B} \int_{\Theta} \frac{(\gamma - \hat{\gamma})' E_{exp,\theta}[X_i X_i'](\gamma - \hat{\gamma})}{n\kappa(n)} d\mu(\theta) \\
&\quad = \text{argmin}_{\hat{\gamma}\in\Gamma:\, c(\mathcal{I}(\hat{\gamma}),n)\leq B} \int_{\Gamma} \frac{(\gamma - \hat{\gamma})' E_{pre,\gamma}[X_i X_i'](\gamma - \hat{\gamma})}{n\kappa(n)} d\mu_{\gamma}(\gamma)
\end{aligned}
$$

which is the desired expression. Q.E.D.

References

Attanasio, Orazio, Ricardo Paes de Barros, Pedro Carneiro, David Evans, Lycia Lima, Rosane Mendonca, Pedro Olinto, and Norbert Schady. 2014. "Free Access to Child Care, Labor Supply, and Child Development." Discussion paper.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2011. "Field Experiments with Firms." *Journal of Economic Perspectives*, 25(3), 63–82.

Banerjee, Abhijit V., Sylvain Chassang, Sergio Montero, and Erik Snowberg. 2017. "A Theory of Experimenters." NBER Working Paper No. 23867.

Banerjee, Abhijit V., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics*, 1(1), 151–78.

Barron, Andrew R., Albert Cohen, Wolfgang Dahmen, and Ronald A. DeVore. 2008. "Approximation and Learning by Greedy Algorithms." *Annals of Statistics*, 36(1), 64–94.

Belloni, Alexandre, and Victor Chernozhukov. 2013. "Least Squares after Model Selection in High-Dimensional Sparse Models." *Bernoulli*, 19(2), 521–47.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives*, 28(2), 29–50.

Bhattacharya, Debopam, and Pascaline Dupas. 2012. "Inferring Welfare Maximizing Treatment Assignment under Budget Constraints." *Journal of Econometrics*, 167(1), 168–96.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(4), 200–32.

Carneiro, Pedro, Oswald Koussihouèdé, Nathalie Lahire, Costas Meghir, and Corina Mommaerts. 2015. "Decentralizing Education Resources: School Grants in Senegal." NBER Working Paper 21063.

Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives*, 29(3), 81–98.

Davis, Geoffrey, Stéphane Mallat, and Marco Avellaneda. 1997. "Adaptive Greedy Approximations." *Constructive Approximation*, 13(1), 57–98.

Dominitz, Jeff, and Charles F. Manski. 2017. "More Data or Better Data? A Statistical Decision Problem." *Review of Economic Studies*, forthcoming.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review*, 101(5), 1739–1774.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools." *Journal of Public Economics*, 123, 92–110.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, Volume 4, ed. T. Paul Schults, and John Strauss, 3895–962. Amsterdam: Elsevier.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph Newhouse, Heidi Allen, Katherine Baicker, Oregon Health Study Group. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics*, 127(3), 1057–1106.

Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Frison, Lars, and Stuart J. Pocock. 1992. "Repeated Measures in Clinical Trials: Analysis Using Mean Summary Statistics and its Implications for Design." *Statistics in Medicine*, 11, 1685–704.

Györfi, László, Michael Kohler, Adam Krzyżak, and Harro Walk. 2002. *A Distribution-Free Theory Of Nonparametric Regression*. Springer-Verlag, New York.

Hahn, Jinyong, Keisuke Hirano, and Dean Karlan. 2011. "Adaptive Experimental Design Using the Propensity Score." *Journal of Business and Economic Statistics*, 29(1), 96–108.

Hamermesh, Daniel S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature*, 51(1), 162–72.

Huang, Junzhou, Tong Zhang, and Dimitris Metaxas. 2011. "Learning with Structured Sparsity." *Journal of Machine Learning Research*, 12, 3371–412.

Imbens, Guido W. and Donald B. Rubin 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Ing, Ching-Kang, and Tze Leung Lai. 2011. "A Stepwise Regression Method and Consistent Model Selection for High-Dimensional Sparse Linear Models." *Statistica Sinica*, 21(4), 1473–513.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review*, 105(5), 491–95.

Kasy, Maximilian. 2016. "Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead." *Political Analysis*, 24(3), 324?-338.

Kremer, Michael, Edward Miguel,, and Rebecca Thornton. 2009. "Incentives to Learn." *Review of Economics and Statistics*, 91(3), 437–456.

Li, Xinran, Peng Ding, and Donald B. Rubin. 2016. "Asymptotic Theory of Rerandomization in Treatment-Control Experiments." arXiv working paper, arXiv:1604.00698.

List, John, Sally Sadoff, and Mathis Wagner. 2011. "So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design." *Experimental Economics*, 14(4), 439–57.

List, John A. 2011. "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off." *Journal of Economic Perspectives*, 25(3), 3–16.

List, John A., and Imran Rasul. 2011. "Field Experiments in Labor Economics." In *Handbook of Labor Economics*, Volume 4A, ed. Orley Ashenfelter, and David Card, 103–228. Amsterdam: Elsevier.

McConnell, Brendon, and Marcos Vera-Hernández. 2015. "Going Beyond Simple Sample Size Calculations: A Practitioner's Guide." Institute for Fiscal Studies (IFS) Working Paper W15/17.

McKenzie, David. 2012. "Beyond Baseline and Follow-up: The Case for More T in Experiments." *Journal of Development Economics*, 99(2), 210–21.

Miguel, Edward, and Michael 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(2), 1263–1282.

Morgan, K. L., and D. B. Rubin 2015: "Rerandomization to Balance Tiers of Covariates," *Journal of the American Statistical Association*, 110(512), 1412–1421.

Natarajan, Balas K. 1995. "Sparse Approximate Solutions to Linear Systems." *SIAM Journal on Computing*, 24(2), 227–34.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-analysis Plans." *Journal of Economic Perspectives*, 29(3), 61–80.

Sancetta, Alessio. 2016. "Greedy Algorithms for Prediction." *Bernoulli*, 22(2), 1227–77.

Temlyakov, Vladimir N. 2011. *Greedy Approximation*. Cambridge: Cambridge University Press.

Tetenov, Aleksey. 2015. "An Economic Theory of Statistical Testing." Discussion Paper.

Tropp, Joel A. 2004. "Greed is Good: Algorithmic Results for Sparse Approximation." *IEEE Transactions on Information Theory*, 50(10), 2231–42.

Tropp, Joel A., and Anna C. Gilbert. 2007. "Signal Recovery from Random Measurements via Orthogonal Matching Pursuit." *IEEE Transactions on Information Theory*, 53(12), 4655–66.

Zhang, Tong. 2009. "On the Consistency of Feature Selection Using Greedy Least Squares Regression." *Journal of Machine Learning Research*, 10, 555–68.

For Online Publication

Appendix A: Theoretical Properties of the Algorithm

In this appendix, we provide theoretical properties of the OGA approximation $\hat{\mathbf{f}} := \mathbf{Z}\hat{\gamma}$. Following Barron, Cohen, Dahmen, and DeVore (2008), we define

$$\|f\|_{\mathcal{L}_1^N} := \inf\Big\{ \sum_{k=1}^{p} |\beta_k|_2 : \beta_k \in \mathbb{R}^{|G_k|} \text{ and } f = \sum_{k=1}^{p} X'_{G_k}\beta_k, \\ \text{where the elements of } X_{G_k} \text{ are normalized in the empirical norm } \|\cdot\|_N \Big\}.$$

When the expression $f = \sum_{k=1}^{p} X'_{G_k}\beta_k$ is not unique, we take the true f_0 to be one with the minimum value of $\|f\|_{\mathcal{L}_1^N}$. This gives $f_0 := \gamma_0' X$ and $\mathbf{f_0} := \mathbf{X}\gamma_0$ for some γ_0. Note that $\mathbf{f_0}$ is defined by $\mathbf{X}$ with the true parameter value γ_0, while $\hat{\mathbf{f}}$ is an OGA estimator of $\mathbf{f_0}$ using only $\mathbf{Z}$.

Define

$$\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}}) := \|\mathbf{Y} - \hat{\mathbf{f}}\|_N^2/\hat{n},$$

which is equal to the objective function in (5). Note that $\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})$ can also be called the "empirical risk". In addition, define

$$MSE_{\hat{n},\infty}(\mathbf{f_0}) := E_{\exp}\left[(Y - f_0)^2\right]/\hat{n}.$$

Note that $E_{\exp}[(Y - f_0)^2]$ is the counterpart of $\|\mathbf{Y} - \mathbf{f_0}\|_N^2$ using the population in the experiment. We assume that f_0 minimizes $f \mapsto E_{\exp}[(Y - f)^2]$, where $f = \gamma' X$. This implies that

$$\mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}}, \mathbf{f_0}) := MSE_{\hat{n},\infty}(\hat{\mathbf{f}}) - MSE_{\hat{n},\infty}(\mathbf{f_0})$$

is always non-negative. For each OGA step $k \geq 1$, let $\mathcal{G}_k$ denote the following class of functions

$$\mathcal{G}_k := \left\{ X_{\mathcal{I}_k} \mapsto \gamma'_{\mathcal{I}_k} X_{\mathcal{I}_k} : \gamma_{\mathcal{I}_k} \in \mathbb{R}^{|\mathcal{I}_k|} \right\}.$$

Let $R(g) := E_{\exp}[(Y - g)^2]$ for any function g. The following theorem gives a bound for $\mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}}, \mathbf{f_0})$.

Theorem A.1. *Assume that* $(\mathbf{X}'_{G_j}\mathbf{X}_{G_j})/N = \mathbf{I}_{|G_j|}$ *for each* $j = 1, \ldots, p$. *Suppose* $\mathcal{N}$ *is a finite subset of* $\mathbb{N}_+$, $c : \{0,1\}^M \times \mathbb{N}_+ \to \mathbb{R}$ *some function, and* $B > 0$ *some constant.*

Further, assume that f_0 minimizes $f \mapsto E_{exp}[(Y-f)^2]$, where $f = \gamma' X$. Then the following bound holds:

$$0 \leq \mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}}, \mathbf{f_0}) \leq T_1 + T_2,$$

where

$$T_1 := \frac{4\|f\|^2_{\mathcal{L}_1^N}}{\hat{n}} \left(\frac{1}{\min\{p, k_{\hat{n}}\}} \right),$$
$$T_2 := \frac{2}{\hat{n}} \sup_{g \in \mathcal{G}_{k_{\hat{n}}}} \left| \left\{ \|\mathbf{Y} - \mathbf{g}\|^2_N - R(g) \right\} \right|.$$

The theorem provides a bound on $\mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}}, \mathbf{f_0})$ using the OGA approximation. On one hand, the square root of $MSE_{\hat{n},\infty}(\mathbf{f_0})$ is the best possible standard error for the estimator of the average treatment effect when the sample size is $\hat{n}$. This corresponds to the oracle case that $f_0 = \gamma_0' X$ is known. On the other hand, the square root of $MSE_{\hat{n},\infty}(\hat{\mathbf{f}})$ is the standard error for the estimator of the average treatment effect using the OGA approximation. Theorem A.1 shows that the difference between these two standard errors is small for a given $\hat{n}$ if both terms T_1 and T_2 are small. For a given $\hat{n}$, T_1 gets smaller as $k_{\hat{n}}$ gets larger; however, T_2 does not decrease if $k_{\hat{n}}$ gets larger. In fact, it may get larger since the complexity of $\mathcal{G}_{k_{\hat{n}}}$ increases as $k_{\hat{n}}$ gets larger.

To understand the former term T_1 intuitively, suppose that $\|f\|_{\mathcal{L}_1^N} < \infty$ (Remark A.2 discusses this condition) and that $p < k_{\hat{n}}$ (that is, some groups of covariates are not selected by the OGA algorithm). Consider, for example, the simple case in which, for a given sample size n, data collection on every covariate incurs the same costs (i.e., $\widetilde{c}(n)$) and each group consists of a single covariate. Then the total data collection costs are equal to the number of covariates selected multiplied by $\widetilde{c}(n)$ (i.e., $c(S,n) = \widetilde{c}(n) \sum_{j=1}^{M} S_j$). Assuming that $\widetilde{c}(n)$ is non-decreasing in n, we then have

$$\frac{1}{k_{\hat{n}}} = \frac{1}{\lfloor B/\widetilde{c}(\hat{n}) \rfloor},$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than x. This shows that for a given $\hat{n}$, as the budget B increases, a larger $k_{\hat{n}}$ will be chosen and the efficiency of the estimator of the average treatment effect will improve.

We now consider the latter term T_2, which is due to the difference between the pre-experimental estimation sample and the population in the experiment. This term

will be small only if the following conditions are met: (i) the population for the pre-experimental sample and the population for the experiment need to have the same expectations for $(Y-g)^2$, (ii) the sample size N in the pre-experimental sample has to be large enough, and (iii) the complexity of $\mathcal{G}_{k_{\hat{n}}}$ is not too large. In the main text of this paper, we have assumed that the pre-experimental and experimental samples are from the same population, N is large, and the budget B is small. In this scenario, the latter term is negligible compared to the former term.

Remark A.1. Suppose that either (i) the pre-experimental sample size N is relatively small, or (ii) the budget B is large enough such that the possibility of overfitting is present. In this scenario, it might be desirable to solve a penalized version of (6) instead of solving (6). For example, in the k-th OGA step, one may solve

$$\min_{n\in\mathcal{N}} \frac{1}{n} \min_{\gamma\in\mathbb{R}^M} \left[\frac{1}{N}\sum_{i=1}^{N}(Y_i-\gamma' X_i)^2 + \kappa \frac{k\log N}{N} \right] \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma),n)\leq B, \tag{A.1}$$

where $\kappa \geq 0$ is an extra tuning parameter that needs to be determined by the researcher. Theoretical properties of this penalized OGA estimator can be obtained using the arguments similar to those used in Barron, Cohen, Dahmen, and DeVore (2008) in conjunction with the use of the truncation operator.

Remark A.2. The condition $\|f\|_{\mathcal{L}_1^N} < \infty$ is trivially satisfied when p is finite. In the case $p \to \infty$, the condition $\|f\|_{\mathcal{L}_1^N} < \infty$ requires that not all groups of covariates are equally important in the sense that the coefficients β_k, when their ℓ_2 norms are sorted in decreasing order, need to converge to zero fast enough to guarantee that $\sum_{k=1}^{\infty} |\beta_k|_2 < \infty$. If suitable laws of large numbers apply, we can also replace the condition $\|f\|_{\mathcal{L}_1^N} < \infty$ by its population counterpart.

Proofs

Proof of Theorem A.1. Write

$$\mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}},\mathbf{f_0}) = T_1 + T_2 + T_3,$$

where

$$T_1 := MSE_{\hat{n},\infty}(\hat{\mathbf{f}}) - \widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}}),$$
$$T_2 := \widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}}) - \widehat{MSE}_{\hat{n},N}(\mathbf{f_0}),$$

$$T_3 := \widehat{MSE}_{\hat{n},N}(\mathbf{f_0}) - MSE_{\hat{n},\infty}(\mathbf{f_0}).$$

Note that for each $k \geq 1$,

$$|T_1 + T_3| \leq \frac{2}{\hat{n}} \sup_{g \in \mathcal{G}_k} \left| \left\{ \|\mathbf{Y} - \mathbf{g}\|_N^2 - R(g) \right\} \right|.$$

Then the desired result follows immediately from Lemma 1, which is given below. Q.E.D.

Lemma 1. *Assume that* $(\mathbf{X}'_{G_j}\mathbf{X}_{G_j})/N = \mathbf{I}_{|G_j|}$ *for each* $j = 1, \ldots, p$. *Suppose* $\mathcal{N}$ *is a finite subset of* $\mathbb{N}_+$, $c : \{0,1\}^M \times \mathbb{N}_+ \to \mathbb{R}$ *some function, and* $B > 0$ *some constant. Then the following bound holds:*

$$\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}}) - \widehat{MSE}_{\hat{n},N}(\mathbf{f_0}) \leq \frac{4\|f_0\|_{\mathcal{L}_1^N}^2}{\hat{n}} \left(\frac{1}{\min\{p, k_{\hat{n}}\}} \right). \tag{A.2}$$

Proof of Lemma 1. This lemma can be proved by arguments similar to those used in the proof of Theorem 2.3 in Barron, Cohen, Dahmen, and DeVore (2008). The main difference between our Lemma 1 and Theorem 2.3 of Barron, Cohen, Dahmen, and DeVore (2008) is that we pay explicit attention to the group structure. In the subsequent arguments, we fix n and leave indexing by n implicit.

First, letting $\hat{\mathbf{r}}_{k-1,i}$ denote the ith component of $\hat{\mathbf{r}}_{k-1}$, we have

$$\begin{aligned}
\|\hat{\mathbf{r}}_{k-1}\|_N^2 &= N^{-1} \sum_{i=1}^N \hat{\mathbf{r}}_{k-1,i} Y_i \\
&= N^{-1} \sum_{i=1}^N \hat{\mathbf{r}}_{k-1,i} U_i + N^{-1} \sum_{i=1}^N \hat{\mathbf{r}}_{k-1,i} \sum_{j=1}^\infty X'_{G_j,i} \beta_j \\
&\leq \|\hat{\mathbf{r}}_{k-1}\|_N \Big\| \mathbf{Y} - \sum_{k=1}^\infty \mathbf{X}'_{G_k} \beta_k \Big\|_N + \Big[\sum_{j=1}^\infty |\beta_j|_2 \Big] N^{-1} |\hat{\mathbf{r}}'_{k-1} \mathbf{X}_{G_k}|_2 \\
&\leq \frac{1}{2} \Big(\|\hat{\mathbf{r}}_{k-1}\|_N^2 + \Big\| \mathbf{Y} - \sum_{k=1}^\infty \mathbf{X}'_{G_k} \beta_k \Big\|_N^2 \Big) + \Big[\sum_{j=1}^\infty |\beta_j|_2 \Big] N^{-1} |\hat{\mathbf{r}}'_{k-1} \mathbf{X}_{G_k}|_2,
\end{aligned}$$

which implies that

$$\|\hat{\mathbf{r}}_{k-1}\|_N^2 - \Big\| \mathbf{Y} - \sum_{k=1}^\infty \mathbf{X}'_{G_k} \beta_k \Big\|_N^2 \leq 2 \Big[\sum_{j=1}^\infty |\beta_j|_2 \Big] N^{-1} |\hat{\mathbf{r}}'_{k-1} \mathbf{X}_{G_k}|_2. \tag{A.3}$$

Note that if the left-hand side of (A.3) is negative for some $k = k_0$, then the conclusion of the theorem follows immediately for all $m \geq k_0 - 1$. Hence, we assume that the left-hand side of (A.3) is positive, implying that

$$\Big(\|\hat{\mathbf{r}}_{k-1}\|_N^2 - \Big\|\mathbf{Y} - \sum_{k=1}^{\infty} \mathbf{X}'_{G_k}\beta_k\Big\|_N^2\Big)^2 \leq 4\Big[\sum_{j=1}^{\infty} |\beta_j|_2\Big]^2 N^{-2}|\hat{\mathbf{r}}'_{k-1}\mathbf{X}_{G_k}|_2^2. \tag{A.4}$$

Let P_k denote the projection matrix $P_k := \mathbf{X}_{G_k}(\mathbf{X}'_{G_k}\mathbf{X}_{G_k})^{-1}\mathbf{X}'_{G_k} = N^{-1}\mathbf{X}_{G_k}\mathbf{X}'_{G_k}$, where the second equality comes from the assumption that $(\mathbf{X}'_{G_k}\mathbf{X}_{G_k})/N = \mathbf{I}_{|G_k|}$. Hence, it follows from the fact that P_k is the projection matrix that

$$\|\hat{\mathbf{r}}_{k-1} - P_k\hat{\mathbf{r}}_{k-1}\|_N^2 = \|\hat{\mathbf{r}}_{k-1}\|_N^2 - \|P_k\hat{\mathbf{r}}_{k-1}\|_N^2. \tag{A.5}$$

Because $\hat{\mathbf{r}}_k$ is the best approximation to $\mathbf{Y}$ from $\hat{\mathcal{I}}_{n,k}$, we have

$$\|\hat{\mathbf{r}}_k\|_N^2 \leq \|\hat{\mathbf{r}}_{k-1} - P_k\hat{\mathbf{r}}_{k-1}\|_N^2. \tag{A.6}$$

Combining (A.6) with (A.5) and using the fact that $P_k^2 = P_k$, we have

$$\begin{aligned} \|\hat{\mathbf{r}}_k\|_N^2 &\leq \|\hat{\mathbf{r}}_{k-1}\|_N^2 - \|P_k\hat{\mathbf{r}}_{k-1}\|_N^2 \\ &= \|\hat{\mathbf{r}}_{k-1}\|_N^2 - \|N^{-1}\mathbf{X}_{G_k}\mathbf{X}'_{G_k}\hat{\mathbf{r}}_{k-1}\|_N^2 \\ &= \|\hat{\mathbf{r}}_{k-1}\|_N^2 - N^{-2}|\hat{\mathbf{r}}'_{k-1}\mathbf{X}_{G_k}|_2^2, \end{aligned} \tag{A.7}$$

Now, combining (A.7) and (A.4) together yields

$$\|\hat{\mathbf{r}}_k\|_N^2 \leq \|\hat{\mathbf{r}}_{k-1}\|_N^2 - \frac{1}{4}\Big(\|\hat{\mathbf{r}}_{k-1}\|_N^2 - \Big\|\mathbf{Y} - \sum_{k=1}^{\infty} \mathbf{X}'_{G_k}\beta_k\Big\|_N^2\Big)^2\Big[\sum_{j=1}^{\infty} |\beta_j|_2\Big]^{-2}. \tag{A.8}$$

As in the proof of Theorem 2.3 in Barron, Cohen, Dahmen, and DeVore (2008), let $a_k := \|\hat{\mathbf{r}}_k\|_N^2 - \|\mathbf{Y} - \sum_{k=1}^{\infty} \mathbf{X}'_{G_k}\beta_k\|_N^2$. Then (A.8) can be rewritten as

$$a_k \leq a_{k-1}\Big(1 - \frac{a_{k-1}}{4}\Big[\sum_{j=1}^{\infty} |\beta_j|_2\Big]^{-2}\Big). \tag{A.9}$$

Then the induction method used in the proof of Theorem 2.1 in Barron, Cohen, Dahmen, and DeVore (2008) gives the desired result, provided that $a_1 \leq 4[\sum_{j=1}^{\infty} |\beta_j|_2]^2$. As discussed at the end of the proof of Theorem 2.3 in Barron, Cohen, Dahmen, and DeVore (2008), the initial condition is satisfied if $a_0 \leq 4[\sum_{j=1}^{\infty} |\beta_j|_2]^2$. If not, we have

that $a_0 > 4[\sum_{j=1}^{\infty} |\beta_j|_2]^2$, which implies that $a_1 < 0$ by (A.9). Hence, in this case, we have that $\|\hat{\mathbf{r}}_1\|_N^2 \leq \|\mathbf{Y} - \sum_{k=1}^{\infty} \mathbf{X}'_{G_k}\beta_k\|_N^2$ for which there is nothing else to prove.

Then, we have proved that the error of the group OGA satisfies

$$\|\hat{\mathbf{r}}_m\|_N^2 \leq \Big\|\mathbf{Y} - \sum_{k=1}^{p} \mathbf{X}'_{G_k}\beta_k\Big\|_N^2 + \frac{4}{m}\Big[\sum_{j=1}^{p} |\beta_j|_2\Big]^2, \; m = 1, 2, \ldots.$$

Equivalently, we have, for any $n \in \mathcal{N}$ and any $k \geq 1$,

$$\|\mathbf{Y} - \hat{\mathbf{f}}_{n,k}\|_N^2 - \|\mathbf{Y} - \mathbf{f_0}\|_N^2 \leq \frac{4\|f_0\|_{\mathcal{L}_1^N}^2}{k}.$$

Because $\mathcal{N}$ is a finite set, the desired result immediately follows by substituting in the definition of $\hat{\mathbf{f}}$ and $k_{\hat{n}}$. Q.E.D.

Appendix B: Cost Functions Used in Section VI

In this appendix, we provide detailed descriptions of the cost functions used in Section VI.

Calibration of the Cost Function in Section VI.A

Here, we give a detailed description of components of the cost function used in Section VI.A.

- **Administration costs.** The administration costs in the survey were R$10,000 and the average survey took two hours per household to conduct (i.e., $T(S) = 120$ measured in minutes). Therefore,

  $$c_{\text{admin}}(S, n) = \phi(120)^{\alpha} = 10{,}000.$$

  If we assume that, say, $\alpha = 0.4$ (which means that the costs of 60 minutes are about 75.8 percent of the costs of 120 minutes), we obtain $\phi \approx 1{,}473$.

- **Training costs.** The training costs in the survey were R$25,000, that is,

  $$c_{\text{train}}(S, n) = \kappa(1,466) \cdot 120 = 25{,}000,$$

so that $\kappa(1,466) \approx 208$. It is reasonable to assume that there exists some lumpiness in the training costs. For example, there could be some indivisibility in hotel rooms that are rented, and in the number of trainers required for each training session. To reflect this lumpiness, we assume that

$$\kappa(n) = \begin{cases} 150 & \text{if} \quad 0 < n \leq 1{,}400 \\ 208 & \text{if} \quad 1{,}400 < n \leq 3{,}000 \\ 250 & \text{if} \quad 3{,}000 < n \leq 4{,}500 \\ 300 & \text{if} \quad 4{,}500 < n \leq 6{,}000 \\ 350 & \text{if} \quad 6{,}000 < n \end{cases}.$$

Note that, in this specification, $\kappa(1{,}466) \approx 208$, as calculated above. We take this as a point of departure to calibrate $\kappa(n)$. Increases in sample size n are likely to translated into increases in the required number of field workers for the survey, which in turn lead to higher training costs. Our experience in the field (based on running surveys in different settings, and on looking at different budgets for different versions of this same survey) suggests that, in our example, there is some concavity in this cost function, because an increase in the sample size, in principle, will not require a proportional increase in the number of interviewers, and an increase in the number of interviewers will probably require a less than proportion increase in training costs. For example, we assume that a large increase in the size of the sample, from 1,500 to 6,000, leads to an increase in $\kappa(n)$ from 208 to 300 (i.e., an increase in overall training costs of about 50 percent).

- **Interview costs.** Interview costs were R$630,000, accounting for the majority of the total survey costs, that is,

$$c_{\text{interv}}(S, n) = 1{,}466 \cdot \eta + 1{,}466 \cdot p \cdot 120 = 630{,}000,$$

so that $\eta + 120p \approx 429.74$. The costs of traveling to each household in this survey were approximately half of the total costs of each interview. If we choose $\eta = 200$, then the fixed costs η amount to about 47 percent of the total interview costs, which is consistent with the actual costs of the survey. Then we obtain the price per unit of survey time as $p \approx 1.91$. It is also reasonable to assume that half of the variable costs per individual are due to the collection of the three outcomes in the survey, because their administration was quite lengthy. The costs of collecting the outcomes could also be seen as fixed costs (equal to $0.955 \times 120 = 114.6$),

which means that the price per unit of survey time for each of the remaining covariates is about 0.955. In sum, we can rewrite interview costs as

$$c_{\text{interv}}(S, n) = 1{,}466 \times (200 + 114.6) + 1,466 \times 0.955 \times 120 = 630{,}000.$$

- **Price per covariate.** We treat the sample obtained from the original experiment as $\mathcal{S}_{\text{pre}}$, a pilot study or the first wave of a data collection process, based on which we want to decide which covariates and what sample size to collect in the next wave. We perform the selection procedure for each outcome variable separately, and thus adjust $T(S) = \tau(1 + \sum_{j=1}^{M} S_j)$. For simplicity, we assume that to ask each question on the questionnaire takes the same time, so that $\tau_0 = \tau_j = \tau$ for every question; therefore, $T(S) = \tau(1 + \sum_{j=1}^{M} S_j) = 120$. Note that we set $\tau_0 = \tau$ here, but the high costs of collecting the outcome variables are reflected in the specification of η above. This results in $\tau = 120/(1 + \sum_{j=1}^{M} S_j)$. The actual number of covariates collected in the experiment was 40; so $\sum_{j=1}^{M} S_j = 40$, and thus $\tau \approx 3$.

- **Rescaled budget.** Because we use only a subsample of the original experimental sample, we also scale down the original budget of R\$665,000 down to R\$569,074, which corresponds to the costs of selecting all 36 covariates in the subsample; that is, $c(\mathbf{1}, 1{,}330)$ where $\mathbf{1}$ is a 36-dimensional vector of ones and $c(S, n)$ is the calibrated cost function.

Calibration of the Cost Function in Section VI.B

Here, we present a detailed description of components of the cost function used in Section VI.B.

- **Administration costs.** The administration costs for the low- and high-cost covariates were estimated to be about \$5,000 and \$24,000, respectively. The high-cost covariates were four tests that took about 15 minutes each (i.e., $T_{\text{high}}(S) = 60$). For the low-cost covariates (teacher and principal survey), the total survey time was around 60 minutes, so $T_{\text{low}}(S) = 60$. High- and low-cost variables were collected by two different sets of enumerators, with different levels of training and skills. Therefore,

$$\phi_{\text{low}}(60)^{\alpha_{\text{low}}} = 5{,}000 \qquad \text{and} \qquad \phi_{\text{high}}(60)^{\alpha_{\text{high}}} = 24{,}000.$$

If we assume that, say, $\alpha_{\text{low}} = \alpha_{\text{high}} = 0.7$, we obtain $\phi_{\text{low}} \approx 285$ and $\phi_{\text{high}} \approx$ 1,366.

- **Training costs.** μ_{high} and μ_{low} are the numbers of enumerators collecting high- and low-cost variables, respectively. The training costs for enumerators in the high and low groups increase by 20 for each set of additional 20 low-cost enumerators, and by 12 for each set of 4 high-cost enumerators:

$$\kappa_{\text{low}}(c, n_c) := 20 \sum_{k=1}^{19} k \cdot \mathbb{1}\{20(k-1) < \mu_{\text{low}}(c, n_c) \leq 20k\}$$

and

$$\kappa_{\text{high}}(c, n_c) := 12 \sum_{k=1}^{17} k \cdot \mathbb{1}\{4(k-1) < \mu_{\text{high}}(c, n_c) \leq 4k\}.$$

This is reasonable because enumerators for low-cost variables can be trained in large groups (i.e., groups of 20), while enumerators for high-cost variables need to be trained in small groups (i.e., groups of 4). However, training a larger group demands a larger room, and, in our experience, more time in the room. The lumpiness comes from the costs of hotel rooms and the time of the trainers. The numbers 20 and 12 as the average costs of each cluster of enumerators were chosen based on our experience with this survey (even if the design of the training and the organization of the survey was not exactly the same as the stylized version presented here), and reflect both the time of the trainer and the costs of hotel rooms for each type of enumerators. Because the low-cost variables are questionnaires administered to principals and teachers, in principle the number of required enumerators only depends on c (i.e., $\mu_{\text{low}}(c, n_c) = \lfloor \lambda_{\text{low}} c \rfloor$). High-cost variables are collected from students, and therefore the number of required enumerators should depend on c and n_c, so $\mu_{\text{high}}(c, n_c) = \lfloor \lambda_{\text{high}} c \mu_{n,\text{high}}(n_c) \rfloor$. We assume that the latter increases again in steps, in this case of 10 individuals per cluster, that is,

$$\mu_{n,\text{high}}(n_c) := \sum_{k=1}^{7} k \cdot \mathbb{1}\{10(k-1) < n_c \leq 10k\}.$$

We let $\lambda_{\text{low}} = 0.14$ (capturing the idea that one interviewer could do about seven schools) and $\lambda_{\text{high}} = 0.019$ (capturing the idea that one enumerator could perhaps work with about 50 children). The training costs in the survey were \$1,600 for

the low-cost group of covariates and \$1,600 for the high-cost group of covariates.

- **Interview costs.** We estimate that interview costs in the survey were \$150,000 and \$10,000 for the high- and low-cost variables, respectively, i.e.

$$\psi_{\text{low}}(350)\eta_{\text{low}} + 350 \cdot p_{\text{low}} \cdot 60 = 10{,}000$$

and

$$\psi_{\text{high}}(350, 24)\eta_{\text{high}} + 350 \cdot 24 p_{\text{high}} \cdot 60 = 150{,}000.$$

We set $\psi_{\text{low}}(c) = \mu_{\text{low}}(c)$ and $\psi_{\text{high}}(c, n_c) = \mu_{\text{high}}(c, n_c)$, the number of required enumerators for the two groups, so that η_{low} and η_{high} can be interpreted as fixed costs per enumerator. From the specification of $\mu_{\text{low}}(c)$ and $\mu_{\text{high}}(c, n_c)$ above, we obtain $\mu_{\text{low}}(350) = 50$ and $\mu_{\text{high}}(350, 24) = 20$. The fixed costs in the survey were about $\psi_{\text{low}}(350)\eta_{\text{low}} = 500$ and $\psi_{\text{high}}(350, 24)\eta_{\text{high}} = 1{,}000$ for low- and high-cost covariates. Therefore, $\eta_{\text{low}} = 500/50 = 10$ and $\eta_{\text{high}} = 1{,}000/20 = 50$. Finally, we can solve for the prices $p_{\text{low}} = (10{,}000 - 500)/(350 \times 60) \approx 0.45$ and $p_{\text{high}} = (150{,}000 - 1{,}000)/(350 \times 24 \times 60) \approx 0.3$.

- **Price per covariate.** For simplicity, we assume that to ask each low-cost question takes the same time, so that $\tau_j = \tau_{\text{low}}$ for every low-cost question (i.e., $j = 1, \ldots, M_{\text{low}}$), and that each high-cost question takes the same time (i.e., $\tau_j = \tau_{\text{high}}$) for all $j = M_{\text{low}} + 1, \ldots, M$. The experimental budget contains funding for the collection of one outcome variable, the high-cost test results at follow-up, and three high-cost covariates at baseline. We modify $T_{\text{high}}(S)$ accordingly: $T_{\text{high}}(S) = \tau_{\text{high}}(1 + \sum_{j=M_{\text{low}}+1}^{M} S_j) = 4\tau_{\text{high}}$ so that $\tau_{\text{high}} = 60/4 = 15$. Similarly, originally there were 255 low-cost covariates, which leads to $\tau_{\text{low}} = 120/255 \approx 0.47$.

- **Rescaled budget.** As in the previous subsection, we use only a subsample of the original experimental sample. Therefore, we scale down the original budget to the amount that corresponds to the costs of collecting all covariates used in the subsample. As a consequence, the rescaled budget is \$25,338 in the case of baseline outcomes and \$33,281 in the case of the follow-up outcomes.

Appendix C: A Simple Formulation of the Problem

Uniform Cost per Covariate

Take the following simple example where: (1) all covariates are orthogonal to each other; (2) all covariates have the same price, and the budget constraint is just $B = nk$, where n is sample size and k is the number of covariates. Order the covariates by the contribution to the MSE, so that the problem is to choose the first k covariates (and the corresponding n).

Define $\sigma^2(k) = (1/N)\sum_{i=1}^{N}(Y_i - \gamma'_{0,k}X_i)^2$, where $\gamma_{0,k}$ is the same as the vector of true coefficients γ_0 except that all coefficients after the $(k+1)$th coefficient are set to be zeros, and let $MSE(k,n) = (1/n)\sigma^2(k)$. For the convenience of using simple calculus, suppose that k is continuous, ignoring that k is a positive integer, and that $\sigma^2(k)$ is twice continuously differentiable. This would be a reasonable first-order approximation when there are a large number of covariates, which is our set-up in the paper. Because we ordered the covariates by the magnitude of their contribution to a reduction in the MSE, we have $\partial\sigma^2(k)/\partial k < 0$, and $\partial^2\sigma^2(k)/\partial k^2 > 0$.

The problem we solve in this case is just

$$\min_{n,k} \frac{1}{n}\sigma^2(k) \quad \text{s.t.} \quad nk \leq B.$$

Assume we have an interior solution and that n is also continuous. Replace the budget constraint in the objective function and we obtain

$$\min_{n,k} \frac{k}{B}\sigma^2(k).$$

This means that k is determined by

$$\sigma^2(k) + k\frac{\partial\sigma^2(k)}{\partial k} = 0,$$

or

$$\frac{\sigma^2(k)}{k} + \frac{\partial\sigma^2(k)}{\partial k} = 0, \tag{C.1}$$

which in this particular case does not depend on B. Then, n is given by the budget constraint (i.e., $n = B/k$).

Another way to see where this condition comes from is just to start from the budget

constraint. If we want to always satisfy it then, starting from a particular choice of n and k yields

$$n \cdot dk + k \cdot dn = 0,$$

or

$$\frac{dn}{dk} = -\frac{n}{k}.$$

Now, suppose we want to see what happens when k increases by a small amount. In that case, keeping n fixed, the objective function falls by

$$\frac{1}{n}\frac{\partial \sigma^2(k)}{\partial k}dk.$$

This is the marginal benefit of increasing k. However, n cannot stay fixed, and needs to decrease by $(n/k)dk$ to keep the budget constraint satisfied. This means that the objective function will increase by

$$\left(-\frac{1}{n^2}\right)\sigma^2(k)\left(-\frac{n}{k}\right)dk.$$

This is the marginal cost of increasing k.

At the optimum, in an interior solution, marginal costs and marginal benefits need to balance out, so

$$\frac{1}{nk}\sigma^2(k)\,dk = -\frac{1}{n}\frac{\partial \sigma^2(k)}{\partial k}dk$$

or

$$\frac{\sigma^2(k)}{k} + \frac{\partial \sigma^2(k)}{\partial k} = 0,$$

which reproduces (C.1).

There are a few things to notice in this simple example.

(1) The marginal costs of an increase in k are increasing in $\sigma^2(k)$. This is because increases in n are more important role for the MSE when $\sigma^2(k)$ is large than when it is small.

(2) The marginal costs of an increase in k are decreasing in k. This is because when k is large, adding an additional covariate does not cost much in terms of reductions in n.

(3) A large n affects the costs and benefits of increasing k in similar way. Having a large n reduces benefits of additional covariates because it dilutes the decrease in

$\sigma^2(k)$. Then, on one hand, it increases costs through the budget constraint, as a larger reduction in n is needed to compensate for the same change in k. However, on the other hand, it reduces costs, because when n is large, a particular reduction in n makes much less difference for the MSE than in the case where n is small.

(4) We can rewrite this condition as

$$\frac{1}{k} + \frac{\partial \sigma^2(k)/\partial k}{\sigma^2(k)} = 0,$$

where the term $(\partial \sigma^2(k)/\partial k)/\sigma^2(k)$ is the percentage change in the unexplained variance from an increase in k.

If we combine

$$\frac{dn}{n} = \frac{dk}{k},$$

which comes from the budget constraint, and

$$\frac{1}{MSE(n,k)} \frac{\partial MSE(n,k)}{\partial n} = -\frac{1}{n},$$

we notice that the percentage decrease in MSE from an increase in n is just $(dn)/n$, the percentage change in n, which in turn is just equal to $(dk)/k$. So what the condition above says is that we want to equate the percentage change in the unexplained variance from a change in k to the percentage change in the MSE from the corresponding change in n.

Perhaps even more interesting is to notice that k is the survey cost per individual in this very simple example. Then this condition says that we want to choose k to equate the percentage change in the survey costs per individual ($(dk)/k$) to the percentage change in the residual variance

$$\frac{\partial \sigma^2(k)/\partial k}{\sigma^2(k)} dk.$$

This condition explicitly links the impacts of k on the survey costs and on the reduction in the MSE.

Adding fixed costs F of visiting each individual is both useful and easy in this very simple framework. Suppose there are a fixed costs F of going to each individual, so the budget constraint is $n(F+k) = B$. Proceeding as above, we can rewrite our problem as

$$\min_{n,k} \frac{F+k}{B} \sigma^2(k).$$

This means that k is determined by

$$\sigma^2(k) + (F+k)\frac{\partial \sigma^2(k)}{\partial k} = 0,$$

or

$$\frac{1}{F+k} + \frac{\partial \sigma^2(k)/\partial k}{\sigma^2(k)} = 0.$$

Note that, when there are large fixed costs of visiting each individual, increasing k is not going to be that costly at the margin. It makes it much easier to pick a positive k. However, other than that, the main lessons (1)–(4) of this simple model remain unchanged.

Variable Cost per Covariate

If covariates do not have uniform costs, then the problem is much more complicated. Consider again a simple set-up where all the regressors are orthogonal, and we order them by their contribution to the MSE. However, suppose that the magnitude of each covariate's contribution the MSE takes a discrete finite number of values. Let $\mathcal{R}$ denote the set of these discrete values. Let r denote an element of $\mathcal{R}$ and $R = |\mathcal{R}|$ (the total number of all elements in $\mathcal{R}$). There are many potential covariates within each r group, each with a different price p. The support of p could be different for each r. So, within each r, we will then order variables by p. The problem will be to determine the optimal k for each r group. Let $\mathbf{k} \equiv \{k_r : r \in \mathcal{R}\}$.

The problem is

$$\min_{n,\mathbf{k}} \frac{1}{n}\sigma^2(\mathbf{k}) \quad \text{s.t.} \sum_{r\in\mathcal{R}} c_r(k_r) \leq B,$$

where $c_r(k_r) = \sum_{l=1}^{k_r} p_l$ are the costs of variables of type r used in the survey. We can also write it as $c_r(k_r) = p_r(k_r)\,k_r$, where $p_r(k_r) = (\sum_{l=1}^{k_r} p_l)/k_r$. Because we order the variables by price (from low to high), $\partial p_r(k_r)/\partial k_r > 0$. Let $\sigma_r^2 = \partial\sigma^2(\mathbf{k})/\partial k_r$, which is a constant (this is what defines a group of variables).

Then, assume we can approximate $p_l(k_r)$ by a continuous function and that we have an interior solution. Then, substituting the budget constraint in the objective function:

$$\min_{n,\mathbf{k}} \frac{1}{B}\left[\sum_{r\in\mathcal{R}} c_r(k_r)\right]\sigma^2(\mathbf{k}).$$

From the first-order condition for k_r,

$$\frac{\partial c_r(k_r)}{\partial k_r}\sigma^2(\mathbf{k}) + \left[\sum_{r\in\mathcal{R}} c_r(k_r)\right]\frac{\partial \sigma^2(\mathbf{k})}{\partial k_r} = 0,$$

or

$$\frac{\partial c_r(k_r)/\partial k_r}{\sum_{r\in\mathcal{R}} c_r(k_r)} = -\frac{\partial \sigma^2(\mathbf{k})/\partial k_r}{\sigma^2(\mathbf{k})}.$$

What this says is that, for each r, we choose variables up to the point where the percent marginal contribution of the additional variable to the residual variance equals the percent marginal contribution of the additional variable to the costs per interview, just as in the previous subsection.

Appendix D: Simulations

In this appendix, we study the finite sample behavior of our proposed data collection procedure, and compare its performance to other variable selection methods. We consider the linear model from above, $Y = \gamma' X + \varepsilon$, and mimic the data-generating process in the day-care application of Section VI.A with the cognitive test outcome variable.

First, we use the dataset to regress Y on X. Call the regression coefficients $\hat{\gamma}_{\text{emp}}$ and the residual variance $\hat{\sigma}^2_{\text{emp}}$. Then, we regress Y on the treatment indicator to estimate the treatment effect $\hat{\beta}_{\text{emp}} = 0.18656$. We use these three estimates to generate Monte Carlo samples as follows. For the pre-experimental data $\mathcal{S}_{\text{pre}}$, we resample X from the empirical distribution of the $M = 36$ covariates in the dataset and generate outcome variables by $Y = \gamma' X + \varepsilon$, where $\varepsilon \sim N(0, \hat{\sigma}^2_{\text{emp}})$ and

$$\gamma = \hat{\gamma}_{\text{emp}} + \frac{1}{2}\text{sign}(\hat{\gamma}_{\text{emp}})\kappa\bar{\gamma}.$$

We vary the scaling parameter $\kappa \in \{0, 0.3, 0.7, 1\}$ and $\bar{\gamma} := (\bar{\gamma}_1, \ldots, \bar{\gamma}_{36})'$ is specified in three different fashions, as follows:

- "lin-sparse", where the first five coefficients linearly decrease from 3 to 1, and all others are zero, that is,

$$\bar{\gamma}_k := \begin{cases} 3 - 2(k-1)/5, & 1 \le k \le 5 \\ 0, & \text{otherwise} \end{cases};$$

- "lin-exp", where the first five coefficients linearly decrease from 3 to 1, and the remaining decay exponentially, that is,

$$\bar{\gamma}_k := \begin{cases} 3 - 2(k-1)/5, & 1 \leq k \leq 5 \\ e^{-k}, & k > 5 \end{cases} ;$$

- "exp", where exponential decay $\bar{\gamma}_k := 10e^{-k}$.

When $\kappa = 0$, the regression coefficients γ are equal to those in the empirical application. When $\kappa > 0$, we add one of the three specifications of $\bar{\gamma}$ to the coefficients found in the dataset, thereby increasing (in absolute value) the first few coefficients[20]

[20]Because all estimated coefficients in the dataset ($\hat{\gamma}_{\text{emp}}$) are close to zero and roughly of the same magnitude, we simply pick the first five covariates that have the highest correlation with the outcome variable.

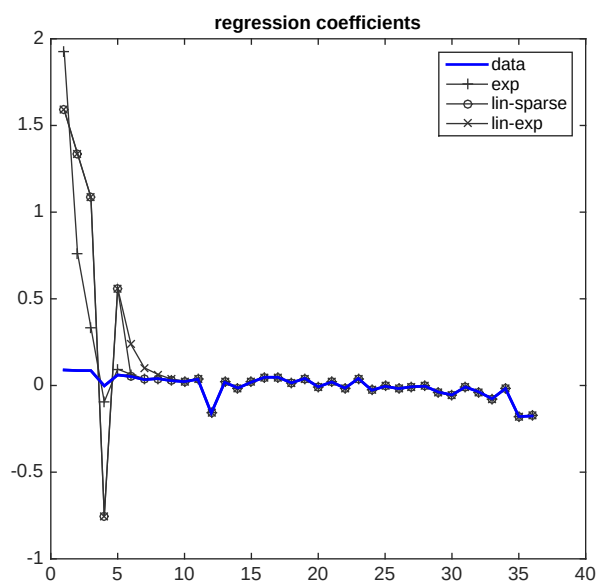

Figure 1: Regression coefficients in the simulation when $\kappa = 0.3$

more than the others, and thus increasing the importance of the corresponding regressors for prediction of the outcome. Figure 1 displays the regression coefficients in the dataset (i.e., when $\kappa = 0$, denoted by the blue line labeled "data"), and γ for the three different specifications when $\kappa = 0.3$.

For each Monte Carlo sample from $\mathcal{S}_{\text{pre}}$, we apply the OGA, LASSO, and POST-LASSO methods, as explained in Section VI.A. The cost function and budget are specified exactly as in the empirical application. We store the sample size and covariate selection produced by each of the three procedures, and then mimic the randomized experiment $\mathcal{S}_{\text{exp}}$ by first drawing a new sample of X from the same data-generating process as in $\mathcal{S}_{\text{pre}}$. Then we generate random treatment indicators D, so that outcomes are determined by

$$Y = \hat{\beta}_{\text{emp}} D + \gamma' X + \varepsilon,$$

where ε is randomly drawn from $N(0, \hat{\sigma}^2_{\text{emp}})$. We then compute the treatment effect estimator $\hat{\beta}$ of β as described in Step 4 of Section III.

The results are based on 500 Monte Carlo samples, $N = 1{,}330$, which is the sample size in the dataset, and $\mathcal{N}$ a fine grid from 500 to 4,000. All covariates, those in the dataset as well as the simulated ones, are studentized so that their variance is equal to one.

For the different specifications of $\bar{\gamma}$, Tables D.1–D.3 report the selected sample size ($\hat{n}$), the selected number of covariates ($|\hat{I}|$), the ratio of costs for that selection divided by the budget B, the square root of the estimated MSE, $\sqrt{\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})}$, the bias and standard deviation of the estimated average treatment effect ($bias(\hat{\beta})$ and $sd(\hat{\beta})$), and the RMSE of $\hat{\beta}$ across the Monte Carlo samples of the experiment.

Overall, all three methods perform similarly well across different designs and the number of selected covariates tends to increase as κ becomes large. No single method dominates other methods, although POST-LASSO seems to perform slightly better than LASSO. In view of the Monte Carlo results, we argue that the empirical findings reported in Section VI.A are likely to result from the lack of highly predictive covariates in the empirical example.

Table D.1: Simulation results: lin-sparse

Scale	Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	$\sqrt{\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})}$	$bias(\hat{\beta})$	$sd(\hat{\beta})$	RMSE($\hat{\beta}$)	EQB
0	Experiment	1,330	36	1	0.02498	−0.0034284	0.049981	0.050048	$56,9074
	OGA	2,508	1.4	0.99543	0.019249	0.00034598	0.038838	0.038801	$34,8586
	LASSO	2,587	0.1	0.99278	0.019418	0.0020874	0.039372	0.039388	$35,6781
	POST-LASSO	2,529	1.0	0.99457	0.019222	0.00069394	0.037758	0.037727	$35,1659
0.3	Experiment	1,330	36	1	0.02494	0.00036992	0.049501	0.049453	$56,9074
	OGA	2,350	3.9	0.99443	0.019751	0.0013905	0.038275	0.038262	$37,7076
	LASSO	2,228	5.9	0.988	0.020346	−0.00093089	0.041713	0.041682	$39,3017
	POST-LASSO	2,320	4.4	0.99321	0.019696	−0.0020751	0.038746	0.038763	$37,1730
0.7	Experiment	1,330	36	1	0.024953	−0.00086563	0.050992	0.050948	$56,9074
	OGA	2,346	4.0	0.99433	0.020552	−0.0020151	0.041475	0.041483	$39,7722
	LASSO	2,218	6.1	0.98747	0.021145	0.00057516	0.043957	0.043917	$42,3971
	POST-LASSO	2,246	5.7	0.98929	0.020177	0.0019693	0.042683	0.042686	$38,7095
1	Experiment	1,330	36	1	0.024938	−0.0021535	0.051146	0.05114	$56,9074
	OGA	2,346	4.0	0.99433	0.021566	0.00044162	0.043389	0.043348	$43,8536
	LASSO	2,172	6.9	0.98513	0.021383	−0.00058106	0.045378	0.045336	$43,1589
	POST-LASSO	2,172	6.9	0.98513	0.019956	−0.0048726	0.040967	0.041215	$38,0053

Table D.2: Simulation results: lin-exp

Scale	Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	$\sqrt{\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})}$	$bias(\hat{\beta})$	$sd(\hat{\beta})$	RMSE($\hat{\beta}$)	EQB
0	Experiment	1,330	36	1	0.024965	0.0027033	0.051564	0.051583	\$569,074
	OGA	2,509	1.3	0.99541	0.019249	−0.00042961	0.03723	0.037195	\$348,682
	LASSO	2,588	0.1	0.99275	0.01941	−0.003374	0.03845	0.03856	\$357,261
	POST-LASSO	2,530	1.0	0.9946	0.019215	0.00076956	0.037924	0.037894	\$351,755
0.3	Experiment	1,330	36	1	0.02492	−0.0015645	0.049457	0.049432	\$569,074
	OGA	2,343	4.0	0.99421	0.019868	−0.0014349	0.040197	0.040182	\$379,540
	LASSO	2,186	6.7	0.98569	0.020652	0.0019377	0.04084	0.040845	\$403,004
	POST-LASSO	2,313	4.5	0.99288	0.019816	−0.0025812	0.039587	0.039631	\$377,876
0.7	Experiment	1,330	36	1	0.024936	0.0041527	0.050436	0.050556	\$569,074
	OGA	2,301	4.7	0.99247	0.020805	−0.0017267	0.041303	0.041297	\$408,990
	LASSO	2,134	7.7	0.98551	0.02162	−0.00071182	0.042716	0.042679	\$440,232
	POST-LASSO	2,206	6.5	0.98955	0.020522	0.0013055	0.043358	0.043334	\$400,219
1	Experiment	1,330	36	1	0.024964	−0.0034064	0.049484	0.049551	\$569,074
	OGA	2,286	5.0	0.99187	0.021874	−0.0025106	0.042304	0.042336	\$451,756
	LASSO	2,080	9.0	0.98793	0.021987	−0.0015746	0.044218	0.044201	\$454,765
	POST-LASSO	2,078	9.0	0.98787	0.020374	0.00077488	0.041977	0.041942	\$396,218

Table D.3: Simulation results: exp

Scale	Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	$\sqrt{\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})}$	$bias(\hat{\beta})$	$sd(\hat{\beta})$	RMSE($\hat{\beta}$)	EQB
0	Experiment	1,330	36	1	0.024953	0.00083077	0.054043	0.053996	\$569,074
	OGA	2,511	1.3	0.99538	0.019234	0.0016616	0.037237	0.037236	\$348,426
	LASSO	2,588	0.1	0.99278	0.019394	−0.00049328	0.038849	0.038813	\$356,941
	POST-LASSO	2,529	1.0	0.99452	0.019203	−0.00044404	0.039549	0.039512	\$351,403
0.3	Experiment	1,330	36	1	0.024947	−0.00089522	0.051246	0.051202	\$569,074
	OGA	2,411	2.9	0.99605	0.019426	0.0016951	0.038729	0.038727	\$359,950
	LASSO	2,291	4.9	0.9911	0.020184	−0.0022094	0.040243	0.040263	\$389,560
	POST-LASSO	2,380	3.5	0.99514	0.019377	0.0014552	0.039996	0.039982	\$359,662
0.7	Experiment	1,330	36	1	0.024946	−0.0012694	0.050947	0.050912	\$569,074
	OGA	2,408	3.0	0.99605	0.019457	0.0015399	0.040789	0.040778	\$362,287
	LASSO	2,279	5.1	0.99039	0.020233	0.0011166	0.042491	0.042463	\$391,128
	POST-LASSO	2,376	3.5	0.99515	0.019405	−0.0023208	0.037252	0.037287	\$361,903
1	Experiment	1,330	36	1	0.024948	−0.0034014	0.051898	0.051957	\$569,074
	OGA	2,407	3.0	0.99603	0.019494	0.0022031	0.038846	0.038869	\$364,015
	LASSO	2,271	5.2	0.99008	0.020298	0.0016393	0.039024	0.039019	\$392,857
	POST-LASSO	2,377	3.5	0.99516	0.019448	−0.00085645	0.039135	0.039106	\$363,023

Appendix E: Variables Selected in the School Grants Example

Table E.1: School grants (outcome: math test): selected covariates in panel (a) of Table 6

OGA	LASSO	POST-LASSO
Child is male	Child is male	Child is male
Village pop.	Dist. to Dakar	Dist. to Dakar
Piped water	Dist. to city	Dist. to city
Teach-stud	Village pop.	Village pop.
No. computers	Piped water	Piped water
Req. (h) teach. qual.	No. computers	No. computers
Req. (h) teach. att.	Req. (h) teach-stud	Req. (h) teach-stud
Obs. (h) manuals	Hrs. tutoring	Hrs. tutoring
Books acq. last yr.	Books acq. last yr.	Books acq. last yr.
Any parent transfer	Provis. struct.	Provis. struct.
Teacher bacc. plus	NGO cash cont.	NGO cash cont.
Teach. train. math	Any parent transfer	Any parent transfer
Obst. (t) class size	NGO promised cash	NGO promised cash
Measure. equip.	Avg. teach. exp.	Avg. teach. exp.
	Teacher bacc. plus	Teacher bacc. plus
	Obs. (t) student will.	Obs. (t) student will.
	Obst. (t) class size	Obst. (t) class size
	Silence kids	Silence kids

Table E.2: Definition of variables in Table E.1

Variable	Definition
Child is male	Male student
Village pop.	Size of the population in the village
Piped water	School has access to piped water
Teach–stud	Teacher–student ratio in the school
No. computers	Number of computers in the school
Req. (h) teach. qual.	Principal believes teacher quality is a major requirement for school success
Req. (h) teach. att.	Principal believes teacher attendance is a major requirement for school success
Obs. (h) manuals	Principal believes the lack of manuals is a major obstacle to school success
Books acq. last yr.	Number of manuals acquired last year
Any parent transfer	Cash contributions from parents
Teacher bacc. plus	Teacher has at least a baccalaureate degree
Teach. train. math	Teacher received special training in math
Obst. (t) class size	Teacher believes class size is a major obstacle to school success
Measure. equip.	There is measurement equipment in the classroom
Dist. to Dakar	Distance to Dakar
Dist. to city	Distance to the nearest city
Req. (h) teach–stud	Principal believe teacher–student ratio is a major requirement for school success
Hrs. tutoring	Hours of tutoring by teachers
Provis. struct.	Number of provisional structures in school
NGO cash cont.	Cash contributions by NGO
NGO promised cash	Promised cash contributions by NGO
Avg. teach. exp.	Average experience of teachers in the school
Obst. (t) student will.	Teacher believes the lack of student willpower is one of the main obstacles to learning in the school
Obst. (t) class size	Teacher believes the lack of classroom size is one of the main obstacles to learning in the school
Silence kids	Teacher has to silence kids frequently

Appendix F: Out-of-Sample Evaluations

In the empirical applications, we performed the covariate selection procedure as well as its evaluation (by RMSE and EQB) on the same pre-experimental sample. In this section, we study the sensitivity of our findings when the covariate selection and evaluation steps are performed on two separate samples.

We partition each of the datasets into five subsamples of equal size. Four of the five subsamples are merged to form the training set on which we perform the covariate selection procedure, and the remaining fifth subsample serves as evaluation sample on which we calculate the performance measures RMSE and EQB. Given the partition into five subsamples, there are five possible ways to combine them into training and evaluation samples. We perform the covariate selection on each of these five training samples using the same calibrated cost functions as in the main text, but adjusting the budget for the drop in sample size by letting the budget be the cost function $c(S, n)$ evaluated at the experimental selection $S = (1, \ldots, 1)'$ and n the length of the training sample. The output of the procedure consists of five sample size selections $\hat{n}$, five covariate selections, i.e. five values of $|\hat{I}|$, and five cost-to-budget ratios. Tables E.3–E.5 show the averages of $\hat{n}$, $|\hat{I}|$, and "Cost/B" over those five different training samples. The RMSE is calculated using the estimate of γ from the training sample and data on Y and X from the evaluation sample. Similarly, the EQB is the budget necessary to achieve the RMSE on the evaluation sample equal to that of the experiment when the covariate selection procedures are applied to the training sample. Both RMSE and EQB are then averaged over the five subsamples.

Overall, the results of this out-of-sample evaluation exercise are similar to those reported in the full-sample analysis of the main text. Qualitatively, in both applications, the covariate selection procedures recommend larger sample sizes than the experiment. The recommended sample size may differ somewhat from those reported in the main text because the budget and training sample size is different, but the orders of magnitude are the same. In the school grants application, we notice that the recommended number of covariates selected tends to be smaller than in the full-sample evaluation of the main text, but if anything the covariate selection procedures manage to achieve even lower relative equivalent budgets and lower RMSE than the experiment.

Table E.3: Day-care (outcome: cognitive test), 5-fold out-of-sample evaluation

method	$\hat{n}$	$\lvert\hat{I}\rvert$	cost/B	RMSE	EQB	relative EQB
experiment	1,330	36	1	0.029068	R$460,809.54	1
OGA	2,209	0.8	0.99503	0.020694	R$235,654.21	0.511
LASSO	2,260	0	0.99392	0.020777	R$237,425.38	0.515
POST-LASSO	2,146	1.8	0.99464	0.020647	R$234,494.95	0.509

Table E.4: Day-care (outcome: health assessment), 5-fold out-of-sample evaluation

method	$\hat{n}$	$\lvert\hat{I}\rvert$	cost/B	RMSE	EQB	relative EQB
experiment	1,330	36	1	0.029313	R$460,809.54	1
OGA	2,221	0.6	0.99495	0.020708	R$232,066.95	0.504
LASSO	2,260	0	0.99392	0.020787	R$233,751.11	0.507
POST-LASSO	2,158	1.6	0.9949	0.020644	R$231,224.87	0.502

Table E.5: School grants (outcome: math test), 5-fold out-of-sample evaluation

Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	RMSE	EQB	Relative EQB
(a) Baseline outcome						
experiment	1,824	142	1	0.0082721	$27,523.74	1
OGA	2,618.4	1.2	0.99823	0.0044229	$16,609.53	0.603
LASSO	2,658	0	0.9991	0.004445	$16,621.60	0.604
POST-LASSO	2,638.2	1.2	0.99927	0.0044291	$16,651.77	0.605
(b) Follow-up outcome						
experiment	609	143	1	0.0098756	$48,856.20	1
OGA	6,132	0	0.99885	0.0028432	$14,664.85	0.300
LASSO	6,132	0	0.99885	0.0028432	$14,664.85	0.300
POST-LASSO	6,132	0	0.99885	0.0028432	$14,664.85	0.300
(c) Follow-up outcome, no high-cost covariates						
experiment	609	143	1	0.0098756	$48,856.20	1
OGA	6,092.8	0.8	0.99893	0.0027807	$14,571.10	0.298
LASSO	6,000.8	5.4	0.99905	0.002795	$14,651.75	0.300
POST-LASSO	6,040.4	2.4	0.99887	0.0027623	$14,532.12	0.297
(d) Follow-up outcome, force baseline outcome						
experiment	609	143	1	0.0098756	$48,856.20	1
OGA	2,035.2	2.4	0.90783	0.0041647	$24,918.89	0.510
LASSO	2,494	1	0.99623	0.0046439	$25,522.72	0.522
POST-LASSO	2,494	1	0.99623	0.0034893	$23,651.72	0.484

Appendix G: The Case of Multivariate Outcomes

In this section, we consider an extension to the case of multivariate outcomes. If data on a particular regressor is collected, then the regressor is automatically available for regressions involving any of the outcomes. Therefore, it is natural to select one common set of regressors for all outcomes. Hence, our regression problem corresponds to the special case of seemingly unrelated regressions (SUR) such that the vector of regressors is identical for each equation. In this case, it is well known that the OLS and GLS estimators are algebraically identical. In other words, there is no loss of efficiency in using the single-equation OLS estimator even if regression errors are correlated.

Suppose there are L outcome variables of interest, say $\{Y_{\ell,i} : \ell = 1, \ldots, L, \text{ and } i = 1, \ldots, N\}$. Then a multivariate analog of (5) can be written as

$$\min_{n \in \mathbb{N}_+, \boldsymbol{\gamma} = (\gamma_1', \ldots, \gamma_L')' \in \mathbb{R}^{ML}} \frac{1}{nNL} \sum_{\ell=1}^{L} \sum_{i=1}^{N} (Y_{\ell,i} - \gamma_\ell' X_i)^2 \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma), n) \leq B. \tag{G.1}$$

In other words, the stacked version of the OLS problem is equivalent to regressing $\boldsymbol{y} := (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_L')'$ on $I_L \otimes \boldsymbol{X}$ conditional on the budget constraint, where $\boldsymbol{y}_\ell = (Y_{\ell,1}, \ldots, Y_{\ell,L})'$, I_L is the L-dimensional identity matrix, and $\boldsymbol{X}$ is $N \times M$ dimensional matrix whose ith row is X_i'. Therefore, the OGA applies to this case as well with minor modifications. First, we need to redefine the outcome vector and the design matrix with the stacked $\boldsymbol{y}$ and the enlarged design matrix $I_L \otimes \boldsymbol{X}$. Suppose that a variable selection problem is on individual components of X_i. Then note that because of the nature of the stacked regressions, we need to apply a group OGA with each group consisting of L columns of $[I_L \otimes \boldsymbol{X}]_k$, where $k = (\ell - 1)M + m$ $(\ell = 1, \ldots, L)$ for each $m = 1, \ldots, M$.

Appendix H: Counterfactual Increases of the Predictive Power of Covariates

In the second empirical application, we increase the correlation of the baseline outcome with the follow-up outcome as follows:

- First, we run a regression of follow-up outcome Y_{2i} on baseline outcome Y_{1i}, yielding regression coefficient $\hat{\rho}$ and residual $\hat{e}_i$.

- We then increase the predictive power of the baseline outcome by multiplying $\hat{\rho}$ by a factor w_1 and the residual by w_2 to define a new follow-up outcome

$$\widetilde{Y}_{2i} = w_1 \hat{\rho} Y_{1i} + w_2 \hat{e}_i.$$

- Then the variance of the original follow-up outcome can be decomposed into an "explained" and "unexplained" part as

$$Var(Y_{2i}) = \hat{\rho}^2 Var(Y_{1i}) + Var(\hat{e}_i),$$

  and similarly for the new follow-up outcome

$$Var(\widetilde{Y}_{2i}) = w_1^2 \hat{\rho}^2 Var(Y_{1i}) + w_2^2 Var(\hat{e}_i).$$

- We choose w_2 so that the two outcomes have the same variance ($Var(Y_{2i}) = Var(\widetilde{Y}_{2i})$), i.e.

$$w_2 = \sqrt{\frac{(1 - w_1^2)\hat{\rho}^2 Var(Y_{1i})}{Var(\hat{e}_i)} + 1}.$$

- In panel (c) of Table 7, we set $w_1 = 1.2$, and in panel (d) of Table 7, $w_1 = 1.3$.