4-1-2018

Experiment-as-Market: Incorporating Welfare into Randomized Controlled Trials

Yusuke Narita

EXPERIMENT-AS-MARKET: INCORPORATING WELFARE INTO RANDOMIZED CONTROLLED TRIALS

By

Yusuke Narita

August 2019

Revised May 2019

COWLES FOUNDATION DISCUSSION PAPER NO. 2127R

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Experiment-as-Market:
Incorporating Welfare into Randomized Controlled Trials*

Yusuke Narita†

May 10, 2019

Abstract

Randomized Controlled Trials (RCTs) enroll hundreds of millions of subjects and involve many human lives. To improve subjects' welfare, I propose a design of RCTs that I call *Experiment-as-Market (EXAM)*. EXAM produces a Pareto efficient allocation of treatment assignment probabilities, is asymptotically incentive compatible for preference elicitation, and unbiasedly estimates any causal effect estimable with standard RCTs. I quantify these properties by applying EXAM to a water cleaning experiment in Kenya (Kremer et al., 2011). In this empirical setting, compared to standard RCTs, EXAM improves subjects' predicted well-being while reaching similar treatment effect estimates with similar precision.

Keywords: Social Experiment, Clinical Trial, A/B Test, Market Design, Competitive Equilibrium from Equal Income, Pareto Efficiency, Causal Inference, Development Economics

*Presentation slides for this paper are available at `https://www.dropbox.com/s/uvlgtxz45zehqtu/EXAMslide90min.pdf?dl=0`. I am grateful to Dean Karlan for a conversation that inspired this project; Jason Abaluck, Josh Angrist, Tim Armstrong, James Berry, Sylvain Chassang, Esther Duflo, Naoki Egami, Peter Hull, Costas Meghir, Bobby Pakzad-Hurson, Amanda Kowalski, Michael Kremer, Daniel Marszalec, Gerard Padro i Miquel, Joseph Moon, Kritika Narula, Rohini Pande, Parag Pathak, Mark Rosenzweig, Don Rubin, Jesse Shapiro, Joseph Shapiro, Erik Snowberg, Suk Joon Son, Seiki Tanaka, Kosuke Uetake, and Glen Weyl for criticisms and encouragement; several research assistants for their help; seminar participants at Penn, Chicago (theory and econometrics), NBER (development and market design), Columbia, Wisconsin, Tokyo, Oxford, Microsoft, AEA, RIKEN Center for Advanced Intelligence Project, Brown, Illinois, Yale, Hitotsubashi, European Summer Symposium in Economic Theory on "An Economic Perspective on the Design and Analysis of Experiments," CyberAgent, Kyoto, International Conference on Experimental Methods in Economics.

†Yale University, Department of Economics and Cowles Foundation. Email: `yusuke.narita@yale.edu`. URL: `www.yusuke-narita.com`. Address: 37 Hillhouse Avenue, New Haven, CT 06511

1 Introduction

Today is the golden age of Randomized Controlled Trials (RCTs). RCTs started as safety and efficacy tests of farming and medical treatments, but have grown to become the society-wide standard of evidence. RCTs are widespread in business and politics, as well as public policy, the social sciences, and engineering.

RCTs are high-stakes. Firstly, a large number of individuals participate in RCTs. I find that over 360 million patients and 22 million individuals participated in registered clinical trials and social RCTs, respectively, during 2007-17. Second, many RCTs randomize high-stakes treatments. For instance, in a glioblastoma therapy trial, the five-year death rate of glioblastoma patients is 97% in the control group but only 88% in the treatment group (Stupp et al., 2009). In expectation, therefore, the lives of up to 9% of the study's 573 participants depend on who receives treatments. Social RCTs also randomize critical treatments such as basic income[1], high-wage job offers (Dal Bó et al., 2013), and HIV testing (Angelucci and Bennett, 2017). This prompted some RCT participants to sue their experimenters.[2]

RCTs thus determine the fate of numerous people, giving rise to a long-standing dilemma:

> *How can a physician committed to doing what he thinks is best for each patient tell a woman with breast cancer that he is choosing her treatment by something like a coin toss? How can he give up the option to make changes in treatment according to the patient's responses?* ("Patients' Preferences in Randomized Clinical Trials" by physician Marcia Angell)

Similar concerns motivated pioneering experimental designs to incorporate participant preferences as a welfare measure into treatment assignment probabilities (Zelen, 1979; Angrist and Imbens, 1991; Chassang et al., 2012). Other prior designs respect another welfare measure, i.e., predicted treatment effects (Zelen, 1969; Wei and Durham, 1978; Hu and Rosenberger, 2003).

This paper develops an experimental design that optimally incorporates both of the two welfare criteria (preferences and predicted effects), thus further alleviating the concern with RCTs. This experimental design not only improves participant welfare, but also always unbiasedly and often precisely estimates treatment effects.

[1] "8 basic income experiments to watch out for in 2017," at `http://www.businessinsider.com/basic-income-experiments-in-2017-2017-1/#finland-2`, retrieved in May 2019.

[2] See, for example, *Gelsinger v. University of Pennsylvania* about a gene-therapy clinical trial and *Grimes v. Kennedy-Krieger Institute* about a social experiment that randomly assigned lead reduction methods to housings. For details, see `https://www.sskrplaw.com/gelsinger-v-university-of-pennsylvania.html` and `https://www.courtlistener.com/opinion/2386331/grimes-v-kennedy-krieger-institute-inc/`, accessed in May 2019.

I start by defining experimental designs as procedures that determine each subject's treatment assignment probabilities based on data about two measures of welfare: (a) the predicted treatment effect of each treatment on each subject and (b) each subject's willingness-to-pay (WTP) for each treatment. These complementary welfare measures are allowed to be heterogeneous and correlated with each other. In practice, the experimenter may estimate them from prior experimental or observational data, or ask subjects to self-report them, especially WTP.

I propose an experimental design that I call *Experiment-as-Market (EXAM)*. I choose this name because EXAM is an experimental design based on an imaginary centralized market, inspired by the idea of competitive equilibrium from equal incomes (Friedman, 1962; Hylland and Zeckhauser, 1979; Budish et al., 2013; He et al., 2017; Mollner and Weyl, 2018). EXAM uses this artificial market to Pareto optimally incorporate both predicted effects and WTP.

Specifically, EXAM randomly assigns treatments to subjects via the following hypothetical market created in the experimenter's computer. EXAM first endows each subject with a common artificial budget and lets her use the budget to purchase the most preferred (highest WTP) bundle of treatment assignment probabilities given their prices. The prices are personalized so that each treatment is cheaper for subjects with better predicted effects of the treatment. EXAM computes its treatment assignment probabilities as what subjects demand at market clearing prices, where subjects' aggregate demand for each treatment is balanced with its supply or capacity (assumed to be exogenously given). EXAM finally requires every subject to be assigned to every treatment with a positive probability.[3]

This virtual-market construction gives EXAM nice welfare and incentive properties. EXAM has a Pareto optimality property, in that no other design makes every subject better-off in terms of expected predicted effects of and WTP for assigned treatment. EXAM also allows the experimenter to elicit WTP in an asymptotically incentive compatible way. That is, when the experimenter asks subjects to self-report their WTP to be used by EXAM, every subject's optimal choice is to report her true WTP, at least for large experiments.[4]

Importantly, EXAM also allows the experimenter to unbiasedly estimate the same treatment effects as standard RCTs do (in a finite sample and for a wide class of treatment effect parameters). To see this, note that in the end, EXAM is an experiment stratified on observable predicted effects and WTP, in which the experimenter observes each subject's

[3] EXAM is executable even without WTP and predicted effects (when WTP and predicted effects are unknown or irrelevant to the experimenter). When the experimenter uses neither WTP nor predicted effects, EXAM reduces to the standard RCT. EXAM therefore nests the standard RCT.

[4] I have to be satisfied with asymptotic incentive compatibility since exact incentive compatibility is known to be incompatible with Pareto efficiency. The incentive analysis owes much to studies on the incentive compatibility of competitive equilibria and experimental designs (Jackson, 1992; Chassang et al., 2012; Azevedo and Budish, 2017; He et al., 2017).

assignment probabilities (propensity scores). As a result, EXAM's treatment assignment is random (independent from potential outcomes) conditional on the observables. The conditionally independent treatment assignment allows the experimenter to unbiasedly estimate the average treatment effects conditional on observables. By integrating such conditional effects, EXAM can unbiasedly estimate the (unconditional) average treatment effect and other effects. This informational virtue materializes regardless of whether the experimenter correctly predicts treatment effects and WTP.[5]

I also characterize the statistical efficiency in EXAM's average treatment effect estimation. EXAM's standard error is potentially smaller than that of RCTs, but in general, the standard error comparison of EXAM and a typical RCT is ambiguous. This motivates an empirical comparison of the two designs, which also allows me to confirm and quantify the other welfare, incentive, and unbiasedness properties.

I apply EXAM to data from a water cleaning experiment in Kenya (Kremer et al., 2011). Compared to RCTs, EXAM turns out to substantially improve participating households' predicted welfare. Here, welfare is measured by predicted effects of clean water on child diarrhea and revealed WTP for water cleaning. EXAM is also found to almost always incentivize subjects to report their true WTP. Finally, EXAM's data produces treatment effect estimates and standard errors similar to those from RCTs. EXAM therefore produces information that is as valuable for the outside society as that from RCTs.[6]

Taken together, EXAM sheds light on a way economic thinking can "*facilitate the advancement and use of complex adaptive (...) and other novel clinical trial designs,*" a performance goal by the US Food and Drug Administration (FDA) for 2018-2022.[7] Experimental design is a potentially life-saving application of economic market design (Roth, 2015). More concretely, my analysis shows how best to use predicted treatment effects for experimental design. The use of predicted effects for new experiments is established in medicine (Food and Drug Administration, 2010) and business (White, 2012), and emerging in the social sciences (Hahn et al., 2011) as important interventions such as deworming and conditional cash transfers ask for repeated evaluations. EXAM combines the predicted-effects consideration with another idea of respecting subjects' WTP for treatments.

After a review of related experimental designs, Section 2 outlines my motivation by

[5] This experimental value of EXAM and competitive equilibrium from equal incomes echoes Abdulkadiroğlu et al. (2017) and Narita (2016), who highlight the informational values of a different sort of mechanism design (centralized school choice with lotteries).

[6] Along the way, I develop C++ and Python computer programs to implement EXAM with little computational cost. A single execution of EXAM on data with 1540 subjects and 2 treatments takes less than a minute on average with a standard personal computer.

[7] See `https://www.fda.gov/downloads/forindustry/userfees/prescriptiondruguserfee/ucm511438.pdf`, retrieved in May 2019

providing facts about the impact of RCTs on participant welfare. Section 3 develops the EXAM experimental design, and Section 4 shows its welfare and incentive properties. Section 5 studies the experimental information embedded in EXAM and explains how to use data from EXAM for causal inference. An empirical application is in Section 6. Finally, Section 7 summarizes my findings, discusses their limitations, and outlines future directions. Proofs are in Appendix A.2.

1.1 Comparison with Existing Designs

Classical Experimental Design

The traditional experimental design literature (Cox and Cochran (1992), Athey and Imbens (2017) Section 7) is as old as the very concept of RCTs. This literature focuses on how to design experiments for maximizing information measured by the power of testing the null hypothesis of no treatment effect and other measures. This focus on information continues in much of the modern literature on sequential and adaptive experimental designs (Hahn et al., 2011). My interest lies more in subject welfare.

Preference- and Response-adaptive Designs

With its interest in subject well-being measured by WTP and predicted effects, EXAM is closer to younger and smaller strands of the literature on preference- and response-adaptive experimental designs. Preference-adaptive designs reflect subject preferences in treatment assignment probabilities. For example, Randomized Consent or Preference Trials (originally proposed by Zelen (1979) and further advocated by Angrist and Imbens (1991)) randomize subjects into two groups. In one group, subjects are allowed to choose the treatment or the control based on their preferences. All subjects in the other group are assigned to the control.

Selective Trials by Chassang et al. (2012) are more general preference-adaptive designs where the treatment assignment probability is increasing in the WTP for the treatment. Chassang et al. show their Selective Trials can be implemented as dominant-strategy mechanisms and are Blackwell more informative than RCTs in the limit. See also Björklund (1988) for a related experimental design proposal. Other examples of preference-adaptive designs are development economics RCTs that elicit and use subject preferences for treatment (Ashraf et al., 2006; Cohen and Dupas, 2010; Ashraf et al., 2010; Devoto et al., 2012; Dupas, 2014; Berry et al., 2018).

In complementary response-adaptive designs (reviewed by Hu and Rosenberger (2006) and Food and Drug Administration (2010)), the experimenter incorporates predicted treat-

ment effects into treatment assignment probabilities. For example, Play-the-Winner Rules (Zelen, 1969; Wei and Durham, 1978) more likely assign a treatment to patients predicted to have better treatment effects.[8]

Building upon these prior ideas, EXAM integrates preference- and response-adaptive designs into a single design. EXAM is formally shown to strike an optimal balance between WTP and predicted effects without compromising incentive compatibility and experimental information. EXAM thereby extends existing preference- and response-adaptive designs: If the experimenter shuts down WTP consideration by assuming constant WTP, EXAM simplifies to a Play-the-Winner Rule. Similarly, EXAM reduces to a Consent or Selective Trial if the experimenter ignores predicted effects.

Multi-Armed Bandit Algorithms

EXAM shares much of its spirit with Multi-Armed Bandit (MAB) algorithms in computer science, machine learning, and statistics (Bubeck and Cesa-Bianchi, 2012; Russo et al., 2018): Both MAB and EXAM strike a balance between exploration (gaining information) and exploitation (improving subject or experimenter welfare). MAB algorithms are popular in the web industry, especially for online ads, news, and recommendations (White, 2012). There are many differences between MAB and EXAM. For example, MAB mostly ignores incentive issues. In contrast, EXAM is formally and empirically shown to be nearly incentive compatible. EXAM may also be easier to implement than MAB since the implementation of MAB often requires a system for frequently observing outcome data and updating treatment assignment.

Clinical Trial Practices and Regulations

Finally, clinical trial practitioners and regulators have long recognized ethical concerns with RCTs, as highlighted in Marcia Angell's quote in the introduction. Their concerns manifested in regulations and practices that safeguard patients from excessive experimentation. Primary examples include the following: informed consent, a "stopping rule" that requires a sequential clinical trial to terminate if it becomes clear that its treatment is sufficiently better or worse than the control (Friedman et al. (1998) chapters 2 and 16), and a "randomized phase-in" design that assigns everybody to the treatment with randomized timing (Duflo et al. (2007) section 3.3.2). EXAM complements these existing practices and specify treatment

[8] The treatment assignment literature in econometrics (Manski, 2008; Kitagawa and Tetenov, 2018) and medicine (Chakraborty and Moodie, 2013) attempts a related but distinct task of using experimental data to optimally assign treatment to maximize welfare alone. See also related biostatistics developments on optimal dynamic treatment regimes by Murphy (2003) and Robins et al. (2008) among others.

assignment probabilities conditional on deciding to conduct a trial at a point in time and having subjects agreeing to participate in the trial.

2 Why Subject Welfare?

My goal is to design an experiment with an emphasis on subject welfare. Why should I study subject well-being? This section provides normative and practical reasons. Here I focus on the internal welfare of experimental subjects, leaving the external welfare of the outside population to Sections 5 and 6.

Normative Considerations

First, RCTs involve a large number of subjects. To demonstrate it, I assembled data on clinical trials registered in the WHO International Clinical Trials Registry Platform (ICTRP).[9] ICTRP is the largest international clinical trial registry and subsumes domestic platforms like ClinicalTrials.gov for the US.[10] Table 1 Panel a shows that the sum of sample sizes of registered trials is over 360 million for 2007-2017. As for social and economic RCTs, I scraped the American Economic Association's registry to find the sum of sample sizes of registered RCTs amounts to above 22 million for the last decade (Table 1 Panel b).

For such a large subject population, RCTs frequently randomize high-stakes treatment. The high-stakes and occasionally life-threatening nature of many RCTs is highlighted by examples in Table 2. In the first clinical trial (row i in Panel a), for example, a cholesterol-lowering drug treatment was found to lower the 5-year death rate of heart disease patients by about 30% relative to the baseline death rate in the control group. Other clinical trials in Table 2 Panel a also report significant impacts on survival and other crucial outcomes. As exemplified in Table 2 Panel b, social and economic RCTs also randomize treatment such as cash transfers, health insurance, HIV testing, and police patrol, as well as other numerous interventions related to childhood development, education, labor, and public finance (Fryer, 2017; Rothstein and von Wachter, 2017). These treatments are often found to have profound treatment effects.

Practical Considerations

Practical considerations also motivate a care for subject welfare. The successful implementation of any RCT depends on subject choices, including whether subjects participate in the

[9] `http://www.who.int/ictrp/en/`, retrieved in May 2019.
[10] `https://clinicaltrials.gov`, retrieved in May 2019.

RCT, whether subjects take up and use the assigned treatment, and whether subjects stay in contact in a follow-up period. The RCT produces useful information only if participants are active in each step. This prerequisite is hard to achieve, however. RCTs often suffer from subject indifference or fear in the form of non-participation, non-compliance, and dropouts before, during, and after experiments (Friedman et al. (1998) chapters 10 and 14, Duflo et al. (2007) sections 4.3 and 6.4, Glennerster (2017) sections 2.1 and 2.2).

A welfare-conscious experimental design could alleviate non-participation, non-compliance, and dropouts. Indeed, King et al. (2005) provide a clinical trial meta-analysis suggesting that incorporating subject preferences makes subject recruitment easier. In a range of econometric and theoretical models, welfare-enhancing treatment assignment is predicted to facilitate compliance with treatment assignment (Björklund and Moffitt, 1987; Heckman and Vytlacil, 2005; Chan and Hamilton, 2006). Chan and Hamilton (2006) use AIDS trial data to find that subjects experiencing better treatment effects are less likely to drop out.[11]

Finally, ethical experimental designs would ease collaboration with partner governments and companies that may have an ethical and reputational concern with involvement in RCTs (Glennerster (2017) section 1).

3 Experiment-as-Market (EXAM)

3.1 Framework

The normative and practical importance of subject well-being prompts me to design an experiment that balances subject welfare with experimental information. An *experimental design problem* consists of:

- Experimental *subjects* $i_1, ..., i_n$.
- Experimental *treatments* $t_0, t_1, ..., t_m$ where t_0 is a placebo or control.
- Each subject i's *preference or WTP* $w_{it} \in \mathbb{R}$ for treatment t where $w_{it} \geq w_{it'}$ means subject i weakly prefers treatment t over t'. Let $w_i \equiv (w_{it})_t$.
- Each treatment t's *predicted treatment effect* $e_{ti} \in \mathbb{R}$ for subject i where $e_{ti} \geq e_{t'i}$ means treatment t is predicted to have a weakly better effect than t' for subject i.

[11] In an effort to maximize the treatment take-up rate and minimize attrition, many field experiments start with an expression-of-interest survey before randomization and recruit only survey respondents who express strong interest. This recruitment practice causes external validity concerns. These concerns may also be alleviated by replacing the experimenter's discretionary selective recruitment with an experimental design respecting subject welfare in a rule-based way. See also Hull (2018) and references therein for other survey designs and analysis methods to deal with attrition.

When multiple outcomes matter, e_{ti} can be set to the predicted effect on a known function of these outcomes. Let $e_i \equiv (e_{ti})_t$.[12]

I assume e_{ti} and w_{it} to be deterministic for simplicity. I normalize e_{ti} and w_{it} by assuming $e_{t_0 i} = w_{it_0} = 0$ for every subject i. e_{ti} and w_{it} are therefore the predicted effect of t and WTP for t, respectively, relative to the control t_0. This normalization is without loss of generality because only differences in WTP and predicted effects matter for subject welfare from treatments $t_0, ..., t_m$. Every experimental design discussed below produces the same assignment probabilities with and without the normalization.

I use e_{ti} and w_{it} as complementary welfare measures, one outcome- or treatment-effect-based and one WTP-based. Each has an established role in economic welfare analysis, especially because WTP is sometimes found to be weakly or even negatively correlated with treatment effects (Walters, 2018). The medical literature more frequently studies treatment effects but also acknowledges that patients often have heterogeneous preferences for treatments (even conditional on treatment effects). This is especially the case for psychologically sensitive treatments like abortion methods (Henshaw et al., 1993) and depression treatments (Chilvers et al., 2001). In response to these findings, a US-government-endorsed movement tries to bridge the gap between evidence-based medicine and patient-preference-centered medicine (Food and Drug Administration, 2016). According to advocates, "*patient-centered care (...) promotes respect and patient autonomy; it is considered an end in itself, not merely a means to achieve other health outcomes*" (Epstein and Peters, 2009). My welfare criterion echoes this trend and accommodates both outcome- and preference-based approaches.

Predicted effects and WTP may also be freely heterogeneous and correlated. This is an important generality since evidence of correlation between treatment effects and WTP is ample both in the social sciences and medicine (Preference Collaborative Review Group, 2008; Swift and Callahan, 2009). To be consistent with the evidence, the above setup allows arbitrary correlation between predicted effects and WTP.

3.1.1 Where Do WTP and Predicted Effects Come From?

It is best to estimate predicted effects e_{ti} from prior experimental or observational data. In particular, the experimenter could use prior data to estimate heterogeneous treatment effects conditional (stratified) on observable subject characteristics, and apply the estimates to each subject i's characteristics, producing predicted effects e_{ti}. The most reliable data source is a prior RCT of the same treatment. Subjects in the prior RCT can be different from those

[12] Here I assume WTP and predicted effects are fixed and with cardinal meaning. See Appendices A.1.3 and A.1.4 for discussions about what to do when WTP and predicted effects are uncertain or ordinal.

in the new experiment to be designed. Such sequential RCTs with the same treatment are common in medicine and business, and are growing in the social sciences (e.g., many RCTs for deworming). I illustrate the use of prior RCT data in my empirical application.

For WTP w_{it}, there are a couple of possible sources. The experimenter may ask each subject i to self-report WTP w_i in an incentive compatible way, as proposed by Zelen (1979) and Chassang et al. (2012).[13] Alternatively, the experimenter may estimate WTP with prior data on subjects' treatment choices and their observable characteristics. Such data allows the experimenter to estimate heterogeneous revealed WTP conditional on subject characteristics. The WTP estimates then provide the experimenter with a prediction for each subject i's WTP given i's characteristics. I conduct such demand estimation with a discrete choice model in my empirical application in Section 6.[14]

3.2 Experimental Designs

Taking any experimental design problem as given, an *experimental design* specifies treatment assignment probabilities (p_{it}) where p_{it} is the probability that subject i is assigned to treatment t under the experimental design. The benchmark design is the standard Randomized Controlled Trial, formalized as follows.

Definition 1 (*Randomized Controlled Trial* a.k.a. *RCT*)**.** *Randomized Controlled Trial* is an experimental design that assigns each subject i to each treatment t with the impersonal treatment assignment probability p_t^{RCT} that is assumed to be written as $p_t^{RCT} = c_t/n$ for some natural number $c_t < n$.

The vast majority of clinical trials use RCT or similarly impersonalized randomization, an empirical fact shown in Appendix A.3.2 and Appendix Table A.1.[15] I call c_t *pseudo capacity* or supply, and require experimental designs to satisfy the pseudo capacity constraint that $\sum_i p_{it} \leq c_t$ for every treatment $t = t_1, ..., t_m$. This pseudo capacity constraint is important when treatment is expensive or hard to make and deliver.

I investigate welfare-enhancement with a design that I call Experiment-as-Market or EXAM in short.

[13] This self-reporting method raises the question of incentive compatibility. I study incentive compatibility theoretically in Section 4.2 and empirically in Section 6.3.

[14] Similar demand elicitation or estimation but for different purposes can be found in Ashraf et al. (2006); Cohen and Dupas (2010); Ashraf et al. (2010); Kremer et al. (2011); Devoto et al. (2012); Dupas (2014); Berry et al. (2018).

[15] A significant fraction of the real-world experiments are stratified. If the benchmark is an experiment stratified on some variables, I would implement EXAM conditional on the stratifying variables, i.e., for each subpopulation of subjects sharing the same stratifying variables. The subsequent comparison between RCT and EXAM holds conditional on the stratifying variables.

Definition 2 (*Experiment-as-Market* a.k.a. *EXAM*)**.** In the experimenter's computer, distribute any common artificial budget $b > 0$ to every subject.[16] Find any price-discriminated competitive market equilibrium, i.e., any treatment assignment probabilities (p^*_{it}) and their prices π_{te} with the following properties:[17]

- Effectiveness-discriminated treatment pricing: There exist $\alpha < 0$ and $\beta_t \in \mathbb{R}$ for each treatment t such that the price of a unit of assignment probability to t for subjects with $e_{ti} = e \in \mathbb{R}$ is

$$\pi_{te} = \alpha e + \beta_t.$$

- Subject utility maximization: For each subject i,

$$(p^*_{it})_t \in \arg\max_{p_i \in P} \textstyle\sum_t p_{it} w_{it} \text{ s.t. } \sum_t p_{it} \pi_{te_{ti}} \leq b,$$

  where $p_i \equiv (p_{it})_t$ and $P \equiv \{p_i \in \mathbb{R}^{m+1} | \sum_{t=t_0}^{t_m} p_{it} = 1 \text{ and } |p_{it}| \leq p\}$ where p is a large enough number. $\pi_{te_{ti}}$ is the price of a unit of the assignment probability to treatment t for subject i. EXAM breaks ties or indifferences so that every subject i's p^*_i solves the above problem with the minimum expenditure $\sum_t p_{it} \pi_{te_{ti}}$ while $(p^*_{it})_t = (p^*_{jt})_t$ for any subjects i and j with $w_i = w_j$ and $e_i = e_j$, where recall $w_i \equiv (w_{it})_t$ and $e_i \equiv (e_{ti})_t$.

- Meeting capacity constraints: $\sum_i p^*_{it} \leq c_t$ for every treatment $t = t_1, ..., t_m$ and $\sum_i p^*_{it} < c_t$ only if $\pi_{te_{ti}} \leq 0$ for every i.[18]

Let ϵ be a non-negative number such that the experimenter would like the assignment probabilities to be always within $[\epsilon, 1 - \epsilon]$. Take any $\epsilon \in [0, \bar{\epsilon}]$ as given, where $\bar{\epsilon} \equiv \min_t p^{RCT}_t$ is the largest possible value of ϵ.[19] I define EXAM's treatment assignment probabilities as

$$p^*_{it}(\epsilon) \equiv (1 - q) p^*_{it} + q p^{RCT}_t,$$

where $q \equiv \inf\{q' \in [0, 1] | (1 - q') p^*_{it} + q' p^{RCT}_t \in [\epsilon, 1 - \epsilon] \text{ for all } i \text{ and } t\}$.

[16] Any b, without loss of generality, results in the same assignment probabilities. It is also possible to let the budget vary across subjects while obtaining the same theoretical results.

[17] There may be multiple equilibria. I fix any equilibrium selection method.

[18] The latter part is necessary to make sure that EXAM wastes treatment t only when there is no enough demand for t even with a nonpositive price.

[19] Why is $\bar{\epsilon}$ the largest possible value of ϵ? Suppose $\epsilon > \bar{\epsilon} \equiv \min_t p^{RCT}_t$. For any $t \in \arg\min_t p^{RCT}_t$, whenever $p^*_{it} \leq p^{RCT}_t$, I have

$$(1 - q') p^*_{it} + q' p^{RCT}_t \notin [\epsilon, 1 - \epsilon]$$

for any $q' \in [0, 1]$. On the other hand, if $\epsilon \leq \bar{\epsilon}$, then $q' = 1$ guarantees that $(1 - q') p^*_{it} + q' p^{RCT}_t = p^{RCT}_t \in [\epsilon, 1 - \epsilon]$ for all i and t. Thus ϵ must be between 0 and $\bar{\epsilon}$.

I name this experimental design Experiment-as-Market (EXAM) because EXAM randomly assigns treatments to subjects via a synthetic centralized market. p^*_{it} can be seen as a generalization or variation of the classic idea of competitive market equilibrium from equal incomes (Friedman, 1962; Hylland and Zeckhauser, 1979; Budish et al., 2013; He et al., 2017; Mollner and Weyl, 2018).

More specifically, in Definition 2, EXAM endows each subject with a common imaginary budget. This budget has nothing to do with economic conditions subjects face in the real world. EXAM then lets each subject use the budget to purchase one of the most preferred bundles of treatment assignment probabilities, taking their prices as given. The prices are personalized so that each treatment is cheaper for subjects predicted to benefit more from the treatment. EXAM computes its treatment assignment probabilities as the resulting personalized-price competitive market equilibrium.[20] EXAM finally requires each subject to get each treatment with a probability strictly between 0 and 1. This requirement is important for EXAM to produce non-degenerate random assignments and unbiasedly estimate causal treatment effects; some foundations for this desire for non-degenerate randomization can be found in Proposition 4 below, Blackwell and Girshick (1954) section 8.7, Imbens and Rubin (2015) chapter 3, and Banerjee et al. (2017).[21]

To sum up, the steps for implementing EXAM are as follows.

(1) Obtain predicted effects e_{ti} if possible and relevant, as described in Section 3.1.1.

(2) Obtain WTP w_{it} if possible and relevant, as described in Section 3.1.1.

(3) Apply Definition 2 of EXAM to the data from steps 1 and 2, producing assignment probabilities $p^*_{it}(\epsilon)$.

EXAM is an enrichment of RCT using simple randomization. To see this, note that EXAM allows the experimenter to turn off welfare considerations. For instance, if the experimenter does not know or care about predicted effects, she would let $e_{ti} = e_{tj}$ for all subjects i and j and treatment t. Similarly, let $w_{it} = w_{jt} > 0$ if WTP is unknown or irrelevant; I make the common WTP positive for a minor technical reason (I need to make every subject

[20] The first step of Definition 2 raises two questions, whether such an equilibrium exists and how to find such an equilibrium. After positively solving the first existence question in Proposition 2 below, I develop and implement a script to find an equilibrium in the empirical application in Section 6. See Budish et al. (2016) for a related algorithmic development on a different problem (MBA course allocation).

[21] See Kasy (2016) for an argument against randomization. Definitions 1 and 2 leave unspecified how to draw a final treatment assignment from p^{RCT}_t and $p^*_{it}(\epsilon)$, respectively. For the moment, my analysis applies to any method to draw a treatment assignment. I impose more structures in Section 5 and implement an algorithm to draw an assignment in the empirical application in Section 6. For EXAM, it is known to be always possible to draw a treatment assignment in a way consistent with $p^*_{it}(\epsilon)$ (Budish et al. (2013)'s Theorem 1, the generalized Birkhoff-von Neumann Theorem).

prefer each treatment over the control). For example, the experimenter may want to exclude WTP when there is a concern that revealed or self-reported WTP may be distorted by ignorance, information frictions, or liquidity constraints. The following fact shows that EXAM is equivalent to RCT when the experimenter ignores both WTP and predicted effects.

Proposition 1 (EXAM nests RCT)**.** *Suppose that WTP and predicted effects are unknown or irrelevant so that $w_{it} = w_{jt} > 0$ and $e_{ti} = e_{tj}$ for all subjects i and j and treatment t. Then EXAM reduces to RCT using simple randomization, i.e., for every $\epsilon \in [0, \bar{\epsilon}]$, subject i, and treatment t, I have*

$$p^*_{it}(\epsilon) = p^{RCT}_t.$$

EXAM also extends other more sophisticated designs, such as the Play-the-Winner Rule (Wei and Durham, 1978), Consent Trials (Zelen, 1979; Angrist and Imbens, 1991), and Selective Trials (Chassang et al., 2012). These designs emerge if EXAM ignores either WTP or predicted effects, but not both, as explained in Section 1.1. The experimenter may want to ignore WTP or predicted effects when they are unknown or unimportant from the experimenter's perspective.

4 Welfare and Incentive

4.1 Welfare

As opposed to the special case in Proposition 1, the experimenter is often concerned about WTP and predicted effects (as in studies reviewed in Section 2). In such cases, EXAM differs from RCT and is welfare-optimal in the following sense.

Proposition 2 (Existence and Welfare)**.** *There exists p^*_{it} that satisfies the conditions in Definition 2. For any such p^*_{it} and any $\epsilon \in [0, \bar{\epsilon}]$, the resulting EXAM assignment probability $p^*_{it}(\epsilon)$ satisfies the following property: There is no other experimental design $(p_{it}) \in P^n$ with $p_{it} \in [\epsilon, 1-\epsilon]$ for all subject i and treatment t, $\sum_i p_{it} \leq c_t$ for all $t = t_1, ..., t_m$, and the following better welfare property:*

$$\sum_t p_{it} w_{it} \geq \sum_t p^*_{it}(\epsilon) w_{it} \text{ and } \sum_t p_{it} e_{ti} \geq \sum_t p^*_{it}(\epsilon) e_{ti}$$

for all i with at least one strict inequality.

Proposition 2 says that no other experimental design ex ante Pareto dominates EXAM in terms of the expected WTP for and predicted effect of assigned treatment (while satisfying

the random assignment and capacity constraints).[22] This ex ante Pareto optimality is known to imply ex post Pareto optimality and "ordinal" ex ante optimality (Bogomolnaia and Moulin, 2001).[23] In contrast, RCT fails to satisfy the welfare property as it ignores WTP and predicted effects. I empirically quantify the welfare gap between RCTs and EXAM in Section 6.3.

4.2 Incentive

Proposition 2 takes WTP w_{it} as given and assumes that it represents true WTP. In practice, the experimenter often needs to elicit the WTP information w_{it} from subjects, raising an incentive compatibility concern. This section shows EXAM allows the experimenter to extract WTP in an almost incentive compatible way. My analysis of incentive compatibility builds upon the literature on incentive compatibility of competitive equilibria and experimental designs (Jackson, 1992; Chassang et al., 2012; Azevedo and Budish, 2017; He et al., 2017).

Unfortunately, it is known that no experimental design satisfies the welfare property in Proposition 2 and exact incentive compatibility for general problems (Hylland and Zeckhauser, 1979). This compels me to investigate approximate incentive compatibility in large experimental design problems. Only for this section, consider a sequence of experimental design problems $(i_1, ..., i_n,\ t_0, t_1, ..., t_m,\ (c_t^n))_{n\in\mathbb{N}}$ indexed by the number of subjects, n. Let $\epsilon^n \in [0, \bar{\epsilon}^n)$ (where $\bar{\epsilon}^n$ is $\bar{\epsilon}$ for the n-th problem) be the value of the bound parameter ϵ the experimenter picks for the n-th problem in the sequence. The set of treatments $t_0, t_1, ..., t_m$ is fixed, but everything else may change as n increases. This modeling with a fixed number of treatments and an increasing number of subjects is consistent with real-world experiments with only a few treatments but with hundreds of subjects or more.

To investigate the incentive structure in EXAM, imagine that subjects report their WTP to EXAM. EXAM then uses the reported WTP to compute treatment assignment probabilities. For the n-th problem in the sequence, let $p_i^{*n}(w_i, e_i, w_{-i}, e_{-i}; \epsilon^n)$ be EXAM's treatment assignment probability vector for subject i when subjects report WTP (w_i, w_{-i}) and predicted effects are (e_i, e_{-i}) where $w_{-i} \equiv (w_j)_{j\neq i}$ and $e_{-i} \equiv (e_j)_{j\neq i}$. I extend this notation to

[22] Proposition 2 implies that EXAM is ex ante Pareto optimal for expected WTP alone if the experimenter shuts down predicted effects by assuming $e_{ti} = e_{tj}$ for all subjects i and j and treatment t. Similarly, EXAM satisfies Pareto optimality for expected predicted effects alone when EXAM ignores WTP.

[23] Ex post optimality means that no other (p_{it}) has the following property: $w_{it_i} \geq w_{it_i^*}$ and $e_{t_i i} \geq e_{t_i^* i}$ always hold for all i with at least one strict inequality, where t_i and t_i^* are treatments ex post assigned to i under the alternative design (p_{it}) and EXAM, respectively. Ordinal ex ante optimality is a stronger property that there is no other (p_{it}) such that for all affine transformations f and g, $\sum_t p_{it} f(w_{it}) \geq \sum_t p_{it}^*(\epsilon) f(w_{it})$ and $\sum_t p_{it} g(e_{ti}) \geq \sum_t p_{it}^*(\epsilon) g(e_{ti})$ for all i with at least one strict inequality.

the case where other subjects' WTP reports and predicted effects are random:

$$p_i^{*n}(w_i, e_i, F; \epsilon^n) \equiv \int_{(w_{-i}, e_{-i}) \in (W \times E)^{n-1}} p_i^{*n}(w_i, e_i, w_{-i}, e_{-i}; \epsilon^n) \times \Pr\{(w_{-i}, e_{-i}) \sim_{iid} F\} d(w_{-i}, e_{-i}).$$

Here $\Pr\{(w_{-i}, e_{-i}) \sim_{iid} F\}$ denotes the probability that vector $(w_{-i}, e_{-i}) \equiv (w_j, e_j)_{j \neq i}$ is realized from $n-1$ iid draws (w_j, e_j) from the distribution $F \in \Delta(W \times E)$. $\Delta(W \times E)$ is the set of full-support distributions over the WTP space W and the predicted effect space E. The iid assumption is based on the idea that there are many subjects, so they do not distinguish other subjects ex ante. Only for this section, I restrict WTP and predicted effects to belong to finite sets W and E, respectively, in any problem along the sequence. It is possible to eliminate this simplifying assumption by using an alternative proof technique like He et al. (2017)'s. This concept allows me to state an asymptotic incentive compatibility property.

Proposition 3 (Incentive)**.** *EXAM with WTP reporting is asymptotically incentive compatible, i.e., for any sequence of experimental design problems with any ϵ^n in $[0, \bar{\epsilon}^n)$, any $F \in \Delta(W \times E)$, any $\delta > 0$, there exists n_0 such that, for any $n \geq n_0$, any subject i, any predicted effect e_i, any true and manipulated WTP values w_i and w_i', I have*

$$\sum_t p_{it}^{*n}(w_i, e_i, F; \epsilon^n) \times w_{it} \geq \sum_t p_{it}^{*n}(w_i', e_i, F; \epsilon^n) \times w_{it} - \delta.$$

Proposition 3 says that EXAM approximately incentivizes every subject to report her true WTP, at least for large enough experimental design problems. The experimenter using EXAM can therefore ask subjects to report their true WTP without any deception. As additional support for incentive compatibility, Section 6.3 shows that EXAM is close to incentive compatible in my empirical application only with a finite number of subjects. This suggests asymptotic Proposition 3 is relevant even for real-scale problems.

For intuition, first consider a case with only one treatment t_1 that subject i prefers over the control t_0. Why is there no incentive for subject i to misreport a larger WTP for t_1? As long as subject i prefers t_1 over t_0, subject i spends her entire budget b into purchasing t_1 and gets an assignment probability of $\min\{b/\pi_{it_1}, 1\}$. Misreporting a larger WTP would not affect this assignment probability, confirming the incentive compatibility. More generally, exact incentive compatibility may break down in small problems; see Hylland and Zeckhauser (1979) for such an example. Nevertheless, EXAM is always asymptotically incentive compatible since there is no incentive to misreport when the prices are exogenously fixed, which is approximately true when the number of subjects is large.

5 Information

Despite the welfare merit, EXAM also lets the experimenter estimate treatment effects as unbiasedly and precisely as they would do in RCTs. To spell it out, I switch back to any given finite problem with fixed WTP and predicted effects. I discuss not only bias but also variance in treatment effect estimation. The finite-sample econometric comparison of EXAM and RCT requires me to specify how each design draws a treatment assignment from its assignment probabilities. For simplicity, only the main text assumes that $p_t n_p$ is an integer for every t and p where $n_p \equiv \sum_{i=1}^{n} 1\{p_i^*(\epsilon) = p\}$ is the number of subjects with assignment probability vector p and p_t is the t-th element of p. Appendix A.1.1 generalizes the definition and argument below to a general setting where $p_t n_p$ is any real number. Consider the following method of drawing a deterministic treatment assignment.

Definition 2 (EXAM Continued). Starting from the end of Definition 2 in Section 3.2, draw a treatment assignment from $p_{it}^*(\epsilon)$ as follows. For each assignment probability vector p,

- Uniformly randomly pick $p_{t_0} n_p$ subjects from $\{i|p_i^*(\epsilon) = p\}$ and assign them to t_0.

For each subsequent step $k = 1, ..., m$,

- Step k: From the remaining $n_p - \sum_{t=t_0}^{t_{k-1}} p_t n_p$ subjects in $\{i|p_i^*(\epsilon) = p\}$, uniformly randomly pick $p_{t_k} n_p$ subjects and assign them to t_k.

I assume that an RCT would draw a deterministic treatment assignment by a specialization of the above method assuming every subject i to have $p_{it}^*(\epsilon) = p_t^{RCT}$.

Suppose the experimenter is interested in the causal effect of each treatment on an outcome Y_i. Following the standard potential outcome framework for causal inference (Imbens and Rubin, 2015), let $Y_i(t)$ denote subject i's potential outcome that would be observed if subject i receives treatment t. Let D_{it} be the binary indicator that subject i is ex post assigned to treatment t. The observed outcome is written as $Y_i = \sum_t D_{it} Y_i(t)$. While $Y_i(t)$ is assumed to be fixed, D_{it} and Y_i are random variables, the distributions of which depend on the experimenter's choice of an experimental design. Let $Y \equiv (Y_i)$, $D_i \equiv (D_{it})_t$, and $D \equiv (D_i)$.

The experimenter would like to learn any parameter of interest θ of the distribution of potential outcomes $Y_i(t)$'s, many of which are unobservable. Formally, θ is any mapping $\theta : \mathbb{R}^{n \times (m+1)} \to \mathbb{R}$ that maps each possible value of $(Y_i(t))$ into the corresponding value of the parameter. For example, θ may be the average treatment effect (ATE_t) of treatment t over control t_0, $\dfrac{\sum_{i=1}^{n}(Y_i(t) - Y_i(t_0))}{n}$. The experimenter estimates θ with an estimator $\hat{\theta}(Y, D)$,

a function only of observed outcomes and treatment assignments. Given any experimental design (p_{it}), I say an estimator $\hat{\theta}(Y, D)$ is *simple* if $\hat{\theta}(Y, D)$ can be written as

$$\hat{\theta}(Y, D) = \sum_i f(Y_i, D_i, p_i) + \sum_t \sum_p \sum_{p'} g_{tpp'}((N_{pt}))\hat{\mu}_p(t)\hat{\mu}_{p'}(t)$$

for some function f, $\hat{\mu}_p(t) \equiv \dfrac{\sum_{i:p_i=p} D_{it}Y_i}{p_t \sum_{i=1}^n 1\{p_i = p\}}$, and weights $g_{tpp'}$, which may depend on $N_{pt} \equiv \sum_{i:p_i=p} D_{it}$ but *not* on individual D_{it}'s.[24] I say parameter θ is *unbiasedly estimable with experimental design* $p \equiv (p_{it})$ *and a simple estimator* if there exists a simple estimator $\hat{\theta}(Y, D)$ such that

$$E(\hat{\theta}(Y, D)|(p_{it})) = \theta,$$

where $E(\cdot|(p_{it}))$ is expectation with respect to the distribution of D_{it} induced by experimental design (p_{it}) given the fixed finite experimental design problem.[25]

EXAM turns out to be as informative as RCT in terms of the set of parameters unbiasedly estimable with each experimental design and a simple estimator. Throughout this section, assume $p_t n_p > 1$ for all t and p for which at least one subject i has $p_i^*(\epsilon) = p$. This assumption is likely to hold when the experimenter uses coarse values of predicted effects and WTP.

Proposition 4 (Unbiased Estimability)**.** *If parameter θ is unbiasedly estimable with RCT p_t^{RCT} and a simple estimator, then θ is also unbiasedly estimable with EXAM $p_{it}^*(\epsilon)$ with any $\epsilon > 0$ and a simple estimator.*[26]

Many key parameters, such as the average treatment effect, the treatment effect on the treated, and the mean and variance of potential outcomes are known to be unbiasedly estimable with RCT and a simple estimator (see Appendix A.2).[27] Proposition 4 implies that these parameters are also unbiasedly estimable with EXAM.

[24] More formally, $f : \mathbb{R} \times \mathcal{D} \times \mathcal{P} \to \mathbb{R}$ where $\mathcal{D} \equiv \{d \in \{0,1\}^{m+1} | \sum_t d_t = 1\}$ and $\mathcal{P} \equiv \{p_i | i = i_1, ..., i_n\}$. $g_{tpp'} : \mathbb{N}^{|\mathcal{P}|(m+1)} \to \mathbb{R}$ for each t, p, and p'. I allow f and $g_{tpp'}$ to use known elements of the experimental design problem such as capacities c_t and treatment assignment probabilities p_{it}. I do not allow $\hat{\theta}(Y, D)$ to use unknown elements, especially potential outcomes.

[25] I use this finite-sample framework throughout this section. The appendix provides an alternative large-sample setting.

[26] On the other hand, EXAM and RCT are not comparable in terms of Blackwell's order (Blackwell and Girshick, 1954) in my finite sample framework. This contrasts to the large sample analysis by Chassang et al. (2012), where they compare their Selective Trial and RCT in terms of Blackwell's order.

[27] I define the treatment effect on the treated for experimental design (p_{it}) as $E(\dfrac{\sum_{i=1}^n (Y_i(t) - Y_i(t_0))D_{it}}{\sum_{i=1}^n D_{it}}|(p_{it}))$ while the mean of potential outcomes as $\dfrac{1}{n}\sum_{i=1}^n Y_i(t)$. I define the variance of potential outcomes as $\dfrac{1}{n}\sum_{i=1}^n (Y_i(t) - \dfrac{1}{n}\sum_{j=1}^n Y_j(t))^2$ or $\dfrac{1}{n-1}\sum_{i=1}^n (Y_i(t) - \dfrac{1}{n}\sum_{j=1}^n Y_j(t))^2$, both of which are unbiasedly estimable with RCT and a simple estimator.

Corollary 1. *The average treatment effect, the treatment effect on the treated, and the mean and variance of potential outcomes are unbiasedly estimable with EXAM.*

5.1 Unbiased ATE Estimation with EXAM Data

I use the average treatment effect (ATE) to illustrate the intuition for and implementation of Proposition 4 and Corollary 1. Why is ATE unbiasedly estimable with EXAM? EXAM makes all subjects share the same budget constraint. As a result, if subjects share the same predicted effects and WTP, these subjects solve the same utility maximization problem and purchase the same vector of treatment assignment probabilities. EXAM therefore produces treatment assignment that is independent from (unconfounded by) potential outcomes conditional on predicted effects and WTP, which are observable to the experimenter:

$$(Y_i(t))_t \perp\!\!\!\perp D_i | (e_{ti}, w_{it})_t. \tag{1}$$

With this conditional independence, EXAM fits into causal inference with stratified experiments, selection-on-observables, and the propensity score, i.e., treatment assignment probabilities conditional on observables (see Imbens and Rubin (2015) for an overview). In particular, conditional independence (1) implies that the same conditional independence holds conditional on the propensity score (Imbens and Rubin (2015) section 12.3), which EXAM computes as $p_i^*(\epsilon) \equiv (p_{it}^*(\epsilon))_t$ and again known to the econometrician:

$$(Y_i(t))_t \perp\!\!\!\perp D_i | p_i^*(\epsilon). \tag{2}$$

This conditionally independent treatment assignment allows the experimenter to unbiasedly estimate the conditional average treatment effects of each t over t_0 conditional on observable propensity scores $p_i^*(\epsilon)$,

$$\frac{\sum_{i=1}^n 1\{p_i^*(\epsilon) = p\}(Y_i(t) - Y_i(t_0))}{\sum_{i=1}^n 1\{p_i^*(\epsilon) = p\}},$$

which I denote by $CATE_{pt}$ and is defined for each p such that at least one subject i has $p_i^*(\epsilon) = p$. These conditional-on-the-propensity-score effects are a version of Marginal Treatment Effects (Björklund and Moffitt, 1987; Heckman and Vytlacil, 2005). Marginal Treatment Effects are therefore estimable with EXAM's data.[28]

[28] To see this, as in Heckman and Vytlacil (2005), focus on an experimental design problem with only one treatment t_1 compared to the control t_0. Given EXAM's assignment probability $p_{it_1}^*(\epsilon)$, let $R_i \sim U[0,1]$ with $R_i \perp\!\!\!\perp (Y_i(t_0), Y_i(t_1))$. Write the treatment assignment as

$$D_{it_1} = 1\{R_i \leq p_{it_1}^*(\epsilon)\}.$$

By summing up such marginal or conditional effects, the experimenter can also back out the (unconditional) ATE, the single most important causal object identified and estimated by RCT. That is, with weights $\delta_p \equiv n_p/n$, I use $CATE_{pt}$'s to get ATE as follows:

$$\sum_p \delta_p CATE_{pt} = ATE_t.$$

Importantly, the key conditional independence properties (1) and (2) hold regardless of whether e_{ti} and w_{it} coincide with the true treatment effects and WTP. In this sense, like RCT, EXAM's informational virtue is robust to any of the experimenter's potential misspecifications about predicted effects and WTP.[29]

The above estimability argument motivates a strategy to estimate ATE with EXAM's data. As a warm-up, focus on $\{i|p_i^*(\epsilon) = p\}$, the subpopulation of subjects with propensity vector p, and consider this regression on the subpopulation:

$$Y_i = \alpha_p + \sum_{t=t_1}^{t_m} \beta_{pt} D_{it} + \epsilon_i.$$

By the conditional independence property (2), OLS estimate $\hat{\beta}_{pt}$ from this regression is unbiased for $CATE_{pt}$ for each treatment $t \neq t_0$. I then aggregate the resulting estimates $\hat{\beta}_{pt}$'s into $\sum_p \delta_p \hat{\beta}_{pt}$, which I denote by $\hat{\beta}_t^*$. This $\hat{\beta}_t^*$ is a multinomial propensity score weighting estimator that unbiasedly estimates the average treatment effect with its variance in an analytical form.

Proposition 5 (Bias and Variance)**.** *Suppose that the data-generating experimental design is EXAM $p^*(\epsilon) \equiv (p_{it}^*(\epsilon))_{it}$ with any given $\epsilon > 0$. $\hat{\beta}_t^*$ is an unbiased estimator of the average treatment effect. In particular,*

$$E(\hat{\beta}_t^*|p^*(\epsilon)) = ATE_t \text{ and } Var(\hat{\beta}_t^*|p^*(\epsilon)) = \sum_p \delta_p^2 \Big(\frac{S_{pt}^2}{p_t n_p} + \frac{S_{pt_0}^2}{p_{t_0} n_p} - \frac{S_{ptt_0}^2}{n_p} \Big),$$

where $\bar{Y}_p(t) \equiv \dfrac{\sum_{i:p_i^*(\epsilon)=p} Y_i(t)}{n_p}$ *is the mean of $Y_i(t)$ in the subpopulation with propensity p,*

Note that $E(D_{it_1}) = p_{it_1}^*(\epsilon)$ as desired. This model is a special case of Heckman-Vytlacil's model with local instrumental variable R_i because R_i is independent of $(Y_i(t_0), Y_i(t_1), p_{it_1}^*(\epsilon))$ by construction while R_i can be correlated with $(Y_i(t_0), Y_i(t_1))$. As a result, Heckman and Vytlacil (2005)'s method allows the experimenter to identify Marginal Treatment Effects with EXAM's data. Chassang et al. (2012) provide a similar discussion about their Selective Trial idea. See also Kowalski (2016); Mogstad and Torgovitsky (2018) for recent developments in the marginal treatment effect method.

[29] On the other hand, the welfare optimality in Proposition 2 is welfare-relevant only if the experimenter predicts treatment effects and WTP well.

$S^2_{pt} \equiv \dfrac{\sum_{i:p^*_i(\epsilon)=p}(Y_i(t)-\bar{Y}_p(t))^2}{n_p-1}$ *is the variance of* $Y_i(t)$ *in the subpopulation, and* $S^2_{ptt'} \equiv \dfrac{\sum_{i:p^*_i(\epsilon)=p}(Y_i(t)-Y_i(t')-(\bar{Y}_p(t)-\bar{Y}_p(t')))^2}{n_p-1}$ *is the variance of* $Y_i(t)-Y_i(t')$ *in the subpopulation.*

Though this variance expression is not identified in general, it provides guidance on when the variance is likely to be large or small. In practice, statistical inference may be based on randomization or large-sample inference.

Alternatively, empirical researchers may prefer a single regression controlling for propensity vectors:

$$Y_i = a + \sum_{t=t_1}^{t_m} b_t D_{it} + \sum_{t=t_1}^{t_m} c_t p^*_{it}(\epsilon) + e_i, \tag{3}$$

producing an alternative estimator $\hat{b}^*_t$. As verified in the appendix, $\hat{b}^*_t$ is an unbiased estimator of a differently weighted treatment effect:

$$E(\hat{b}^*_t | p^*(\epsilon)) = \frac{\sum_p \lambda_{pt} CATE_{pt}}{\sum_p \lambda_{pt}} \text{ with weights } \lambda_{pt} \equiv \delta_p p_t(1-p_t). \tag{4}$$

Estimators like $\hat{b}^*_t$ and $\hat{\beta}^*_t$ allow the experimenter to unbiasedly estimate key causal effects with EXAM.

5.2 Power Comparison of EXAM and RCT

Does EXAM compete with RCT in terms of statistical efficiency in ATE estimation? With RCT's data, the most standard estimator of ATE of treatment t over control t_0 is the difference in the average outcome between subjects assigned to treatment t and those assigned to control t_0:

$$\hat{\beta}^{RCT}_t \equiv \frac{\sum_i D_{it} Y_i}{\sum_i D_{it}} - \frac{\sum_i D_{it_0} Y_i}{\sum_i D_{it_0}}.$$

This $\hat{\beta}^{RCT}_t$ is a special case of $\hat{\beta}^*_t$ when $p^*_{it}(\epsilon) = p^{RCT}_t$. By Proposition 5, therefore, $\hat{\beta}^{RCT}_t$ is unbiased for ATE with the following variance, confirming a classic result about RCT.

Corollary 2 (Imbens and Rubin (2015)'s Theorem 6.2)**.**

$$E(\hat{\beta}^{RCT}_t | p^{RCT}) = ATE_t \text{ and } V(\hat{\beta}^{RCT}_t | p^{RCT}) = \frac{S^2_t}{c_t} + \frac{S^2_{t_0}}{c_{t_0}} - \frac{S^2_{tt_0}}{n},$$

where $S^2_t \equiv \dfrac{\sum_i (Y_i(t)-\bar{Y}(t))^2}{n-1}$ *and* $S^2_{tt'} \equiv \dfrac{\sum_i (Y_i(t)-Y_i(t')-(\bar{Y}(t)-\bar{Y}(t')))^2}{n-1}$.

Proposition 5 and Corollary 2 imply that EXAM may produce more precise ATE estimates ($V(\hat{\beta}_t^*|p^*(\epsilon)) < V(\hat{\beta}_t^{RCT}|p^{RCT})$). Such a situation occurs if potential outcomes are well correlated (positively or negatively) with EXAM's treatment assignment probabilities, as illustrated by the following example.

Example 1. Suppose there is only one treatment t_1, $n = 40$, and $c_{t_0} = c_{t_1} = 20$. Every subject has $Y_i(t_0) = 1$. The subjects are divided into four groups A, B, C, and D of the same size (10) based on their potential outcomes $Y_i(t_1)$. Let $Y_i(t_1) = 1, 2, 3$, and 4 for anybody in group A, B, C, and D, respectively. Assume the experimenter imperfectly predicts treatment effects: $e_{t_1 i} = 0$ for every i in group A or B while $e_{t_1 i} = 2$ for group C or D. Let $w_{it_1} > 0$ for all subjects. EXAM with $\epsilon < .2$ gives the following treatment assignment probabilities[30]: $p^*_{it_1}(\epsilon) = 0.2$ for every i in groups A and B while $p^*_{it_1}(\epsilon) = 0.8$ for groups C and D. Under RCT, $p_{t_1}^{RCT} = p_{t_0}^{RCT} = 20/40 = 0.5$ for all subjects. Applying Proposition 5 and Corollary 2 to this example, I have

$$V(\hat{\beta}_t^*|p^*(\epsilon)) = 0.013... < 0.032... = V(\hat{\beta}_t^{RCT}|p^{RCT}).$$

This example makes clear that information production in EXAM is not a diluted version of that in RCT. EXAM's ATE estimation is not only unbiased but also potentially more precise than RCT's; this is true even if the experimenter's prediction of treatment effects is imperfect. Appendix A.1.2 provides further support for this point by showing it remains true in an asymptotic framework.

In general, however, the precision comparison of EXAM and RCT is ambiguous. There are other examples with $V(\hat{\beta}_t^{RCT}|p^{RCT}) < V(\hat{\beta}_t^*|p^*(\epsilon))$; one such example with a binary treatment t_1 vs. t_0 is where $p_{t_0}^{RCT} = p_{t_1}^{RCT} = 0.5$ for every i, $p^*(\epsilon) \neq p^{RCT}$, and there is no correlation between potential outcomes and $p^*(\epsilon)$. This ambiguity is common in precision comparisons of experimental designs. This motivates me to empirically compare EXAM and RCT's estimation precision. The empirical application also allows me to verify and quantify the welfare, incentive, and unbiasedness properties of EXAM.

[30] EXAM outputs these treatment assignment probabilities if I set $\alpha = -\frac{15b}{8}$, $\beta_{t_1} = 5b$, and $\beta_{t_0} = 0$ given any budget b.

6 Empirical Application

6.1 Overview

My empirical test bed for EXAM is an application to a spring protection experiment in Kenya. Waterborne diseases, especially diarrhea, remain the second leading cause of death among children, comprising about 17% of child deaths under age five (about 1.5 million deaths each year).[31] The only quantitative United Nations Millennium Development Goal is in terms of "the proportion of the population without sustainable access to safe drinking water and basic sanitation," such as protected springs.[32] Yet there is controversy about the health impacts of spring protection. Experts argue that improving source water quality may only have limited effects, since, for example, water is likely recontaminated in transport and storage. These arguments were made in the absence of any randomized experiment.

This controversy motivated Kremer et al. (2011) to analyze randomized spring protection conducted by an NGO (International Children Support) in Kenya in the mid 2000s. This experiment randomly selected springs to receive protection from the universe of 200 unprotected springs. The experimenter selected and followed a representative sample of about 1500 households that regularly used some of the 200 springs before the experiment; these households are experimental subjects. Kremer et al. (2011) found that spring protection substantially improves source water quality and is moderately effective at improving household water quality after some recontamination. Diarrhea among children in treatment households fell by about a quarter of the baseline level. I call this real experiment "Kremer et al. (2011)'s experiment" and distinguish it from EXAM and RCT as formal concepts in my model.

Kremer et al. (2011)'s experiment provides an ideal setup for empirically evaluating EXAM. Their experiment is about a high-stakes treatment and produces rich data that allows me to measure not only treatment effects but also subjects' WTP for the treatment. I consolidate Kremer et al. (2011)'s experimental data and my methodological framework to empirically evaluate EXAM. Applying the language and notation of my model, experimental subjects are households in Kremer et al. (2011)'s sample.[33] The protection of the spring each household uses at baseline is a single treatment t_1 while no protection is the control t_0. Each

[31] See UNICEF and WHO's joint document "Diarrhoea: Why Children Are Still Dying and What Can be Done," at `http://apps.who.int/iris/bitstream/10665/44174/1/9789241598415_eng.pdf`, retrieved in May 2019.

[32] See `http://www.un.org/millenniumgoals/`, retrieved in May 2019. Spring protection encases the source of a natural spring in concrete, allowing water to flow from a pipe rather than seeping from the ground. In this way, the water source is protected from human or animal waste.

[33] Alternatively, it's possible to interpret springs as the subjects in my model. The resulting analysis produces similar results (available upon request).

household i's WTP for better water access t_1 is denoted by w_{it_1}, which I estimate below. I also estimate the heterogeneous treatment effect $e_{t_1 i}$ of spring protection t_1 on household i's child diarrhea outcome. Using this embedding, I implement EXAM and compare it with RCT to see which is a better design of a hypothetical future experiment on the spring protection treatment.

6.2 Treatment Effects and WTP

Treatment Effects

For executing EXAM, I need to measure w_{it_1} and $e_{t_1 i}$ and substitute them into EXAM. I estimate heterogeneous treatment effects $e_{t_1 i}$ of access to better water in a similar way as Kremer et al. (2011). This treatment effect estimation exploits additional details of Kremer et al. (2011)'s experiment. The experimenter NGO aspired to eventually protect all the 200 springs but planned for the protection intervention to be phased in over four years due to financial and administrative constraints. In each round, a subset of springs were randomly picked to be protected. Figure I in Kremer et al. (2011) details the timeline of the experiment. This experimental scheme legitimizes the following OLS regression at the (child i, spring j, survey round t)-level:

$$Y_{ijt} = (\phi_1 + \phi_2 X_i) T_{jt} + \alpha_i + \alpha_t + u_{ij} + \epsilon_{ijt}, \tag{5}$$

where Y_{ijt} is the binary outcome indicating that child i in a household drawing water from spring j at baseline has diarrhea in survey round t. X_i contains covariates of child i's household (baseline latrine or sanitation density, diarrhea prevention knowledge score, mother's years of education). Every covariate is normalized to be mean zero so that the coefficient ϕ_1 can be interpreted as the average treatment effect. T_{jt} is the binary treatment indicating that spring j is treated in survey round t. α_i, α_t, and u_{ij} are fixed effects. The treatment effect is $\phi_1 + \phi_2 X_i$ and is heterogeneous across subjects with different covariates X_i.

Estimates from the OLS regression (5) are in Table 3. The average treatment effect is about 4.5% absolute reduction or about 25% relative reduction in the diarrhea outcome Y_{ijt}. Households with higher scores in diarrhea prevention knowledge or mother's education level tend to have better treatment effects, although the relatively large standard errors argue for caution in interpretation. This heterogeneity may be present because such households are more likely to prefer and use protected springs, as suggested by a revealed preference analysis below.

I then use the OLS estimates to predict the treatment effect for each household i with

$\hat{e}_{t_1 i} \equiv \hat{\phi}_1 + \hat{\phi}_2 X_i$, where $\hat{\phi}_1$ and $\hat{\phi}_2$ are OLS estimates of ϕ_1 and ϕ_2, respectively. Kremer et al. (2011)'s experiment randomized T_{jt} and gives its coefficient estimate $\hat{e}_{t_1 i}$ an interpretation as a causal effect. Estimated treatment effects $\hat{e}_{t_1 i}$ exhibit significant heterogeneity, as illustrated in Figure 1 Panel a.

WTP

I estimate heterogeneous WTP w_{it_1} for the treatment as follows. In the experimental target area, each household draws water from a water source the household chooses among multiple sources in the neighborhood. This fact motivates a discrete choice model of households' water source choices, in which households trade off water quality against other source characteristics such as proximity. This model produces revealed preference estimates of household WTP for the spring protection treatment as a spring characteristic, which is identified by exogenous variation in the treatment generated by Kremer et al. (2011)'s experiment.

Specifically, I use a mixed or random-coefficient logit model (Train (2003), chapter 6):

$$U_{ijt} = (\beta_i + \gamma_1 X_i) T_{jt} - c_i D_{ij} + \delta_j + \epsilon_{ijt}, \tag{6}$$

where U_{ijt} is household i's utility from source j in survey round t, and D_{ij} is household i's roundtrip distance to spring j (measured in terms of minutes of walking time). β_i and c_i are random preference coefficients assumed to be distributed according to normal and triangular distributions, respectively, with unknown parameters to be estimated. I restrict the triangular distribution of c_i to have the same mean and standard deviation, making sure every household prefers proximity. δ_j's are spring-type fixed effects. δ_j's attempt to capture the average preference for potentially unobserved spring type characteristics other than treatment T_{jt} and distance D_{ij}. ϵ_{ijt} is logit utility shocks iid according to the type I extreme value distribution with usual variance normalization to $\pi^2/6$. I estimate the model with data on households' spring choices (in the final survey round after random spring protection) and a standard maximum simulated likelihood method, which I detail in Appendix A.3.3.

The mixed logit preference estimates are in Table 4. Households have significant distaste for distance and significant preferences for protected treatment springs (other characteristics being equal). Not surprisingly, households with better diarrhea prevention knowledge scores or higher education levels of mothers tend to have stronger revealed preferences for the spring protection treatment. This heterogeneity is expected if such households are more conscious of water quality.[34]

[34] Tables 3 and 4 show slight differences from Kremer et al.'s estimates. It is because I include the same

I then exploit the mixed logit estimates to estimate household i's WTP for treatment t_1 as $\hat{w}'_{it_1} \equiv \hat{\beta}_i + \hat{\gamma}_1 X_i$, where $\hat{\gamma}_1$ is the mixed logit estimate of γ_1. $\hat{\beta}_i$ is a value drawn from the estimated distribution of the random coefficient β_i. The identification of γ_1 and the distribution of β_i is helped by Kremer et al. (2011)'s experimental variation in protection treatment T_{jt} since otherwise T_{jt} is likely correlated with unobserved spring characteristics ϵ_{ijt}, making it impossible to identify the WTP for spring protection alone.

Since $\hat{w}'_{it_1}$ is in an elusive utility unit, I convert it into a more easily interpreted measure in terms of time cost of water collection. To do that, I first compute $\hat{w}'_{it_1}/\hat{c}_i$, where $\hat{c}_i$ is the mixed logit estimate of c_i (the distaste coefficient on distance). Again, I bootstrap the random coefficient $\hat{c}_i$ from its estimated distribution. I then multiply it by each household's self-reported time cost of traveling for a unit of distance. This procedure gives me a time cost measure of WTP for the treatment, $\hat{w}_{it_1}$. This $\hat{w}_{it_1}$ is measured by workdays utility-equivalent to $\hat{w}'_{it_1}$.

The estimated WTP $\hat{w}_{it_1}$ is displayed in Figure 1 Panel b, showing the histogram of simulated values of $\hat{w}_{it_1}$. The median WTP is about 25 workday-equivalent. While both WTP $\hat{w}_{it_1}$ and treatment effects $\hat{e}_{t_1 i}$ show sizable heterogeneity, there turns out to be only limited correlation between the two. This fact can be seen in the joint density plot in Figure 1 Panel c, where there is a positive correlation between WTP $\hat{w}_{it_1}$ and treatment effects $\hat{e}_{t_1 i}$, but the magnitude of the correlation is small (R^2 is lower than 0.12 when I regress one on the other). This demonstrates that WTP $\hat{w}_{it_1}$ and treatment effects $\hat{e}_{t_1 i}$ contain different types of information about subject welfare, suggesting the importance of respecting both WTP and predicted effects separately. This is what EXAM attempts to do, as I explain next.

6.3 EXAM vs RCT

Now imagine somebody is planning a new experiment for further investigating the same spring protection treatment. What experimental design should she use? Specifically, which is better between RCT and EXAM? A full-fledged comparison of experimental designs requires a meta-experiment that randomly assigns different designs to many experimental studies. To circumvent the difficulties of such a meta-experiment, I resort to an alternative approach exploiting the above WTP and treatment effect estimates.

My approach is to use the estimated WTP $\hat{w}_{it_1}$ and predicted effects $\hat{e}_{t_1 i}$ to simulate EXAM and compare EXAM with RCT in terms of welfare, information, and incentive properties. Throughout the comparison, I fix the set of subjects and treatments as in Kremer et al. (2011)'s experiment. That is, there are 1540 households as subjects to be assigned

set of a small number of covariate interactions both in the OLS and mixed logit models while Kremer et al. include different sets of covariate interactions and other controls in their models.

either to the single water source protection treatment t_1 or the control t_0. Set the treatment capacity c_{t_1} to be the number of households assigned to the treatment t_1 in Kremer et al.'s experiment (by the end of their survey period). I set the bound parameter ϵ to be 0.2; I investigate how the results change under another value of ϵ at the end. I fix predicted effects $e_{t_1 i}$ to their point estimate $\hat{e}_{t_1 i}$.

I simulate WTP with parametric bootstrap from the estimated distribution of $\hat{w}_{it_1}$, i.e., the estimated mixed logit model (6) (conditional on each household's fixed characteristics X_i). In this WTP simulation, I require all families with the same characteristics X_i to share the same WTP. After simulating $\hat{w}_{it_1}$, I compute treatment assignment probabilities $p^*_{it}(\epsilon)$ by running EXAM on the bootstrapped data along with other fixed parameters such as the treatment capacity.[35] The algorithm I use for executing EXAM is described in Appendix A.3.4.

The simulation process for RCT is analogous except that the treatment assignment probability is fixed at $p^{RCT}_{t_1} \equiv c_{t_1}/n = .43$. Note that this RCT is a hypothetical experimental design in line with my Definition 1 and different from Kremer et al. (2011)'s experiment involving additional real-world complications.

Welfare

I start with evaluating EXAM's welfare performance. Use EXAM's treatment assignment probabilities $p^*_{it_1}(\epsilon)$ to calculate two welfare measures for each household i:

$$w^*_i \equiv \sum_t p^*_{it}(\epsilon) w_{it} \text{ and } e^*_i \equiv \sum_t p^*_{it}(\epsilon) e_{ti}.$$

w^*_i and e^*_i are empirical analogues of the two welfare measures in my theoretical welfare analysis (Proposition 2).

I find that EXAM improves on RCT in terms of the welfare measures w^*_i and e^*_i, a result reported in Figure 2. The figure draws the distribution of w^*_i and e^*_i over households and 1000 bootstrap samples. Among other things, the mean of average WTP w^*_i for assigned treatments is about 89% or 9.4 workday-equivalent utilities higher under EXAM than it is under RCT. Another interpretation of this WTP improvement is about 37% of the average WTP for the treatment. Similarly, EXAM improves the mean of e^*_i by about 0.8% absolute reduction or 42% reduction relative to RCT's level. This predicted effect benefit amounts to

[35] To make treatment assignment probabilities take a modest number of values, I coarsen the values of WTP and predicted effects. Specifically, for each simulation and each of WTP and predicted effects, I first group its values into four quartiles and then replace each household's value by the median value within the quantile group to which the household belongs.

about 17% of the average treatment effect of the spring protection found by Kremer et al. (2011) and Table 3.

Information

Data from EXAM also allows me to obtain more or less the same conclusion about treatment effects as RCT. To see this, I augment the above counterfactual simulation with average treatment effect estimation as follows: I first simulate w_{it_1} and run EXAM to get treatment assignment probabilities $p^*_{it}(\epsilon)$. I use $p^*_{it}(\epsilon)$ to draw a final deterministic treatment assignment, denoted by a binary indicator D_i indicating i is ex post assigned to t_1. I then simulate counterfactual or predicted outcome Y_i under D_i by simulating the OLS model I estimate in the last section:

$$Y_i \equiv (\hat{\phi}_1 + \hat{\phi}_2 X_i) D_i + \hat{\alpha}_i + (\text{average of } \hat{\alpha}_t \text{ across all } t) + (\text{average of } \hat{u}_{ij} \text{ across all } j),$$

where objects with a hat mean estimates of the corresponding parameters in regression (5). I take the average of $\hat{\alpha}_t$'s and $\hat{u}_{ij}$'s to adapt regression (5) at the (i, j, t)-level to my counterfactual simulation setting at the household-i-level. Note that the above expression is the definition of Y_i, not a regression. Finally, I use the above simulated Y_i and D_i to estimate treatment effects with $\hat{b}^*$ from this OLS regression:

$$Y_i = a + bD_i + cp^*_{it_1}(\epsilon) + e_i,$$

where I control for propensity score $p^*_{it_1}(\epsilon)$ to make treatment assignment D_i random. This regression is a stripped-down version of the regression strategy (3) in Section 5. I also implement the other propensity-score-weighting estimator $\hat{\beta}^*$, again following the description in Section 5. The procedure for RCT is analogous except that the treatment assignment probability is fixed at p_t^{RCT}.

Program evaluation with EXAM turns out to be as unbiased and precise as that with RCT. Figure 3 plots the distribution of the resulting treatment effect estimates $\hat{b}^*$ and $\hat{\beta}^*$ over 1000 simulations. In line with Propositions 4 and 5, the means of $\hat{b}^*$ and $\hat{\beta}^*$ for EXAM are indistinguishable from those under RCT. Both experimental designs successfully recover Kremer et al. (2011)'s average treatment effect estimate (4.5% reduction in diarrhea; recall column 1 in Table 3).

Perhaps more importantly, the distributions of $\hat{b}^*$ and $\hat{\beta}^*$ for EXAM have similar standard deviations as those for RCT. This means that the two experimental designs produce similar exact, finite-sample standard errors in their estimates $\hat{b}^*$ and $\hat{\beta}^*$. Variations of this

observation are in Figure 4, which shows the distribution of p values for the estimates $\hat{b}^*$. The four panels use p values based on exact, non-robust, robust, and Abadie et al. (2017)'s finite population causal standard errors, respectively, where the exact standard error means the standard deviation in the distribution of $\hat{b}^*$ in Figure 3. RCT produces slightly smaller p values than EXAM, but the median p value is about 0.03 for RCT and about 0.04 for EXAM. Both EXAM and RCT therefore detect a significant average treatment effect for a majority of cases. Overall, EXAM appears to succeed in its informational mission of eliminating selection bias and recovering ATE precisely enough.

Incentive

EXAM's WTP benefits can be regarded as welfare-relevant only if EXAM provides subjects with incentives to reveal their true WTP. I conclude my empirical analysis with an investigation of the incentive compatibility of EXAM. I repeat the following procedure many times: As before, I simulate w_{it_1} and run EXAM to get treatment assignment probabilities $p^*_{it}(\epsilon)$. I then randomly pick one subject j as a WTP manipulator and one potential WTP manipulation w'_{jt_1} by j. I choose the manipulator j uniformly randomly among all subjects. The manipulation w'_{jt_1} is either from $N(w_{jt_1}, 100), N(w_{jt_1}, 1000), U(w_{jt_1}, w_{jt_1} + 100)$, or $U(w_{jt_1} - 100, w_{jt_1})$ where w_{jt_1} is j's true WTP. These computational scenarios cover different types of misreporting, that is, both over-reporting and under-reporting with different magnitudes. I run EXAM on the simulated data but with the WTP manipulation w'_{jt_1} to get treatment assignment probabilities $p'_{it}(\epsilon)$. I finally compute the true WTP gain from the manipulation w'_{jt_1}:

$$\Delta w \equiv \sum_t p'_{it}(\epsilon) w_{jt} - \sum_t p^*_{it}(\epsilon) w_{jt}.$$

EXAM is found to give subjects little incentive for WTP misreporting, empirically verifying Proposition 3. Figure 5 shows this by drawing the distribution of Δw over 1000 simulations and households. Across all scenarios, the WTP gain Δw from misreporting is mostly negative and well below zero on average.

Ideally, I would like to compute the gains from the optimal (as opposed to random) WTP manipulations. The optimal WTP manipulations are hard to find, however, since equilibrium prices endogenously respond to WTP manipulations in a complex, unknown manner. As a feasible exercise, Table 5 shows that even the most profitable manipulations in Figure 5 lead to normalized gains $\Delta w / w_{it_1}$ smaller than 0.021. This result suggests that there are unlikely to be manipulations that produce large gains.

Overall, in this empirical setting, EXAM provides subjects with stronger average incentives for truthful WTP reporting than RCT does (because subjects in RCT are indifferent

among all possible WTP reports). EXAM may therefore be better at eliciting reliable WTP data.

Role of Design Parameters

Finally, I analyze how the results depend on the choice of design parameters, especially ϵ, which governs how close EXAM must be to RCT. With a smaller value of $\epsilon = 0.1$, the same set of results as in Figures 2-5 and Table 5 are reported in Appendix Figures A.1-A.4 and Table A.3. The results stay qualitatively the same between the two analyses. This confirms the above baseline empirical analysis is robust.

Yet there is a key quantitative difference: Appendix Figure A.1 with $\epsilon = 0.1$ finds better welfare performance of EXAM compared to Figure 2 with $\epsilon = 0.2$. On the other hand, Appendix Figure A.3 and Figure 4 suggest EXAM's statistical efficiency deteriorates as ϵ drops from 0.2 to 0.1. This tradeoff is intuitive as smaller values of ϵ allow EXAM's assignment probabilities to get away from RCT's and focus more on welfare enhancement. This welfare enhancement may come at the cost of diluted information. The parameter ϵ thus embodies the welfare vs information tradeoff among different versions of EXAM. This observation raises an intriguing yet challenging methodological question of how to optimally specify ϵ. I leave this direction for future research.

7 Takeaway and Future Directions

Motivated by the high-stakes nature of many RCTs, I propose a data-driven, stratified experiment dubbed Experiment-as-Market (EXAM). EXAM is a solution to a hybrid experimental-design-as-market-design problem of maximizing participants' welfare subject to the constraint that the experimenter must produce as much information and incentives as in RCTs (Propositions 2-5). These properties are then verified and quantified in an empirical application where I simulate my design on a water source protection experiment. Taken together, the body of evidence suggests that EXAM improves subject well-being with little information and incentive costs. The demonstrated benefits are conservative in that they do not incorporate potential additional benefits from EXAM for improving recruitment, compliance with assigned treatment, and attrition (recall the discussion in Section 2).

This paper takes a step toward introducing welfare and ethics into experimental design. This opens the door to several open questions. In ongoing work, I am implementing EXAM in the field. This implementation raises practical questions, such as how to design an easy-to-use interface through which EXAM interacts with subjects, as well as a fast and scalable algorithm to implement EXAM, and how best to obtain predicted effects and WTP.

The empirical and computational analysis in Section 6 is an effort to tackle these practical challenges.

Econometrically and theoretically, this paper's analysis is simplistic in many respects, asking for a variety of extensions. Key extensions include introducing a decision-theoretic framework with an explicit social welfare function for the experimenter; analyzing EXAM in an instrumental variable setting where subjects may not comply with treatment assignment; analyzing experimental designs with endogenous subject participation and dropout; introducing monetary compensation and other contracts like informed consent; analyzing EXAM's dynamic or sequential properties; optimally choosing sample size and treatment definitions (in addition to designing treatment assignment probabilities given the sample size and treatment definition); considering information frictions and psychological elements in patient preferences; and analyzing games among experimenters with experimental design as an action or strategy. I leave these challenging directions for future research.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge, "Sampling-based vs. Design-based Uncertainty in Regression Analysis," 2017. Working Paper.

Abdulkadiroğlu, Atila, Joshua Angrist, Yusuke Narita, and Parag A. Pathak, "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation," *Econometrica*, 2017, *85* (5), 1373–1432.

Angelucci, Manuela and Daniel Bennett, "The Marriage Market for Lemons: HIV Testing and Marriage in Rural Malawi," 2017. Working Paper.

Angrist, Joshua and Guido Imbens, "Sources of Identifying Information in Evaluation Models," 1991. Working Paper.

Ashraf, Nava, Dean Karlan, and Wesley Yin, "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," *Quarterly Journal of Economics*, 2006, *121* (2), 635–672.

__, James Berry, and Jesse M Shapiro, "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia," *American Economic Review*, 2010, *100* (5), 2383–2413.

Athey, Susan and Guido W Imbens, "The Econometrics of Randomized Experiments," *Handbook of Economic Field Experiments*, 2017, *1*, 73–140.

Azevedo, Eduardo and Eric Budish, "Strategyproofness in the Large," 2017. Working Paper.

Baicker, Katherine, Sarah L Taubman, Heidi L Allen, Mira Bernstein, Jonathan H Gruber, Joseph P Newhouse, Eric C Schneider, Bill J Wright, Alan M Zaslavsky, and Amy N Finkelstein, "The Oregon Experiment — Effects of Medicaid on Clinical Outcomes," *New England Journal of Medicine*, 2013, *368* (18), 1713–1722.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg, "Decision Theoretic Approaches to Experiment Design and External Validity," *Handbook of Economic Field Experiments*, 2017, *1*, 141–174.

Berry, James, Greg Fischer, and Raymond P Guiteras, "Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana," 2018. Working Paper.

Billingsley, Patrick, *Probability and Measure*, John Wiley & Sons, 2008.

Björklund, Anders, "Essays in Social Experimentation: What Experiments are Needed for Manpower Policy?," *Journal of Human Resources*, 1988, pp. 267–277.

__ and Robert Moffitt, "The Estimation of Wage Gains and Welfare Gains in Self-selection Models," *Review of Economics and Statistics*, 1987, pp. 42–49.

Blackwell, David and MA Girshick, *Theory of Games and Statistical Decisions*, Dover Publications, 1954.

Bó, Ernesto Dal, Frederico Finan, and Martín A Rossi, "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service," *Quarterly Journal of Economics*, 2013, *128* (3), 1169–1218.

Bogomolnaia, Anna and Hervé Moulin, "A New Solution to the Random Assignment Problem," *Journal of Economic Theory*, 2001, *100* (2), 295–328.

Bubeck, Sébastien and Nicolo Cesa-Bianchi, "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems," *Foundations and Trends in Machine Learning*, 2012, *5* (1), 1–122.

Budish, Eric, Gérard P Cachon, Judd B Kessler, and Abraham Othman, "Course Match: A Large-Scale Implementation of Approximate Competitive Equilibrium from Equal Incomes for Combinatorial Allocation," *Operations Research*, 2016.

_ , Yeon-Koo Che, Fuhito Kojima, and Paul Milgrom, "Designing Random Allocation Mechanisms: Theory and Applications," *American Economic Review*, 2013, *103* (2), 585–623.

Chakraborty, Bibhas and Erica EM Moodie, *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*, Springer, 2013.

Chan, Tat Y and Barton H Hamilton, "Learning, Private Information, and the Economic Evaluation of Randomized Experiments," *Journal of Political Economy*, 2006, *114* (6), 997–1040.

Chassang, Sylvain, Gerard Padró i Miquel, and Erik Snowberg, "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, 2012, *102* (4), 1279–1309.

Chilvers, Clair, Michael Dewey, Katherine Fielding, Virginia Gretton, Paul Miller, Ben Palmer, David Weller, Richard Churchill, Idris Williams, Navjot Bedi et al., "Antidepressant Drugs and Generic Counselling for Treatment of Major Depression in Primary Care: Randomised Trial with Patient Preference Arms," *British Medical Journal*, 2001, *322* (7289), 772.

Cohen, Jessica and Pascaline Dupas, "Free Distribution or Cost-sharing? Evidence from a Randomized Malaria Prevention Experiment," *Quarterly Journal of Economics*, 2010, pp. 1–45.

Cohen, Myron S, Ying Q Chen, Marybeth McCauley, Theresa Gamble, Mina C Hosseinipour, Nagalingeswaran Kumarasamy, James G Hakim, Johnstone Kumwenda, Beatriz Grinsztejn, Jose HS Pilotto et al., "Prevention of HIV-1 Infection with Early Antiretroviral Therapy," *New England Journal of Medicine*, 2011, *365* (6), 493–505.

Cox, Gertrude M and WG Cochran, *Experimental Designs*, Wiley, 1992.

Devoto, Florencia, Esther Duflo, Pascaline Dupas, William Parienté, and Vincent Pons, "Happiness on Tap: Piped Water Adoption in Urban Morocco," *American Economic Journal: Economic Policy*, 2012, *4* (4), 68–99.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics Research: A Toolkit," *Handbook of Development Economics*, 2007, *4*, 3895–3962.

Dupas, Pascaline, "Short-run Subsidies and Long-run Adoption of New Health Products: Evidence from a Field Experiment," *Econometrica*, 2014, *82* (1), 197–228.

Epstein, Ronald M and Ellen Peters, "Beyond Information: Exploring Patients' Preferences," *Journal of American Medical Association*, 2009, *302* (2), 195–197.

Food and Drug Administration, "Adaptive Design Clinical Trials for Drugs and Biologics," 2010.

__ , "Patient Preference Information," 2016.

Friedman, Lawrence M, Curt Furberg, David L DeMets, David M Reboussin, and Christopher B Granger, *Fundamentals of Clinical Trials*, Vol. 3, Springer, 1998.

Friedman, Milton, *Capitalism and Freedom*, University of Chicago Press, 1962.

Fryer, Roland G, "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments," *Handbook of Economic Field Experiments*, 2017, *2*, 95–322.

Glennerster, Rachel, "The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency," *Handbook of Economic Field Experiments*, 2017, *1*, 175–243.

Grant, Robert M, Javier R Lama, Peter L Anderson, Vanessa McMahan, Albert Y Liu, Lorena Vargas, Pedro Goicochea, Martín Casapía, Juan Vicente Guanira-Carranza, Maria E Ramirez-Cardich et al., "Preexposure Chemoprophylaxis for HIV Prevention in Men who Have Sex with Men," *New England Journal of Medicine*, 2010, *2010* (363), 2587–2599.

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

Haushofer, Johannes and Jeremy Shapiro, "The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya," *Quarterly Journal of Economics*, 2016, *131* (4), 1973–2042.

He, Yinghua, Antonio Miralles, Marek Pycia, and Jianye Yan, "A Pseudo-Market Approach to Allocation with Priorities," *American Economic Journal: Microeconomics*, 2017.

Heckman, James J and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

Henshaw, RC, SA Naji, IT Russell, and AA Templeton, "Comparison of Medical Abortion with Surgical Vacuum Aspiration: Women's Preferences and Acceptability of Treatment," *British Medical Journal*, 1993, *307* (6906), 714–717.

Hu, Feifang and William F Rosenberger, "Optimality, Variability, Power: Evaluating Response-adaptive Randomization Procedures for Treatment Comparisons," *Journal of the American Statistical Association*, 2003, *98* (463), 671–678.

__ **and** __ , *The Theory of Response-adaptive Randomization in Clinical Trials*, Vol. 525, John Wiley & Sons, 2006.

Hull, Peter, "IsoLATEing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons," 2018. Working Paper.

Hylland, Aanund and Richard J. Zeckhauser, "The Efficient Allocation of Individuals to Positions," *Journal of Political Economy*, 1979, *87(2)*, 293–314.

Imbens, Guido W and Donald B Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, 2015.

Jackson, Matthew O, "Incentive Compatibility and Competitive Allocations," *Economics Letters*, 1992, *40* (3), 299–302.

Kass, Michael A, Dale K Heuer, Eve J Higginbotham, Chris A Johnson, John L Keltner, J Philip Miller, Richard K Parrish, M Roy Wilson, and Mae O Gordon, "The Ocular Hypertension Treatment Study: A Randomized Trial Determines That Topical Ocular Hypotensive Medication Delays or Prevents the Onset of Primary Open-angle Glaucoma," *Archives of Ophthalmology*, 2002, *120* (6), 701–713.

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

King, Michael, Irwin Nazareth, Fiona Lampe, Peter Bower, Martin Chandler, Maria Morou, Bonnie Sibbald, and Rosalind Lai, "Impact of Participant and Physician Intervention Preferences on Randomized Trials: A Systematic Review," *Journal of American Medical Association*, 2005, *293* (9), 1089–1099.

Kitagawa, Toru and Aleksey Tetenov, "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 2018, *86* (2), 591–616.

Kowalski, Amanda E, "How to Examine External Validity Within an Experiment," 2016. Working Paper.

Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Peterson Zwane, "Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions," *Quarterly Journal of Economics*, 2011, *126* (1), 145–205.

Manski, Charles, *Identification for Prediction and Decision*, Cambridge: Harvard University Press, 2008.

Mogstad, Magne and Alexander Torgovitsky, "Identification and Extrapolation with Instrumental Variables," *Annual Review of Economics*, 2018.

Mollner, Joshua and E Glen Weyl, "Lottery Equilibrium," 2018. Working Paper.

Murphy, Susan A, "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2003, *65* (2), 331–355.

Narita, Yusuke, "(Non)Randomization: A Theory of Quasi-Experimental Evaluation of School Quality," 2016. Working Paper.

Preference Collaborative Review Group, "Patients' Preferences within Randomised Trials: Systematic Review and Patient Level Meta-analysis," *British Medical Journal*, 2008, *337*.

Robins, James, Liliana Orellana, and Andrea Rotnitzky, "Estimation and Extrapolation of Optimal Treatment and Testing Strategies," *Statistics in Medicine*, 2008, *27* (23), 4678–4721.

Roth, Alvin E, *Who Gets What and Why: The New Economics of Matchmaking and Market Design*, Houghton Mifflin Harcourt, 2015.

Rothstein, Jesse and Till von Wachter, "Social Experiments in the Labor Market," *Handbook of Economic Field Experiments*, 2017, *2*, 555–637.

Russo, Daniel J, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen et al., "A Tutorial on Thompson Sampling," *Foundations and Trends® in Machine Learning*, 2018, *11* (1), 1–96.

Scandinavian Simvastatin Survival Study Group and Others, "Randomised Trial of Cholesterol Lowering in 4444 Patients with Coronary Heart Disease: the Scandinavian Simvastatin Survival Study (4S)," *Lancet*, 1994, *344* (8934), 1383–1389.

Sherman, Lawrence W and David Weisburd, "General Deterrent Effects of Police Patrol in Crime "Hot Spots": A Randomized, Controlled Trial," *Justice Quarterly*, 1995, *12* (4), 625–648.

Stupp, Roger, Monika E Hegi, Warren P Mason, Martin J van den Bent, Martin JB Taphoorn, Robert C Janzer, Samuel K Ludwin, Anouk Allgeier, Barbara Fisher, Karl Belanger et al., "Effects of Radiotherapy with Concomitant and Adjuvant Temozolomide versus Radiotherapy Alone on Survival in Glioblastoma in a Randomised Phase III Study: 5-year Analysis of the EORTC-NCIC Trial," *Lancet Oncology*, 2009, *10* (5), 459–466.

Swift, Joshua K and Jennifer L Callahan, "The Impact of Client Treatment Preferences on Outcome: A Meta-analysis," *Journal of Clinical Psychology*, 2009, *65* (4), 368–381.

Train, Kenneth, *Discrete Choice Methods with Simulation*, Cambridge University Press, 2003.

Walters, Christopher, "The Demand for Effective Charter Schools," *Journal of Political Economy*, 2018.

Wei, LJ and S Durham, "The Randomized Play-the-winner Rule in Medical Trials," *Journal of the American Statistical Association*, 1978, *73* (364), 840–843.

White, John, *Bandit Algorithms for Website Optimization*, O'Reilly, 2012.

Writing Group for the Women's Health Initiative Investigators and Others, "Risks and Benefits of Estrogen plus Progestin in Healthy Postmenopausal Women: Principal Results from the Women's Health Initiative Randomized Controlled Trial," *Journal of American Medical Association*, 2002, *288* (3), 321–333.

Zelen, Marvin, "Play the Winner Rule and the Controlled Clinical Trial," *Journal of the American Statistical Association*, 1969, *64* (325), 131–146.

__ , "A New Design for Randomized Clinical Trials," *New England Journal of Medicine*, 1979, *300* (22), 1242–1245.

Table 1: Magnitude of the RCT Landscape

(a) Registered Medical Clinical Trials & Sample Sizes

	Sample Period 2007-2017 May
Total Number of Clinical Trials Registered	296,597
Sum of Sample Sizes	367,902,580

(b) Registered Social and Economic Experiments & Sample Sizes

	Sample Period 2007-2017 May
Total Number of Economic RCTs Registered	1055
Sum of Sample Sizes	22,190,304

Notes: Panel a provides summary statistics of clinical trials registered in the WHO International Clinical Trials Registry Platform (ICTRP, `http://www.who.int/ictrp/en/`, retrieved in May 2019). The sample consists of clinical trials registered there between January 1st 2007 to May 30th 2017. I exclude trials with registered sample size larger than five millions. Panel b provides summary statistics of economic RCTs registered in the American Economic Association RCT Registry (`https://www.socialscienceregistry.org`, retrieved in May 2019). The sample consists of RCTs registered there between January 1st 2007 to May 30th 2017 and where the unit of outcome measurement is an individual or a household. I focus on RCTs with individual or household subjects in order to make it possible to sum up sample sizes. See Section 2 for discussions about this exhibit and Appendix A.3.1 for the detailed computational procedure.

Table 2: A Selection of High-stakes Randomized Controlled Trials

(a) Medical Clinical Trials

	Treatment	Outcome	Treatment Effect
i	Cholesterol Lowering Drug	Mortality (in 5 Years)	30% Reduction
ii	Medication to Reduce Interocular Pressure	Visual Field Abnormality (in 6 Years)	59% Reduction
iii	HIV Prevention Drug	HIV Infection Rate (in 1 Year)	44% Reduction
iv	Antiretroviral Therapy	HIV Transmission Rate (in 1.7 Years)	96% Reduction
v	Hormone Therapy	Coronary Heart Disease (in 5 Years)	29% Increase

(b) Social and Economic Experiments

	Treatment	Outcome	Treatment Effect
I	Unconditional Cash Transfers	Consumption (9 Months Later)	22% Increase
II	Police Patrol	Crime Calls	10% Reduction
III	HIV Testing	Fertility (in 2-3 Years)	18% Increase
IV	Health Insurance (Medicaid)	Depression Rate (in 2 Years)	30% Reduction
V	High Wage Job Offer	Offer Acceptance Rate	29% Increase

Notes: This table lists examples illustrating the high-stakes nature of certain RCTs. Following the convention in the medical literature, treatment effects are measured relative to the average outcome in the control group, which I normalize to 100%. Every treatment effect is statistically significant at the 5% or lower level. The control is a placebo or the absence of any treatment unless otherwise noted below. See the following references for the details of each RCT:

Panel a Study i: Scandinavian Simvastatin Survival Study Group and Others (1994)
Panel a Study ii: Kass et al. (2002)
Panel a Study iii: Grant et al. (2010)
Panel a Study iv: Cohen et al. (2011)
Panel a Study v: Writing Group for the Women's Health Initiative Investigators and Others (2002)
Panel b Study I: Haushofer and Shapiro (2016)
Panel b Study II: Sherman and Weisburd (1995)
Panel b Study III: Angelucci and Bennett (2017)
Panel b Study IV: Baicker et al. (2013)
Panel b Study V: Dal Bó et al. (2013), where the control is a lower wage job offer.

See Section 2 for discussions about this table. Appendix Table A.2 provides further details.

Table 3: OLS Regression Estimates of Heterogeneous Treatment Effects

Dependent Variable: Incidence of Child Diarrhea in Past Week					
	(1)	(2)	(3)	(4)	(5) Main
Treatment	-0.045***	-0.045***	-0.046***	-0.044***	-0.045***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Treatment * latrine density		-0.061			-0.046
		(0.069)			(0.068)
Treatment * diarrhea prevention			-0.012***		-0.010**
			(0.004)		(0.004)
Treatment * mother's education				-0.007**	-0.006*
				(0.003)	(0.003)
Observations	6,750	6,750	6,750	6,742	6,742
Mean of dependent variable in comparison group	0.193	0.193	0.193	0.193	0.193

Notes: This table shows OLS regression estimates of heterogeneous treatment effects of spring protection. Data from all four survey rounds (2004, 2005, 2006, 2007), sample restricted to children under age three at baseline (in 2004) and children born since 2004 in sample households. Diarrhea defined as three or more "looser than normal" stools within 24 hours at any time in the past week. Different columns differ in the set of baseline household characteristics interacted with the treatment indicator. Every household characteristic is normalized to be mean zero so that the coefficient on the treatment indicator can be interpreted as the average treatment effect. The gender-age controls include linear and quadratic current age (by month), and these terms interacted with a gender indicator. I use specifications without additional controls. Stars *, **, and *** mean significance at 90%, 95%, and 99%, respectively, based on Huber-White robust standard errors clustered at the spring level. See Section 6.2 for the model description and discussions about this table.

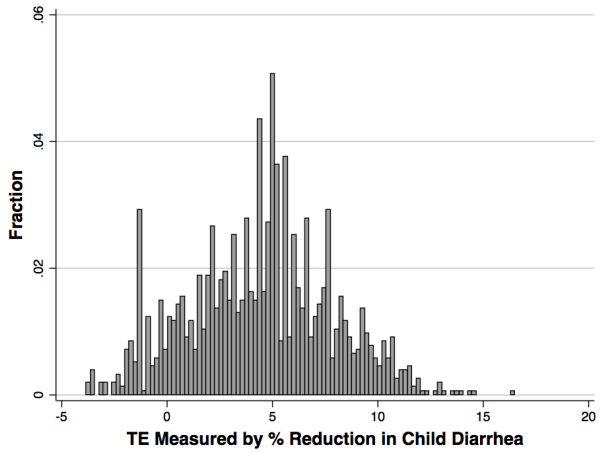
Table 4: Maximum Simulated Likelihood Estimates of Mixed Logit Model of Spring Choice

	(1)	(2)	(3)	(4) Main
Spring protection treatment indicator (Normal)				
Mean	2.205***	3.163***	2.999***	3.516***
	(0.213)	(0.235)	(0.288)	(0.308)
Standard Deviation	5.426***	5.702***	5.557***	5.741***
	(0.298)	(0.291)	(0.405)	(0.305)
Treatment * latrine density	7.533***			2.751**
	(0.939)			(1.178)
Treatment * diarrhea prevention		1.080***		0.565***
		(0.104)		(0.095)
Treatment * mother's education			0.650***	0.609***
			(0.066)	(0.069)
Distance to source, minutes walk (Restricted triangular)				
Mean	0.222***	0.220***	0.220***	0.221***
	(0.010)	(0.010)	(0.010)	(0.010)
Standard Deviation	0.222***	0.220***	0.220***	0.221***
	(0.010)	(0.010)	(0.010)	(0.010)
Source type: borehole/piped	-1.079***	-1.047***	-1.055***	-1.054***
	(0.135)	(0.136)	(0.139)	(0.133)
Source type: well	-1.924***	-1.954***	-1.943***	-1.944***
	(0.137)	(0.131)	(0.134)	(0.131)
Source type: stream/river	-1.422***	-1.387***	-1.443***	-1.393***
	(0.144)	(0.141)	(0.148)	(0.143)
Source type: lake/pond	-0.312	-0.313	-0.333	-0.299
	(0.269)	(0.273)	(0.274)	(0.406)
Number of observations (water collection choice situations)	53,427	53,427	53,427	53,427

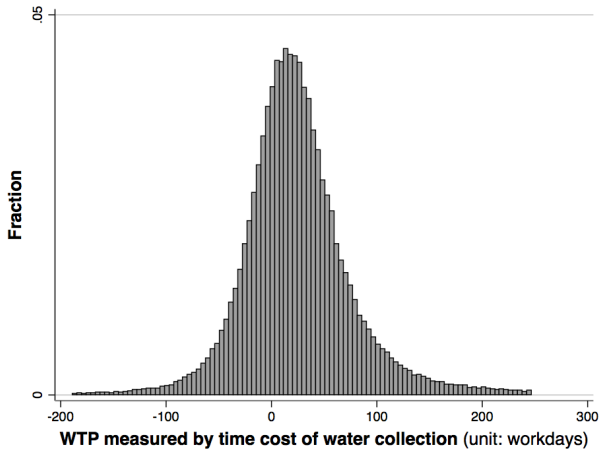
Notes: This table shows mixed logit estimates used for estimating heterogeneous WTP for the treatment. Each observation is a unique water collection trip recorded in the final round of household surveys (2007). The omitted water source category is non-program springs outside the target area of the experiment. Different columns differ in the set of baseline household characteristics interacted with the treatment indicator. The indicator for the spring that each household used at baseline is in the models, but its coefficient estimate is not shown in the table. Standard errors are based on the information matrix with the Hessian being estimated by the outer product of the gradient of the simulated likelihood at the estimated parameter value. Stars *, **, and *** mean significance at 90%, 95%, and 99%, respectively. See Section 6.2 for the model description and discussions about this table. See Appendix A.3.3 for the estimation procedure to produce these estimates.

Figure 1: Treatment Effects and WTP for the Treatment

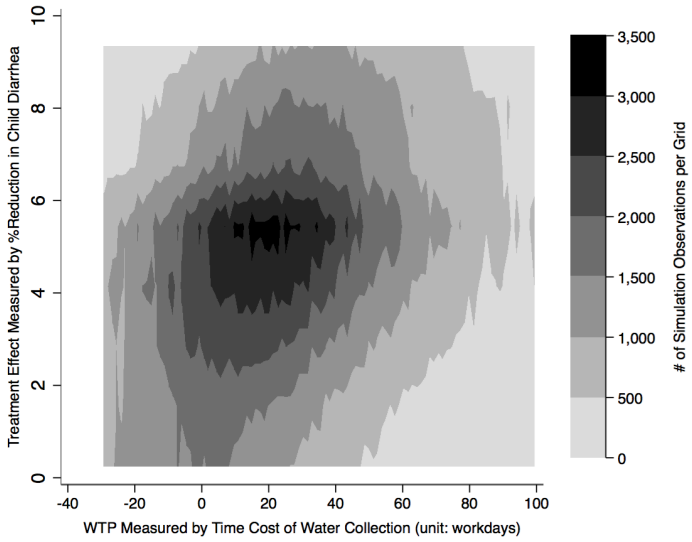
(a) Heterogeneity in Treatment Effects $\hat{e}_{t_1 i}$



(b) Heterogeneity in WTP $\hat{w}_{it_1}$



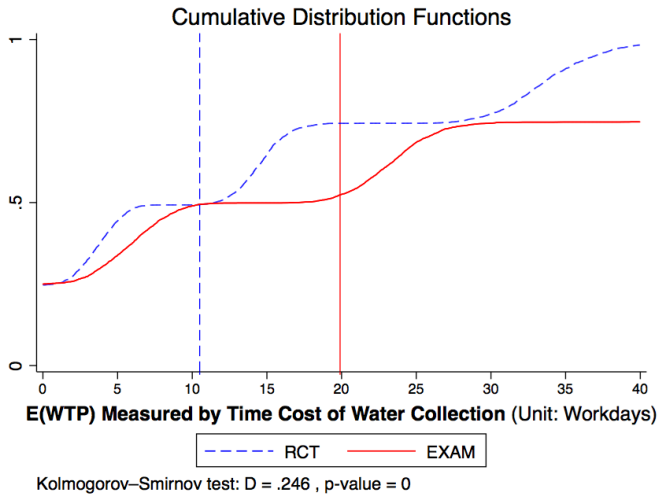
(c) Correlation between Treatment Effects & WTP



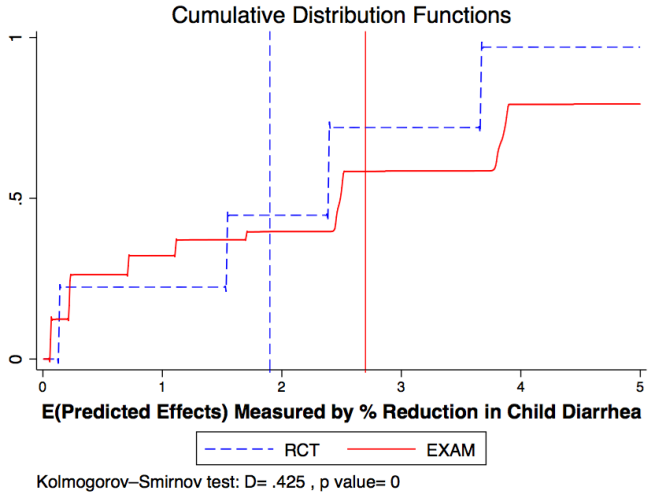
Notes: This figure shows the pattern of heterogeneity in estimated WTP $\hat{w}_{it_1}$ and predicted treatment effects $\hat{e}_{t_1 i}$. Panel a is about the predicted treatment effects $\hat{e}_{t_1 i}$ measured in percentage point reduction in the incidence of child diarrhea in the past week, while Panel b is about WTP for the spring protection treatment $\hat{w}_{it_1}$, measured by time cost of water collection in the unit of workdays. Both predicted effects $\hat{e}_{t_1 i}$ and WTP $\hat{w}_{it_1}$ are based on the main statistical specifications including all of the interactions between the treatment indicator and household characteristics (baseline latrine density, diarrhea prevention knowledge score, and mother's years of education). Panel c demonstrates the correlation between WTP $\hat{w}_{it_1}$ and predicted treatment effects $\hat{e}_{t_1 i}$. For the sake of visibility, I focus on the three standard deviations around the mean. See Section 6.2 for discussions about this figure. See Appendix A.3.3 for the detailed computational procedure.

Figure 2: EXAM vs RCT: Welfare

(a) Average WTP for Assigned Treatments w_i^*



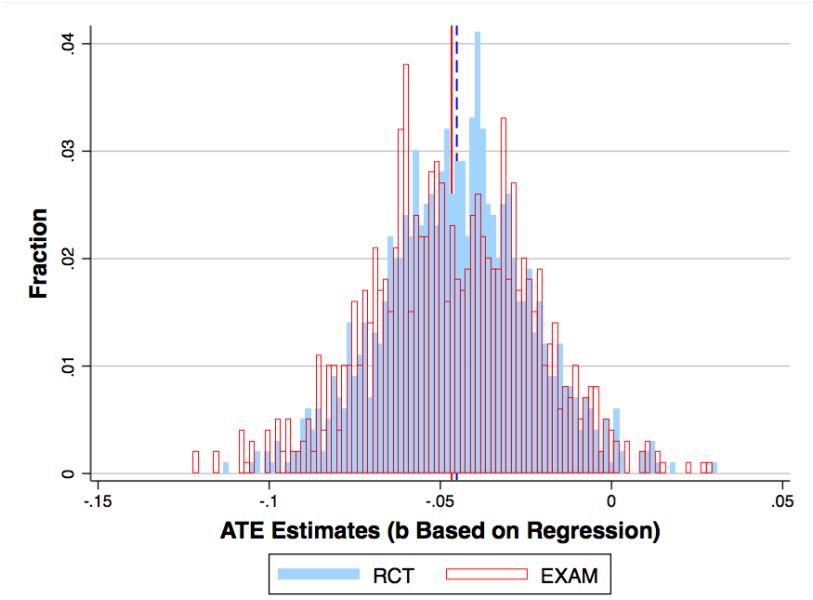
(b) Average Predicted Effects of Assigned Treatments e_i^*



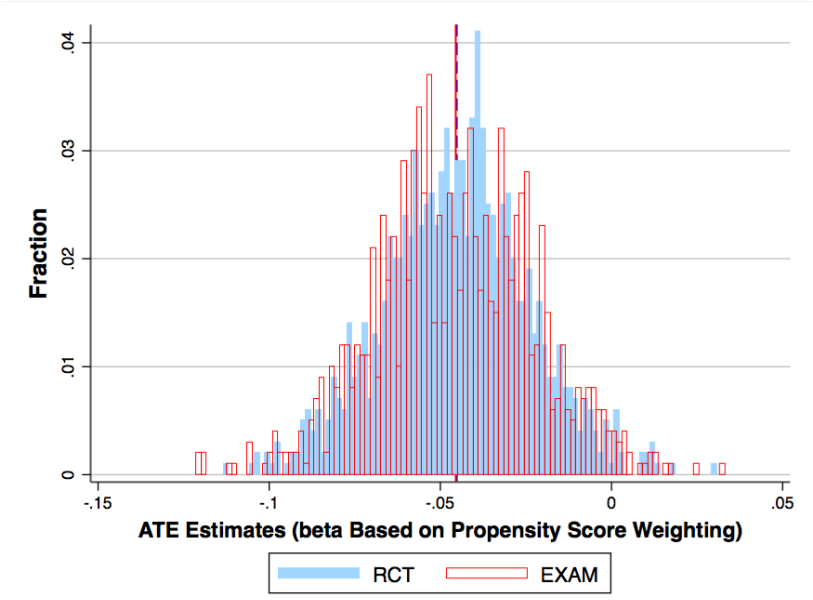
Notes: To compare EXAM and RCT's welfare performance, this figure shows the distribution of average subject welfare over 1000 bootstrap simulations under each experimental design. Panel a measures welfare with respect to average WTP w_i^* for assigned treatments while Panel b with respect to average predicted effects e_i^* of assigned treatments. A dotted line indicates the distribution of each welfare measure for RCT while a solid line indicates that for EXAM. Each vertical line represents mean. Kolmogorov-Smirnov tests find the EXAM and RCT distributions to be significantly different both for w_i^* and e_i^*. Both predicted effects $\hat{e}_{t_1 i}$ and WTP $\hat{w}_{it_1}$ are based on the main statistical specifications including all of the interactions between the treatment indicator and household characteristics (baseline latrine density, diarrhea prevention knowledge score, and mother's years of education). See Section 6.3 for discussions about this figure.

Figure 3: EXAM vs RCT: Average Treatment Effect Estimates

(a) Distribution of Treatment Effect Estimates $\hat{b}^*$



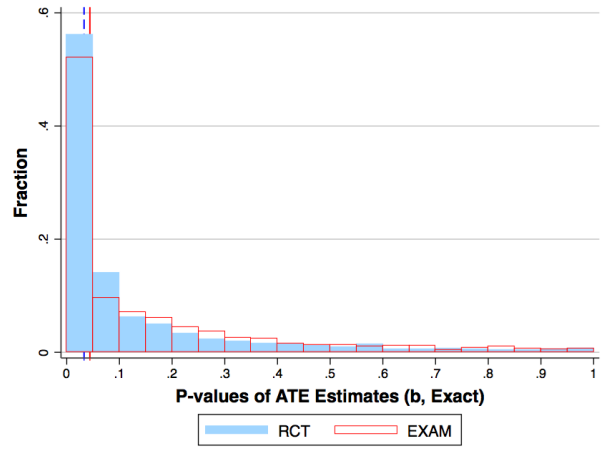
(b) Distribution of Average Treatment Effect Estimates $\hat{\beta}^*$



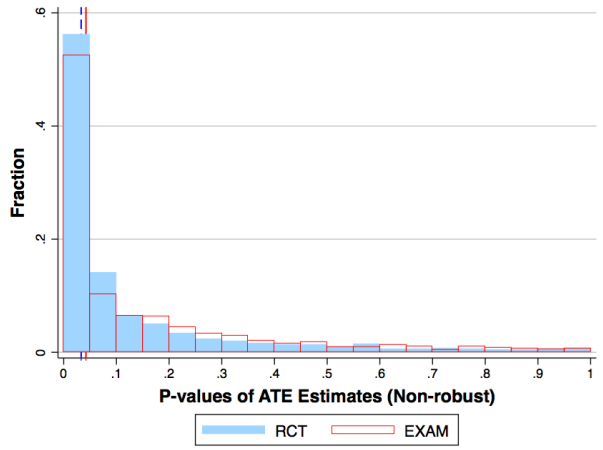
Notes: This figure compares EXAM and RCT's causal inference performance by showing the distribution of average treatment effect estimates under each experimental design. Grey bins indicate average treatment effect estimates for RCT while transparent bins with black outlines indicate those for EXAM. The solid vertical line indicates the mean for EXAM while the dashed vertical line indicates that for RCT. See Section 6.3 for discussions about this figure.

Figure 4: EXAM vs RCT: p Values for $\hat{b}^*$

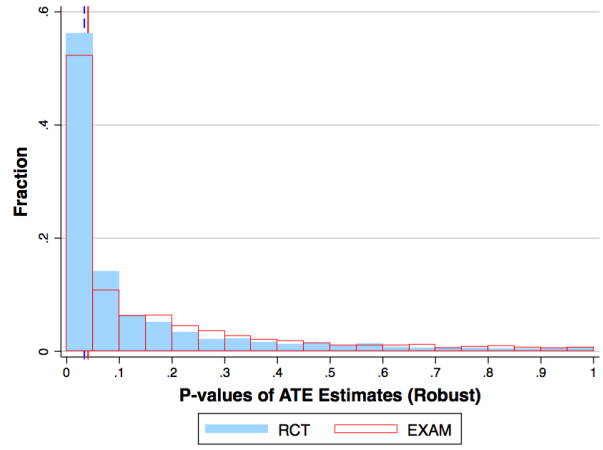
(a) Exact, Finite Sample p Values



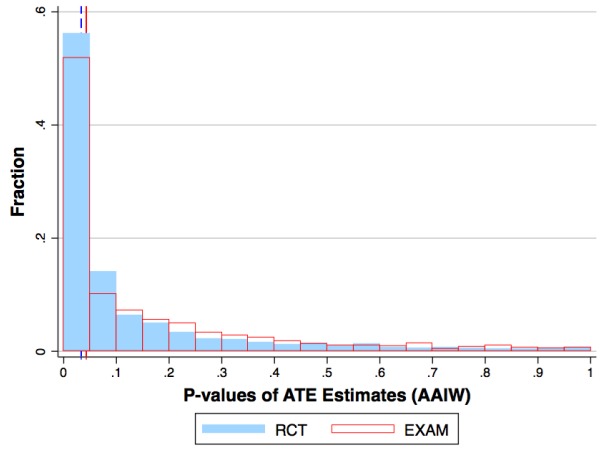
(b) Non-robust p Values



(c) Robust p Values



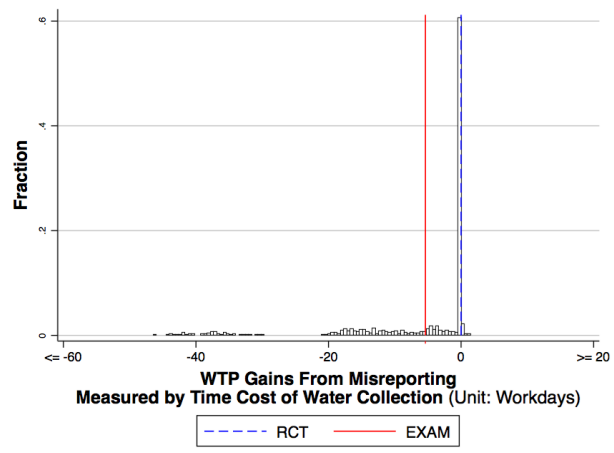
(d) Abadie et al. (2017)'s p Values

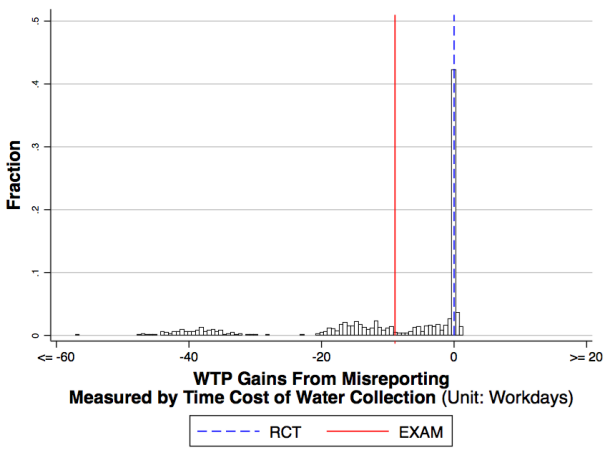


Notes: This figure compares EXAM and RCT's causal inference performance by showing the distribution of p values accompanying treatment effect estimates $\hat{b}^*$ under each experimental design. The p values are based on exact, non-robust, robust, or Abadie et al. (2017)'s finite population causal standard errors. Grey bins indicate p values for RCT while transparent bins with black outlines indicate those for EXAM. The solid vertical line indicates median for EXAM while the dashed vertical line indicates that for RCT. See Section 6.3 for discussions about this figure.

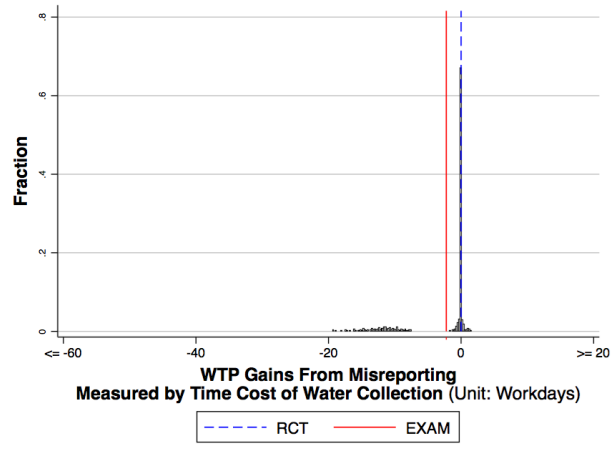
Figure 5: EXAM vs RCT: Incentive

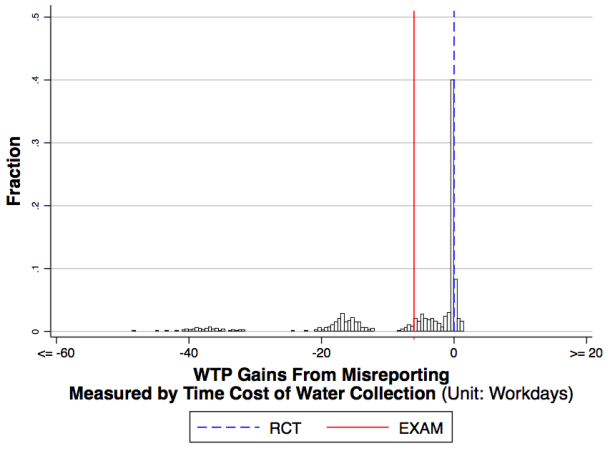
(a) WTP manipulation $\sim$ true WTP+$N(0, 100)$ (b) WTP manipulation $\sim$ true WTP+$N(0, 1000)$





(c) WTP manipulation $\sim$ true WTP+$U(0, 100)$ (d) WTP manipulation $\sim$ true WTP+$U(-100, 0)$





Notes: This figure shows the histogram of true WTP gains from potential WTP misreporting to EXAM, quantifying the incentive compatibility of EXAM. Different panels use different ways of drawing WTP manipulations indicated by the panel titles. Each solid vertical line represents the mean WTP gain from potential WTP misreporting to EXAM. The dash vertical line is for RCT, where the true WTP gain from any WTP misreport is zero. See Section 6.3 for discussions about this figure.

Table 5: EXAM vs RCT: Incentive Details

	$\Delta w/w_{it_1}$ True WTP Gains from Misreporting Relative to True WTP for Treatment t_1					
	95 Percentile	96 Percentile	97 Percentile	98 Percentile	99 Percentile	Max
Figure 5 Panel a Manipulation $\sim N(0, 100)$	0.00%	0.00%	0.10%	0.34%	0.63%	1.55%
Figure 5 Panel b Manipulation $\sim N(0, 1000)$	0.92%	1.12%	1.32%	1.42%	1.55%	1.82%
Figure 5 Panel c Manipulation $\sim U(0, 100)$	0.69%	0.89%	1.20%	1.41%	1.54%	2.04%
Figure 5 Panel d Manipulation $\sim U(-100, 0)$	0.76%	0.98%	1.24%	1.46%	1.64%	1.87%

Notes: This table provides additional details about Figure 5. In particular, for each scenario in Figure 5, this table shows the most profitable true WTP gains from potential WTP misreporting to EXAM, quantifying the incentive compatibility of EXAM. See Section 6.3 for discussions about this table.

A Appendix (For Online Publication)

A.1 Methodological Details

Throughout Appendix A.1, I impose the simplifying assumption that $\sum_i p^*_{it}(\epsilon) \in \mathbb{Z}$ for every t. It is possible to dispense with this assumption with additional notational burden.

A.1.1 Propositions 4 and 5: Generalizations

This section extends Proposition 5 to a general case where $p_t n_p$ (the expected number of subjects with propensity vector p and assigned to treatment t under EXAM) may not be an integer. Let $N_{pt} \equiv \sum_i 1\{p^*_i(\epsilon) = p\} D_{it}$ be a random variable that stands for the number of subjects with propensity vector p and assigned to treatment t. Denote the realization of N_{pt} by $n_{pt} \equiv \sum_i 1\{p^*_i(\epsilon) = p\} d_{it}$ where d_{it} is the realization of D_{it}. Let $\underline{n}_{pt}$ be the greatest integer less than or equal to $p_t n_p$. With this regularity condition, I extend Definition 2 as follows to use EXAM to draw a deterministic treatment assignment and the associated n_{pt}'s.

Definition 2 (EXAM Continued; Generalization). Starting from the end of Definition 2 in Section 3.2, draw a treatment assignment from $p^*_{it}(\epsilon)$ as follows. First apply Budish et al. (2013)'s algorithm (in their Appendix B) to draw $(n_{pt}) \in \mathbb{N}$ that satisfy the following properties (I detail their algorithm and its use below):

- $n_{pt} = p_t n_p$ for all p and t such that $p_t n_p \in \mathbb{N}$.
- $n_{pt} \in \{\underline{n}_{pt}, \underline{n}_{pt} + 1\}$ for all p and t such that $p_t n_p \notin \mathbb{N}$.
- $\sum_t n_{pt} = n_p$ for all p.
- $\sum_p n_{pt} = \sum_i p^*_{it}(\epsilon)$ for all t.
- $E(n_{pt}) = p_t n_p$ for all p and t.

Given the drawn values of (n_{pt}), for each propensity vector p,

- I uniformly randomly pick n_{pt_0} subjects from $\{i | p^*_i(\epsilon) = p\}$ and assign them to t_0.

For each subsequent step $k = 1, ..., m$,

- Step k: From the remaining $n_p - \sum_{t=t_0}^{t_{k-1}} n_{pt}$ subjects, I uniformly randomly pick n_{pt_k} subjects and assign them to t_k.

When $p_t n_p$ is an integer for all p and t, $n_{pt} = p_t n_p$ always holds for all p and t so that this generalized definition reduces to Definition 2 in Section 5. With this extended Definition 2, Proposition 4 holds as it is in the main text. I obtain the following characterization of the variance of ATE estimator $\hat{\beta}_t^*$, which nests Proposition 5 in Section 5.

Proposition 5 (Generalization). Suppose that the data-generating process is EXAM $p^*(\epsilon) \equiv (p_{it}^*(\epsilon))_{it}$ with any given $\epsilon > 0$. $\hat{\beta}_t^*$ is an unbiased estimator of the average treatment effect with the following variance.

$$V(\hat{\beta}_t^* | p^*(\epsilon)) = \sum_p \delta_p^2 \Big\{ \sum_{t' \in \{t_0, t\}} \Big[\Big(\frac{S_{pt'}^2}{\underline{n}_{pt'}} \Big)(1 - p_{t'} n_p + \underline{n}_{pt'}) + \Big(\frac{S_{pt'}^2}{\underline{n}_{pt'} + 1} \Big)(p_{t'} n_p - \underline{n}_{pt'}) \Big] - \frac{S_{ptt_0}^2}{n_p} \Big\}.$$

where recall that $\delta_p \equiv n_p / n$, $\bar{Y}_p(t) \equiv \dfrac{\sum_{i: p_i^*(\epsilon) = p} Y_i(t)}{n_p}$ is the mean of $Y_i(t)$ in the subpopulation with propensity p, $S_{pt}^2 \equiv \dfrac{\sum_{i: p_i^*(\epsilon) = p} (Y_i(t) - \bar{Y}_p(t))^2}{n_p - 1}$ is the variance of $Y_i(t)$ in the subpopulation, and $S_{ptt'}^2 \equiv \dfrac{\sum_{i: p_i^*(\epsilon) = p} (Y_i(t) - Y_i(t') - (\bar{Y}_p(t) - \bar{Y}_p(t')))^2}{n_p - 1}$ is the variance of $Y_i(t) - Y_i(t')$ in the subpopulation.

Using Budish et al. (2013)'s Algorithm

In the above generalized Definition 2, I use Budish et al. (2013)'s algorithm to draw (n_{pt}). To do so, I embed my setting into their notation as follows: $N \equiv \{p \in [0,1]^{m+1} |$ there exists some subject i such that $p_i^*(\epsilon) = p\}$ is the set of "agents" in their terminology. Let $O \equiv \{t_0, t_1, ..., t_m\}$ be the set of "objects." $\mathcal{H} \equiv \{\mathcal{H}_1, \mathcal{H}_2\}$ is a "constraint structure" where $\mathcal{H}_1 \equiv \mathcal{H}_0 \cup \mathcal{H}_1'$, $\mathcal{H}_0 \equiv \{\{(p,t)\}_{p \in N, t \in O}\}$, $\mathcal{H}_1' \equiv \{(p,t) | t \in O\}_{p \in N}$, and $\mathcal{H}_2 \equiv \{(p,t) | p \in N\}_{t \in O}$. Upper and lower constraints are as follows.

- $\bar{q}_s = 1$ and $\underline{q}_s = 0$ if $s \in \mathcal{H}_0$.
- $\bar{q}_s = \underline{q}_s = n_p - \sum_t \underline{n}_{pt}$ if $s \in \mathcal{H}_1'$.
- $\bar{q}_s = \underline{q}_s = \sum_i p_{it}^*(\epsilon) - \sum_p \underline{n}_{pt}$ if $s \in \mathcal{H}_2$.

Budish et al. (2013) show that applying their algorithm to this problem produces (x_{pt}) such that $n_{pt} \equiv \underline{n}_{pt} + x_{pt}$ satisfies the properties in Definition 2. For completeness, I define their algorithm; see their Appendix B for more details. I first construct a network flow as follows. Let $\Omega \equiv \{\{(p,t)\}_{p \in N, t \in O}\}$ and $X = (x_\omega)_{\omega \in \Omega}$. The set of vertices is composed of the

source s and the sink s', two vertices v_ω and $v_{\omega'}$ for each element $\omega \in \Omega$, and v_S for each $S \in \mathcal{H} \setminus [(\cup_{\omega\in\Omega}\{\omega\}) \cup (N \times O)]$. I place (directed) edges according to the following rule.

- For each $\omega \in \Omega$, an edge $e = (v_\omega, v_{\omega'})$ is placed from v_ω to $v_{\omega'}$.
- For each $k = 1, 2$, an edge $e = (v_S, v_{S'})$ is placed from S to $S' \neq S$ where $S, S' \in \mathcal{H}_k$, if $S' \subset S$ and there is no $S'' \in \mathcal{H}_k$ where $S' \subset S'' \subset S$.
- An edge $e = (s, v_S)$ is placed from the source s to v_S if $S \in \mathcal{H}_1$ and there is no $S' \in \mathcal{H}_1$ where $S \subset S'$.
- An edge $e = (v_S, s')$ is placed from v_S to the sink s' if $S \in \mathcal{H}_2$ and there is no $S' \in \mathcal{H}_2$ where $S \subset S'$.

I associate flow with each edge as follows. For each edge $e = (v_\omega, v_{\omega'})$, I associate flow $x_e = x_\omega$. For each e that is not of the form $(v_\omega, v_{\omega'})$ for some $\omega \in \Omega$, the flow x_e is (uniquely) set to satisfy the flow conservation, that is, for each vertex v different from s and s', the sum of flows into v is equal to the sum of flows from v. I use this network flow to define the following algorithm.

Definition 3 (Budish et al. (2013)'s Algorithm)**.** If $\deg[X(\mathcal{H})] \equiv |\{S \in \mathcal{H} | x_s \in \mathbb{Z}\}| = |\mathcal{H}|$, then stop the algorithm. Otherwise, move on to the following steps:

(1) *Cycle-Finding Procedure*

(a) *Step 0*: Since $\deg[X(\mathcal{H})] < |\mathcal{H}|$ by assumption, there exists an edge $e_1 = (v_1, v_1')$ such that its associated flow x_{e_1} is fractional. Define an edge $f_1 = (v_1, v_1')$ from v_1 to v_1'.

(b) *Step* $t = 1, ...$: Consider the vertex v_t' that is the destination of edge f_t.

i) If v_t' is the origin of some edge $f_{t'} \in \{f_1, ..., f_{t-1}\}$, then stop. The procedure has formed a cycle $(f_{t'}, f_{t'+1}, ..., f_t)$ composed of edges in $\{f_1, ..., f_t\}$. Proceed to *Termination-Cycle Procedure* below.

ii) Otherwise, since the flow associated with f_t is fractional by construction and the flow conservation holds at v_t', there exists an edge $e_{t+1} = (u_{t+1}, u_{t+1}') \neq e_t$ with fractional flow such that v_t' is either its origin or destination. Draw an edge f_{t+1} by $f_{t+1} = e_{t+1}$ if v_t' is the origin of e_{t+1} and $f_{t+1} = (u_{t+1}', u_{t+1})$ otherwise. Denote $f_{t+1} = (v_{t+1}, v_{t+1}')$.

(2) *Termination-Cycle Procedure*

(a) Construct a set of flows associated with edges (x_e^1) which is the same as (x_e), except for flows $(x_{e_\tau})_{t'\leq\tau\leq t}$, that is, flows associated with edges that are involved in the cycle from the last step. For each edge e_τ such that $f_\tau = e_\tau$, set $x^1_{e_\tau} = x_{e_\tau} + \alpha$, and each edge e_τ such that $f_\tau \neq e_\tau$, set $x^1_{e_\tau} = x_{e_\tau} - \alpha$, where $\alpha > 0$ is the largest real number such that the induced expected assignment $X^1 = (x^1_\omega)_{\omega\in\Omega}$ still satisfies all constraints in $\mathcal{H}$. By construction, $x^1_S = x_S$ if x_S is an integer, and there is at least one constraint set $S \in \mathcal{H}$ such that x^1_S is an integer while x_S is not. Thus $\deg[X^1(\mathcal{H})] > \deg[X(\mathcal{H})]$.

(b) Construct a set of flows associated with edges (x_e^2) which is the same as (x_e), except for flows $(x_{e_\tau})_{t'\leq\tau\leq t}$, that is, flows associated with edges that are involved in the cycle from the last step. For each edge e_τ such that $f_\tau = e_\tau$, set $x^1_{e_\tau} = x_{e_\tau} - \beta$, and each edge e_τ such that $f_\tau \neq e_\tau$, set $x^1_{e_\tau} = x_{e_\tau} + \beta$, where $\beta > 0$ is the largest real number such that the induced expected assignment $X^2 = (x^2_\omega)_{\omega\in\Omega}$ still satisfies all constraints in $\mathcal{H}$. By construction, $x^2_S = x_S$ if x_S is an integer, and there is at least one constraint set $S \in \mathcal{H}$ such that x^2_S is an integer while x_S is not. Thus $\deg[X^2(\mathcal{H})] > \deg[X(\mathcal{H})]$.

(c) Set γ by $\gamma\alpha + (1-\gamma)(-\beta) = 0$, i.e., $\gamma = \dfrac{\beta}{\alpha+\beta}$.

(d) Decompose X into $X = \gamma X^1 + (1-\gamma)X^2$.

Note that $\deg[X^k(\mathcal{H})] > \deg[X(\mathcal{H})]$ for both $k = 1, 2$, implying that repeating the above algorithm transforms the original X into a distribution over deterministic (x_{pt})'s where every x_{pt} is an integer. The induced distribution can then be used to draw deterministic (x_{pt}) consistent with X. Budish et al. (2013)'s Theorem 1 and Appendix B show that the resulting (x_{pt}) has the property that $n_{pt} \equiv \underline{n}_{pt} + x_{pt}$ satisfies the conditions in Definition 2.

A.1.2 Asymptotic Power Comparison of EXAM and RCT

Section 5 illustrates that EXAM's ATE estimation is potentially more precise than RCT's in a finite sample. This appendix shows the same point in an asymptotic framework.

Sequence of Experimental Design Problems

Following Abadie et al. (2017), I consider a sequence of finite populations of potential subjects indexed by population size N. For each population N, I randomly sample subjects who participate in the experiment. Let $R_{N,i}$ denote the indicator of subject i being sampled from population N, i.e., $R_{N,i} = 1$ if i is sampled and $R_{N,i} = 0$ otherwise. Denote the number

of subjects by $n = \sum_{i=1}^{N} R_{N,i}$. Given each finite population N, I consider a sequence of experimental design problems, each of which consists of

- A set of n experimental subjects $\{i|R_{N,i} = 1\}$.
- Experimental treatments $t_0, t_1, ..., t_m$.
- Each treatment t's pseudo capacity $c_{N,t} \in \mathbb{N}$ with $\Sigma_{t=t_0}^{t_m} c_{N,t} = n$.
- Each subject i's WTP $w_{N,it}$ for each $i \in \{j|R_{N,j} = 1\}$.
- Each treatment t's predicted treatment effect $e_{N,ti}$ for each $i \in \{j|R_{N,j} = 1\}$.

Among these components, experimental treatments are nonrandom and do not depend on N or n. The other elements are random because $R_{N,i}$ is random. I allow $c_{N,t}$ to be random even conditional on $\{i|R_{N,i} = 1\}$.

I study a sequence of experimental designs $p_N = (p_{N,it})_{i:R_{N,i}=1,t=t_0,...,t_m}$ along with the sequence of experimental design problems. p_N is random because some of the components of an experimental design problem is random. For each sampled experimental design problem and each $i \in \{j|R_{N,j} = 1\}$, I use $D_{N,it} = 1$ to indicate that subject i is assigned to treatment t, and $D_{N,it} = 0$ to indicate that subject i is assigned to any other treatment or control. The distribution of $(D_{N,it})$ depends on the algorithm to draw deterministic treatment assignments. Let $Y_{N,i}(t)$ be the fixed potential outcome of subject i that would be observed if i is sampled from population N and assigned to treatment t. The observed outcome of subject i in the sample is $Y_{N,i} = \sum_{t=t_0}^{t_m} D_{N,it} Y_{N,i}(t)$. I observe $(Y_{N,i}, D_{N,i}, w_{N,i}, e_{N,i})$ for each subject i in the sample.

Sequence of Parameters and Estimators

I consider a sequence of two parameters as estimands, the *population average treatment effect* and the *sample average treatment effect*, defined as follows:

$$\beta_{N,t}^{pop} = \frac{1}{N} \sum_{i=1}^{N} (Y_{N,i}(t) - Y_{N,i}(t_0)) \text{ and } \beta_{N,t}^{sample} = \frac{1}{n} \sum_{i=1}^{N} R_{N,i}(Y_{N,i}(t) - Y_{N,i}(t_0)).$$

Let $\beta_N^{pop} = (\beta_{N,t_1}^{pop}, ..., \beta_{N,t_m}^{pop})'$ and $\beta_N^{sample} = (\beta_{N,t_1}^{sample}, ..., \beta_{N,t_m}^{sample})'$. Note that $\beta_{N,t}^{pop}$ is nonrandom while $\beta_{N,t}^{sample}$ is random due to the random sampling of a subject sample. I put the following assumption.

Assumption 1. *There exist G sequences of nonempty subpopulations, $\{P_{N,1}\}, ..., \{P_{N,G}\}$, such that for all N, (i) $P_{N,1}, ..., P_{N,G}$ form a partition of population N, (ii) for all g, for all $i, j \in P_{N,g}$, I have $(w_{N,it}, e_{N,ti})_t = (w_{N,jt}, e_{N,tj})_t$, and (iii) if $(w_{N,it}, e_{N,ti})_t = (w_{N,jt}, e_{N,tj})_t$ for some $i \in P_{N,g}$ and $j \in P_{N,g'}$, then $g = g'$.*

Denote the size of $P_{N,1}, ..., P_{N,G}$ by $N_1, ..., N_G$, respectively. Let $n_g = \sum_{i \in P_{N,g}} R_{N,i}$ be the number of subjects sampled from subpopulation $P_{N,g}$. Now consider two parameters defined on subpopulation $P_{N,g}$:

$$\beta_{N,gt}^{pop} = \frac{1}{N_g} \sum_{i \in P_{N,g}} (Y_{N,i}(t) - Y_{N,i}(t_0)) \text{ and } \beta_{N,gt}^{sample} = \frac{1}{n_g} \sum_{i \in P_{N,g}} R_{N,i}(Y_{N,i}(t) - Y_{N,i}(t_0)).$$

Let $\beta_{N,g}^{pop} = (\beta_{N,gt_1}^{pop}, ..., \beta_{N,gt_m}^{pop})'$ and $\beta_{N,g}^{sample} = (\beta_{N,gt_1}^{sample}, ..., \beta_{N,gt_m}^{sample})'$. The population average treatment effect $\beta_{N,t}^{pop}$ and the sample average treatment effect $\beta_{N,t}^{sample}$ can be written as the weighted average of $\beta_{N,gt}^{pop}$ and $\beta_{N,gt}^{sample}$, respectively:

$$\beta_{N,t}^{pop} = \sum_{g=1}^{G} \frac{N_g}{N} \beta_{N,gt}^{pop} \text{ and } \beta_{N,t}^{sample} = \sum_{g=1}^{G} \frac{n_g}{n} \beta_{N,gt}^{sample}.$$

As in Section 5, I estimate both $\beta_{N,t}^{pop}$ and $\beta_{N,t}^{sample}$ with

$$\hat{\beta}_{N,t}^{*} = \sum_{g=1}^{G} \frac{n_g}{n} \hat{\beta}_{N,gt}^{*},$$

where

$$\hat{\beta}_{N,gt}^{*} = \frac{\sum_{i \in P_{N,g}} R_{N,i} D_{N,it} Y_{N,i}}{\sum_{i \in P_{N,g}} R_{N,i} D_{N,it}} - \frac{\sum_{i \in P_{N,g}} R_{N,i} D_{N,it_0} Y_{N,i}}{\sum_{i \in P_{N,g}} R_{N,i} D_{N,it_0}}.$$

I assume that if two subjects are in different subpopulations, EXAM gives them different assignment probabilities and puts them in different subsamples. Let $\hat{\beta}_{N,g}^{*} = (\hat{\beta}_{N,gt_1}^{*}, ..., \hat{\beta}_{N,gt_m}^{*})'$.

Asymptotic Distribution of $\hat{\beta}_{N,t}^{*}$

To derive the asymptotic distribution of $\hat{\beta}_{N,t}^{*}$, I need a series of regularity conditions. I first assume that each subject is sampled independetly with the same sampling probability, and the expected sample size of each subsample goes to infinity as N goes to infinity. Let $\delta_{N,g} = N_g/N$.

Assumption 2. *(i) There is a sequence of sampling probabilities, ρ_N, such that for all $r \in \{0,1\}^N$,*

$$\Pr(R_N = r) = \rho_N^{\sum_{i=1}^N r_i}(1-\rho_N)^{N-\sum_{i=1}^N r_i}.$$

(ii) For all g, $N_g\rho_N \to \infty$, $\rho_N \to \rho \in [0,1]$ and $\delta_{N,g} \to \delta_g \in [0,1]$ as $N \to \infty$.

I apply EXAM to each realized experimental design problem to obtain treatment assignment probabilities. Denote the assignment probabilities by $p_N^*(\epsilon)$. I impose the following restriction on the distribution of capacities conditional on sample. Below expectations are taken over $R_N \equiv (R_{N,1}, ..., R_{N,N})'$, $(c_{N,t})$ and $(D_{N,it})$.

Assumption 3. *For all g, there is a sequence of constant vectors of size $m+1$, $q_{N,g}$, such that $E[p_{N,it}^*(\epsilon)|R_N = r] = q_{N,g,t}$ for all t, all $r \in \{0,1\}^N$, and all $i \in P_{N,g} \cap \{j|r_j = 1\}$.*

For each subject $i \in P_{N,g} \cap \{k|R_{Nk} = 0\}$, define $p_{N,it}^*(\epsilon)$ as $p_{N,it}^*(\epsilon) = p_{N,jt}^*(\epsilon)$ for an arbitrary $j \in P_{N,g} \cap \{k|R_k = 1\}$. I also define random variable $D_{N,it}$ with $i \in \{k|R_k = 0\}$ such that the following assumption is true and treatment assignments are independent across subjects given assignment probabilities $p_N^*(\epsilon)$.

Assumption 4. *$(D_{N,it})_{t=t_0,...,t_m}$ and $(D_{N,jt})_{t=t_0,...,t_m}$ are independent for any $i \neq j$ conditional on $R_N = r$.*

I put a few additional regularity conditions.

Assumption 5. *For all g, there exists some $\delta > 0$ such that the sequence $\frac{1}{N_g}\sum_{i\in P_{N,g}} E[|Y_{N,i}|^{4+\delta}]$ is bounded.*

Now let $X_{N,it} = D_{N,it} - E[D_{N,it}]$, $D_{N,i} = (D_{N,i,t_1}, ..., D_{N,i,t_m})'$, $X_{N,i} = (X_{N,i,t_1}, ..., X_{N,i,t_m})'$, and for each g,

$$\Omega_{N,g} = \frac{1}{N_g}\sum_{i\in P_{N,g}} E\left[\begin{pmatrix} Y_{N,i} \\ X_{N,i} \\ 1 \end{pmatrix}\begin{pmatrix} Y_{N,i} \\ X_{N,i} \\ 1 \end{pmatrix}'\right].$$

Assumption 6. *For all g, $\Omega_{N,g} \to \Omega_g$, where the limit is full rank.*

Let $\beta_{N,it} = Y_{N,i}(t) - Y_{N,i}(t_0)$ and $\beta_{N,i} = (\beta_{N,i,t_1}, ..., \beta_{N,i,t_m})'$. For all g and all $i \in P_{N,g}$, let

$$\epsilon_{N,i} = \sum_{t=t_0}^{t_m} D_{N,it}\big(Y_{N,i}(t) - \frac{1}{N_g}\sum_{i\in P_{N,g}} Y_{N,i}(t)\big).$$

Let $\Delta_g^{cond} = \lim_{N\to\infty} \frac{1}{N_g} \sum_{i\in P_{N,g}} Var(X_{N,i}\epsilon_{N,i})$ and $\Delta_g^{ehw} = \lim_{N\to\infty} \frac{1}{N_g} \sum_{i\in P_{N,g}} E(X_{N,i}\epsilon_{N,i}^2 X_{N,i}')$.

Assumption 7. *For all g, Δ_g^{cond} and Δ_g^{ehw} exist and are positive definite.*

Assumption 8. *For all g, $\sqrt{n}(\frac{n_g}{n} - \frac{N_g}{N})\beta_{N,g}^{pop} \xrightarrow{p} 0$ as $N \to \infty$.*

Proposition 6 (Asymptotic Distribution of $\hat{\beta}_{N,t}^*$)**.** *Suppose Assumptions 1, 2, 3, 4, 5, 6, 7 and 8 hold. Let $H_g = \lim_{N\to\infty} \frac{1}{N_g} \sum_{i\in P_{N,g}} E(X_{N,i}X_{N,i}')$. Then,*

$$\sqrt{n}(\hat{\beta}_{N,t_j}^* - \beta_{N,t_j}^{pop}) \xrightarrow{d} \mathcal{N}(0, V_{1,jj}^*),$$

where $V_{1,jj}^*$ is the j-th diagonal element of $V_1^* \equiv \sum_{g=1}^G \delta_g H_g^{-1}(\rho\Delta_g^{cond} + (1-\rho)\Delta_g^{ehw})H_g^{-1}$.

$$\sqrt{n}(\hat{\beta}_{N,t_j}^* - \beta_{N,t_j}^{sample}) \xrightarrow{d} \mathcal{N}(0, V_{2,jj}^*),$$

where $V_{2,jj}^*$ is the j-th diagonal element of $V_2^* \equiv \sum_{g=1}^G \delta_g H_g^{-1}\Delta_g^{cond}H_g^{-1}$.

Asymptotic Efficiency Comparison of EXAM and RCT

How does EXAM compare to RCT in terms of asymptotic standard errors? With RCT, the estimator for $\beta_{N,t}^{pop}$ and $\beta_{N,t}^{sample}$ is

$$\hat{\beta}_{N,t}^{RCT} = \frac{\sum_{i:R_{N,i}=1} D_{N,it}Y_{N,i}}{\sum_{i:R_{N,i}=1} D_{N,it}} - \frac{\sum_{i:R_{N,i}=1} D_{N,it_0}Y_{N,i}}{\sum_{i:R_{N,i}=1} D_{N,it_0}}.$$

$\hat{\beta}_{N,t}^{RCT}$ is a special case of $\hat{\beta}_{N,t}^*$ when $w_{N,it} = w_{N,jt} > 0$ and $e_{N,ti} = e_{N,tj}$ for all i, j and t (recall Proposition 1). For this RCT special case, Assumption 1 holds with $G = 1$ and $P_{N,1}$ being the set of all subjects. Assumption 8 also holds since $\sqrt{n}(\frac{n_g}{n} - \frac{N_g}{N})\beta_{N,g}^{pop} = 0$ for all g and N. Let $\epsilon_{N,i} = D_{N,i}'(\beta_{N,i} - \beta_N^{pop}) + Y_{N,i}(t_0) - \frac{1}{N}\sum_{i=1}^N Y_{N,i}(t_0)$, $\Delta^{cond} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N Var(X_{N,i}\epsilon_{N,i})$ and $\Delta^{ehw} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N E(X_{N,i}\epsilon_{N,i}^2 X_{N,i}')$. Proposition 6 therefore implies the following result.

Corollary 3 (Asymptotic Distribution of $\hat{\beta}_{N,t}^{RCT}$)**.** *Suppose Assumptions 2, 3, 4, 5, 6 and 7 hold with $G = 1$ and $p_{N,it}^*(\epsilon) = c_{N,t}/N$ for all i, t, and N. Let $H = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N E(X_{N,i}X_{N,i}')$.*

$$\sqrt{n}(\hat{\beta}_{N,t_j}^{RCT} - \beta_{N,t_j}^{pop}) \xrightarrow{d} \mathcal{N}(0, V_{1,jj}^{RCT}),$$

where $V_{1,jj}^{RCT}$ *is the* j*-th diagonal element of* $V_1^{RCT} \equiv H^{-1}(\rho\Delta^{cond} + (1-\rho)\Delta^{ehw})H^{-1}$.

$$\sqrt{n}(\hat{\beta}_{N,t_j}^{RCT} - \beta_{N,t_j}^{sample}) \xrightarrow{d} \mathcal{N}(0, V_{2,jj}^{RCT}),$$

where $V_{2,jj}^{RCT}$ *is the* j*-th diagonal element of* $V_2^{RCT} \equiv H^{-1}\Delta^{cond}H^{-1}$.

To compare EXAM and RCT by their asymptotic variances, consider a simple situation where there is only one treatment. For EXAM, let

$$q_{g,t} = \lim_{N\to\infty} q_{N,g,t}, S_{gt}^2 = \lim_{N\to\infty} \frac{\sum_{i\in P_{N,g}}(Y_{N,i}(t) - \bar{Y}_{N,g}(t))^2}{N_g} \text{ and}$$

$$S_{gtt'}^2 = \lim_{N\to\infty} \frac{\sum_{i\in P_{N,g}}(Y_{N,i}(t) - Y_{N,i}(t') - (\bar{Y}_{N,g}(t) - \bar{Y}_{N,g}(t')))^2}{N_g},$$

where $\bar{Y}_{N,g}(t) = \dfrac{\sum_{i\in P_{N,g}} Y_{N,i}(t)}{N_g}$. For RCT, let

$$q_t = \lim_{N\to\infty} q_{N,t}, S_t^2 = \lim_{N\to\infty} \frac{\sum_{i=1}^{N}(Y_{N,i}(t) - \bar{Y}_N(t))^2}{N} \text{ and}$$

$$S_{tt'}^2 = \lim_{N\to\infty} \frac{\sum_{i=1}^{N}(Y_{N,i}(t) - Y_{N,i}(t') - (\bar{Y}_N(t) - \bar{Y}_N(t')))^2}{N},$$

where $q_{N,t} = E[c_{N,t}/n]$ and $\bar{Y}_N(t) = \dfrac{\sum_{i=1}^{N} Y_{N,i}(t)}{N}$. I assume these limits exist.

Assumption 9. *For all* g, q_{g,t_1}, $S_{gt_1}^2$, $S_{gt_0}^2$ *and* $S_{gt_1t_0}^2$ *exist with* $q_{g,t_1} \in (0,1)$.

Assumption 10. q_{t_1}, $S_{t_1}^2$, $S_{t_0}^2$ *and* $S_{t_1t_0}^2$ *exist with* $q_{t_1} \in (0,1)$.

Corollary 4 (Binary Treatment Case)**.** *Suppose there is only one treatment* t_1 *to be compared to the control* t_0*. For EXAM, under Assumptions 1, 2, 3, 4, 5, 6, 7, 8 and 9,*

$$\sqrt{n}(\hat{\beta}_{N,t_1}^{*} - \beta_{N,t_1}^{pop}) \xrightarrow{d} \mathcal{N}(0, \sum_{g=1}^{G} \delta_g(\frac{S_{gt_1}^2}{q_{g,t_1}} + \frac{S_{gt_0}^2}{1-q_{g,t_1}} - \rho S_{gt_1t_0}^2)). \tag{7}$$

$$\sqrt{n}(\hat{\beta}_{N,t_1}^{*} - \beta_{N,t_1}^{sample}) \xrightarrow{d} \mathcal{N}(0, \sum_{g=1}^{G} \delta_g(\frac{S_{gt_1}^2}{q_{g,t_1}} + \frac{S_{gt_0}^2}{1-q_{g,t_1}} - S_{gt_1t_0}^2)). \tag{8}$$

For RCT, suppose Assumptions 2, 3, 4, 5, 6, 7 and 10 hold with $G = 1$ *and* $p_{N,it}^{*}(\epsilon) = c_{N,t}/N$ *for all* i *and* t*. Then,*

$$\sqrt{n}(\hat{\beta}_{N,t_1}^{RCT} - \beta_{N,t_1}^{pop}) \xrightarrow{d} \mathcal{N}(0, \frac{S_{t_1}^2}{q_{t_1}} + \frac{S_{t_0}^2}{1-q_{t_1}} - \rho S_{t_1t_0}^2). \tag{9}$$

$$\sqrt{n}(\hat{\beta}^{RCT}_{N,t_1} - \beta^{sample}_{N,t_1}) \xrightarrow{d} \mathcal{N}(0, \frac{S^2_{t_1}}{q_{t_1}} + \frac{S^2_{t_0}}{1-q_{t_1}} - S^2_{t_1 t_0}). \tag{10}$$

The asymptotic variance comparison of RCT and EXAM depends on the limiting distribution of potential outcomes and treatment assignment probabilities. As is the case with other stratified experiments, EXAM may produce more precise ATE estimates than RCT if potential outcomes are well correlated with EXAM's treatment assignment probabilities. The following example illustrates this possibility, providing an asymptotic analogue of Example 1.

Example 2. Suppose there is only one treatment t_1, and $\rho_N = 1$ for all N so that $n = N$ with probability one. Every subject has $Y_{N,i}(t_0) = 0$ for all N. The subjects in every population are divided into four groups A, B, C and D based on their potential outcomes $Y_{N,i}(t_1)$. Let $Y_{N,i}(t_1) = 1, 2, 3$, and 4 for anybody in group A, B, C and D, respectively. Denote the number of subjects in group A, B, C and D by N_A, N_B, N_C and N_D, respectively. The sequences of pseudo capacities and the size of groups A, B, C and D are as follows:

- If $N = 4k$ for some $k \in \mathbb{N}$, $c_{N,t_0} = c_{N,t_1} = 2k$ and $N_A = N_B = N_C = N_D = k$.
- If $N = 4k+1$ for some $k \in \mathbb{N}$, $c_{N,t_0} = 2k$, $c_{N,t_1} = 2k+1$, $N_A = k+1$ and $N_B = N_C = N_D = k$.
- If $N = 4k+2$ for some $k \in \mathbb{N}$, $c_{N,t_0} = 2k+1$, $c_{N,t_1} = 2k+1$, $N_A = N_B = k+1$ and $N_C = N_D = k$.
- If $N = 4k+3$ for some $k \in \mathbb{N}$, $c_{N,t_0} = 2k+1$, $c_{N,t_1} = 2k+2$, $N_A = N_B = N_C = k+1$ and $N_D = k$.

Assume the experimenter imperfectly predicts treatment effects: $e_{N,t_1 i} = 0$ for every i in group A or B while $e_{N,t_1 i} = 2$ for every i in group C or D. Let $w_{N,it_1} = 1$ for every i in group A or B and $w_{N,it_1} = 2$ for every i in group C or D. There are two subpopulations, $P_{N,1}$ and $P_{N,2}$, such that $i \in P_{N,1}$ for every i in group A or B and $i \in P_{N,2}$ for every i in group C or D. For every N, EXAM with $\epsilon < .2$ gives the following treatment assignment probabilities[36]: $p^*_{N,it_1}(\epsilon) = 0.2$ for every $i \in P_{N,1}$ while $p^*_{N,it_1}(\epsilon) = 0.8$ for every $i \in P_{N,2}$. Under RCT, $p^{RCT}_{N,t_1} = c_{N,t_1}/N = q_{N,t_1}$. Note that $\delta_{N,1} = N_1/N = (N_A + N_B)/N \to 0.5$, $\delta_{N,2} = N_2/N = (N_C + N_D)/N \to 0.5$, $q_{N,1,t_1} \to 0.2$, $q_{N,2,t_1} \to 0.8$ and $q_{N,t_1} \to 0.5$ as $N \to \infty$.

[36] EXAM outputs these treatment probabilities if I set $\alpha = -\frac{15b}{8}$, $\beta_{t_1} = 5b$, and $\beta_{t_0} = 0$ given any budget b. Note that the capacity constraint holds, i.e., $\sum_{i=1}^{N} p^*_{N,it_1} \leq c_{N,t_1}$.

Applying Corollary 4 to this example, I have

$$aVar(\hat{\beta}^{*}_{N,t_1}) = \frac{0.53125}{N} < \frac{1.25}{N} = aVar(\hat{\beta}^{RCT}_{N,t_1}).$$

where $aVar$ is the asymptotic variance relative either β^{pop}_{N,t_1} or β^{sample}_{N,t_1}, both of which produce the same asymptotic variance by $\rho = 1$. The above inequality means that EXAM's ATE estimation may be asymptotically more precise than RCT's even if the experimenter imperfectly predicts potential outcomes.

A.1.3 Uncertainty in Predicted Effects and Preferences

Unlike the baseline setting in the main body, the experimenter's information about preferences and predicted effects may be uncertain and probabilistic. What experimental design should the experimenter use with uncertain preferences and predicted effects? An *uncertain experimental design problem* consists of experimental subjects, treatments, pseudo capacities, and the following objects.

- Each subject i's *preference or WTP* $\tilde{w}_{it}$ for treatment t where $\tilde{w}_{it}$ is a random variable.
- Each treatment t's *predicted treatment effect* $\tilde{e}_{ti}$ for subject i where $\tilde{e}_{ti}$ is a random variable.

$\tilde{w}_{it}$ and $\tilde{e}_{ti}$ are the experimenter's statistical perceptions about WTP and predicted treatment effect, respectively. Denote $w_{it} \equiv E(\tilde{w}_{it})$ and $e_{ti} \equiv E(\tilde{e}_{ti})$ where each expectation is with respect to the distribution of $\tilde{w}_{it}$ and $\tilde{e}_{ti}$, respectively.

When I apply EXAM to (w_{it}, e_{ti}), the resulting EXAM nests RCT, is efficient with respect to (w_{it}, e_{ti}), is approximately incentive compatible, and is as informative as RCT in the same senses as in Propositions 1-5.

A.1.4 Ordinal Predicted Effects and Preferences

The experimenter's information about preferences and predicted effects may be ordinal. What experimental design should the experimenter use with ordinal preferences and predicted effects? An *ordinal experimental design problem* consists of experimental subjects, treatments, pseudo capacities, and the following objects.

- Each subject i's *ordinal preference* $\succsim_i$ for treatment t where $t \succsim_i t'$ means subject i weakly prefers treatment t over t'. $\succsim_i$ may involve ties and indifferences.

- Each treatment t's *ordinal predicted treatment effect* $\succsim_t$ for subject i where $i \succsim_t i'$ means treatment t is predicted weakly more effective for subject i than for subject i'. Again, $\succsim_t$ may involve ties and indifferences.

I consider the following adaptation of EXAM to this ordinal experimental design problem.

Definition 4 (Ordinal EXAM)**.** (1) Create any cardinal WTP w'_{it} of each subject i for each treatment t so that $w'_{it} > w'_{it'}$ if and only if $t \succ_i t'$.

(2) Create any cardinal predicted effect of each treatment t for each subject i so that $e'_{ti} > e'_{ti'}$ if and only if $i \succ_t i'$.

(3) Run EXAM (as defined in Definition 2) on (w'_{it}, e'_{ti}) to get treatment assignment probabilities $p^{*o}_{it}(\epsilon)$

Ordinal EXAM nests RCT, is approximately incentive compatible, and is as informative as RCT in the same senses as in Propositions 1, 3, and 4, respectively. For approximate incentive compatibility, I modify the setting so that subjects report ordinal preferences $\succsim_i$ instead of cardinal WTP w'_{it}. Moreover, ordinal EXAM has the following nice welfare property with respect to ordinal preferences and predicted effects.

Proposition 7. *$p^{*o}_{it}(\epsilon)$ is ordinally efficient in the following sense. There is no other experimental design (p_{it}) with $p_{it} \in [\epsilon, 1-\epsilon]$ for all subject i and treatment t, $\sum_i p_{it} \leq c_t$ for all $t = t_1, ..., t_m,$, and with the following better welfare property: For all cardinal WTP w_{it} consistent with ordinal $\succsim_i$ and all cardinal predicted effects e_{ti} consistent with ordinal $\succsim_t$, I have*

$$\sum_t p_{it} w_{it} \geq \sum_t p^{*o}_{it}(\epsilon) w_{it} \text{ and } \sum_t p_{it} e_{ti} \geq \sum_t p^{*o}_{it}(\epsilon) e_{ti}$$

for all i with at least one strict inequality.

A.2 Proofs

Proof of Proposition 1

Suppose to the contrary that there exist some $\epsilon \in [0, \bar{\epsilon}]$, i, and t such that $p^*_{it}(\epsilon) \neq p^{RCT}_t$. Since $e_{ti} = e_{tj}$ for all subjects i and j and treatment t, I have $\pi_{te_{ti}} \equiv \alpha e_{ti} + \beta_t = \alpha e_{tj} + \beta_t \equiv \pi_{te_{tj}}$ for all subjects i and j and treatment t. Combined with $w_{it} = w_{jt}$ for all subjects i and j and treatment t, this implies that any subjects i and j face the same utility maximization problem:

$$\arg\max_{p_i \in P}(\textstyle\sum_t p_{it} w_{it} \text{ s.t. } \sum_t p_{it}\pi_{te_{ti}} \leq b) = \arg\max_{p_j \in P}(\sum_t p_{jt} w_{jt} \text{ s.t. } \sum_t p_{jt}\pi_{te_{tj}} \leq b).$$

This implies $p^*_{jt}(\epsilon) = p^*_{it}(\epsilon) \neq p^{RCT}_t = c_t/n$ by the requirement in Definition 2 that $(p^*_{it})_t = (p^*_{jt})_t$ for any i and j with $w_i = w_j$ and $e_i = e_j$.

If $p^*_{jt}(\epsilon) = p^*_{it}(\epsilon) > c_t/n$ for some $t \neq t_0$, then $\sum_{j=1}^n p^*_{jt}(\epsilon) = np^*_{it}(\epsilon) > nc_t/n = c_t$, which implies $\sum_{j=1}^n p^*_{jt} > c_t$ (since $\sum_{j=1}^n p^{RCT}_t = c_t$). This contradicts the capacity constraint in the definition of p^*_{it}. If $p^*_{jt}(\epsilon) = p^*_{it}(\epsilon) < c_t/n$, then there is another treatment $t' \neq t$ for which $p^*_{jt'}(\epsilon) = p^*_{it'}(\epsilon) > c_t/n$ since $\sum_t c_t/n = \sum_t p^*_{jt}(\epsilon) = 1$ for any subject j. This implies that $\sum_{j=1}^n p^*_{jt'}(\epsilon) = np^*_{it'}(\epsilon) > nc_{t'}/n = c_{t'}$, again contradicting the capacity constraint if $t' \neq t_0$.

The only remaining possibility is $p^*_{jt_0}(\epsilon) = p^*_{it_0}(\epsilon) > c_{t_0}/n \geq \epsilon > 0$. This implies $p^*_{jt}(\epsilon) = p^*_{it}(\epsilon) < c_t/n \leq 1 - \epsilon$ for some $t \neq t_0$ and so $p^*_{jt} = p^*_{it} < c_t/n$ (since $p^{RCT}_t = c_t/n$ for any i). But this is a contradiction since j can increase the value of her objective function $\sum_t p_{jt} w_{jt}$ by changing $p^*_{jt_0}$ and p^*_{jt} to $p^*_{jt_0} - \delta$ and $p^*_{jt} + \delta$, respectively, for small enough $\delta > 0$, since $w_{jt} > w_{jt_0} = 0$. Such $p^*_{jt_0} - \delta$ and $p^*_{jt} + \delta$ satisfy the budget constraint since $\pi_{te_{ti}} \leq 0$ for every i and so

$$\sum_{t' \neq t_0, t} p^*_{jt'}\pi_{t'e_{t'j}} + (p^*_{jt_0} - \delta)\pi_{t_0 e_{t_0 j}} + (p^*_{jt} + \delta)\pi_{te_{tj}} \leq \sum_{t'} p^*_{jt'}\pi_{t'e_{t'j}} \leq b.$$

Therefore, it cannot be the case that $p^*_{jt_0}(\epsilon) = p^*_{it_0}(\epsilon) > c_{t_0}/n$. Thus, for every $\epsilon \in [0, \bar{\epsilon}], i$, and t, it must be the case that $p^*_{it}(\epsilon) = p^{RCT}_t$.

Proof of Proposition 2

EXAM always exists: It is enough to find (p^*_{it}) that satisfies the conditions in Step (1) of Definition 2.

Lemma 1. *There exists (p^*_{it}) that satisfies a weaker version of Definition 2 that is the same as Definition 2 except that EXAM breaks ties or indifferences so that every subject i's p_i solves the utility maximization problem with the minimum expenditure $\sum_t p_{it}\pi_{te_{ti}}$ (but it is not necessarily the case $(p^*_{it})_t = (p^*_{jt})_t$ for any i and j with $w_i = w_j$ and $e_i = e_j$).*

Proof of Lemma 1. Fix α at any negative constant $\alpha^* < 0$. Fix $\beta_{t_0} = 0$. Define a space of possible values of $\beta \equiv (\beta_t)_t$ by $B \equiv \beta_{t_0} \times [0, nb - \alpha^*\bar{e}1\{\bar{e} > 0\}]^m$ where $\bar{e} = \max\{e_{ti}\}$. For any given $\gamma \geq 0$, define the demand correspondence for each subject i by $p^*_i(\beta, \gamma) \equiv \text{argmax}_{p_i \in P} \sum_t (p_{it} w_{it} - \gamma p_{it}(\alpha^* e_{ti} + \beta_t))$ s.t. $\sum_t p_{it}(\alpha^* e_{ti} + \beta_t) \leq b$. Define the excess demand correspondence $z_\gamma(\cdot) : B \to \mathbb{R}^{m+1}$ by

$$z_\gamma(\beta) = \{\sum_i p_i - c | p_i \in p^*_i(\beta, \gamma) \text{ for every } i\} \equiv \sum_i p^*_i(\beta, \gamma) - c$$

where $c \equiv (c_t)$. This correspondence $z_\gamma(\cdot)$ is upper hemicontinuous in β and convex-valued because it is a linear finite sum of $p_i^*(\beta, \gamma)$'s, which are upper hemicontinuous and convex-valued as shown below.

Step 1. *For every subject i and $\gamma \geq 0$, her demand correspondence $p_i^*(\beta, \gamma)$ is nonempty, convex-valued, and upper hemicontinuous in β.*

Proof of Step 1. $p_i^*(\beta, \gamma)$ is convex-valued since for any $p_i, p_i' \in p_i^*(\beta, \gamma) \subset P$ and $\delta \in [0, 1]$, it holds that $p_{\delta,i} = \delta p_i + (1-\delta)p_i'$ is in $p_i^*(\beta, \gamma)$ because the objective $\sum_t p_{\delta,it} w_{it} - \gamma p_{\delta,it}(\alpha^* e_{ti} + \beta_t) = \sum_t (w_{it} - \gamma(\alpha^* e_{ti} + \beta_t)) p_{\delta,it}$ is linear and $p_{\delta,i}$ satisfies the budget constraint because $\sum_t p_{\delta,it}(\alpha^* e_{ti} + \beta_t) = \delta \sum_t p_{it}(\alpha^* e_{ti} + \beta_t) + (1-\delta) \sum_t p_{it}'(\alpha^* e_{ti} + \beta_t) \leq \delta b + (1-\delta) b = b$. $p_i^*(\beta, \gamma)$ is non-empty and upper-hemicontinuous by the maximum theorem. To see this, note that (1) the utility function is linear and (2) the correspondence from β to the choice set $\{p_i \in P | \sum_t p_{it}(\alpha^* e_{ti} + \beta_t) \leq b\}$ is both upper-hemicontinuous and lower-hemicontinuous as well as compact-valued and nonempty ($(p_{it})_{t=t_0,t_1,...,t_m} = (1, 0, ..., 0)$ is for free and always in the choice set). Thus the maximum theorem implies that $p_i^*(\beta, \gamma)$ is non-empty and upper-hemicontinuous, completing the proof of Step 1. □

Let $\bar{c} \equiv \max c_t$ and $\tilde{B} \equiv 0 \times [-\bar{c}, n(b + \bar{c}) - \alpha^* \bar{e} 1\{\bar{e} > 0\}]^m$. Define a truncation function $f : \tilde{B} \to B$ by $f(\beta) \equiv 0 \times (\max\{0, \min\{\beta_t, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}\}\})_{t=t_1,...,t_m}$. Define correspondence $g_\gamma : \tilde{B} \to B$ by $g_\gamma(\beta) \equiv f(\beta) + z_\gamma(f(\beta))$.

Step 2. *For all $\gamma \geq 0$, g_γ has a fixed point $\beta_\gamma^* \in g_\gamma(\beta_\gamma^*)$.*

Proof of Step 2. $z_\gamma(f(\beta))$ is upper hemicontinuous and convex-valued as a function of $\beta \in \tilde{B}$ because $f(\cdot)$ is continuous and $z_\gamma(\cdot)$ is an upper hemicontinuous and convex-valued correspondence, as explained above. This implies that $g_\gamma(\beta)$ is upper hemicontinuous and convex-valued as well. The range of $g_\gamma(\beta)$ lies in $\tilde{B}$, i.e., $g_\gamma : \tilde{B} \to \tilde{B}$. It is because

- $f(\beta) \equiv 0 \times (\max\{0, \min\{\beta_t, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}\}\})_{t=t_1,...,t_m} \in [0, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}]^{m+1}$, which is by $nb - \alpha^* \bar{e} 1\{\bar{e} > 0\} \geq 0$ (recall $\alpha^* < 0$).
- $\bar{c} \equiv \max c_t \geq 1$.
- $z_\gamma(f(\beta)) \in [-\bar{c}, n]^{m+1}$ because, for any $\beta \in \tilde{B}$ and t, the excess demand $z_{t,\gamma}(\beta)$ is at least $-\bar{c}$ (since the supply of any treatment t is $c_t \leq \bar{c}$ by definition) and at most n (since there are n subjects and the demand for any treatment t by any subject i is at most 1).

Finally, $\tilde{B}$ is nonempty by $-\bar{c} < 0 < n(b+\bar{c}) \leq n(b+\bar{c}) - \alpha^* \bar{e} 1\{\bar{e} > 0\}$. $g_\gamma(\beta) \equiv f(\beta) + z_\gamma(f(\beta))$ is therefore an upper hemicontinuous, nonempty, and convex-valued correspondence defined on the non-empty, compact, and convex set $\tilde{B}$. By Kakutani's fixed point theorem, there exists a fixed point $\beta^*_\gamma \in g_\gamma(\beta^*_\gamma)$, proving Step 2. □

Step 3. *For any sequence of $\gamma_n > 0$ with $\lim_{n\to\infty} \gamma_n = 0$, consider the associated sequence of fixed points $\beta^*_{\gamma_n} \in g_{\gamma_n}(\beta^*_{\gamma_n})$. There exists a subsequence of $(\beta^*_{\gamma_n})$ that converges to some β^*. Any such limit β^* is a fixed point of g_0 in the sense that $\beta^* \in g_0(\beta^*)$.*

Proof of Step 3. The space of possible values of $\beta^*_{\gamma_n}$ is $B \equiv \beta_{t_0} \times [0, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}]^m$, which is compact. The Bolzano-Weierstrass Theorem therefore implies the existence of a convergent subsequence of $(\beta^*_{\gamma_n})$.

The last part follows from $\beta^* \in \lim_{n\to\infty} g_{\gamma_n}(\beta^*) \subset g_0(\beta^*)$, which I show below, where $\lim_{n\to\infty} g_{\gamma_n}(\beta^*)$ is the set-theoretic limit define by $\{\beta | \lim_{n\to\infty} 1\{\beta \in g_{\gamma_n}(\beta^*)\} = 1\}$. This set-theoretic limit exists by the following reason: Since $\beta^* \in B$ and $f(\beta^*) = \beta^*$, I have

$$g_{\gamma_n}(\beta^*) \equiv f(\beta^*) + z_{\gamma_n}(f(\beta^*)) = \beta^* + z_{\gamma_n}(\beta^*) = \beta^* + \sum_i p^*_i(\beta^*, \gamma_n) - c.$$

For proving the existence of $\lim_{n\to\infty} g_{\gamma_n}(\beta^*)$, it is enough to show $\lim_{n\to\infty} p^*(\beta^*, \gamma_n)$ exists. To show it, note that if $p^1 \in p^*(\beta^*, \gamma_j)$ and $p^1 \notin p^*(\beta^*, \gamma_k)$ for $\gamma_j > \gamma_k > 0$, then for all γ_l with $\gamma_k > \gamma_l > 0$, I have

$$p^1 \notin p^*(\beta^*, \gamma_l),$$

which is true by the following reason: $p^1 \in p^*(\beta^*, \gamma_j)$ and $p^1 \notin p^*(\beta^*, \gamma_k)$ imply there exists some p^2 satisfying the budget constraint and such that

$$\sum_t p^2_{it} w_{it} - \gamma_k p^2_{it}(\alpha^* e_{ti} + \beta_t) > \sum_t p^1_{it} w_{it} - \gamma_k p^1_{it}(\alpha^* e_{ti} + \beta_t)$$

while

$$\sum_t p^2_{it} w_{it} - \gamma_j p^2_{it}(\alpha^* e_{ti} + \beta_t) \leq \sum_t p^1_{it} w_{it} - \gamma_j p^1_{it}(\alpha^* e_{ti} + \beta_t).$$

Taking the difference between the last two equations results in

$$\sum_t p^2_{it}(\alpha^* e_{ti} + \beta_t) > \sum_t p^1_{it}(\alpha^* e_{ti} + \beta_t).$$

Plugging this inequality back into an earlier expression, I get $\sum_t p^2_{it} w_{it} > \sum_t p^1_{it} w_{it}$. Therefore, $p^1 \notin p^*(\beta^*, \gamma_l)$. Note also that for any small $\epsilon > 0$ there exists infinitely many n such that $\gamma_n < \epsilon$ and finitely many m such that $\gamma_m > \epsilon$. By a result from measure theory (Billingsley

(2008) p.52), therefore, $\liminf_{n\to\infty} p^*(\beta^*, \gamma_n) = \limsup_{n\to\infty} p^*(\beta^*, \gamma_n)$ and the set-theoretic limit $\lim_{n\to\infty} p^*(\beta^*, \gamma_n)$ exists, implying by the above argument that $\lim_{n\to\infty} g_{\gamma_n}(\beta^*)$ also exists.

It only remains to prove $\lim_{n\to\infty} g_{\gamma_n}(\beta) \subset g_0(\beta)$ for all $\beta \in \tilde{B}$. Suppose to the contrary there exist some b and β such that $b \in \lim_{n\to\infty} g_{\gamma_n}(\beta)$ but $b \notin g_0(\beta)$. Thus $b \in \lim_{n\to\infty} \text{argmax}_{p_i \in P}(\sum_t p_{it} w_{it} - \gamma_n p_{it}(\alpha^* e_{ti} + \beta_t)$ s.t. $\sum_t p_{it}(\alpha^* e_{ti} + \beta_t) \leq b)$ but there exists some b^* satisfying the budget constraint such that $\sum_t b^*_{it} w_{it} > \sum_t b_{it} w_{it}$. But this implies $\lim_{n\to\infty} \sum_t b^*_{it} w_{it} - \gamma_n b^*_{it}(\alpha^* e_{ti} + \beta_t) > \lim_{n\to\infty} \sum_t b_{it} w_{it} - \gamma_n b_{it}(\alpha^* e_{ti} + \beta_t)$ since $\gamma_n \to 0$. This is a contradiction to the assumption $b \in \lim_{n\to\infty} g_{\gamma_n}(\beta)$. It is thus true that $\beta^* \in \lim_{n\to\infty} g_{\gamma_n}(\beta^*) \subset g_0(\beta^*)$, implying $\beta^* \in g_0(\beta^*)$. □

Step 4. *For the fixed point $\beta^* = \lim_{\gamma_n \to 0} \beta^*_{\gamma_n}$ of $g(\cdot)$, the associated price function vector $(\pi_{te} \equiv \alpha^* e + f_t(\beta^*))_t$, where $f_t(\beta^*)$ is the t-th element of $f(\beta^*)$, satisfies the conditions in Lemma 1.*

Proof of Step 4. By the definition of a fixed point and correspondence $g(\cdot)$, there exists $z^* \equiv (z^*_t) \in z(f(\beta^*))$ such that $\beta^*_t = f_t(\beta^*) + z^*_t$ for all t. Fix any such z^* and the associated β^*. It is enough to show that the associated equilibrium treatment assignment probability vector (p^*_{it}) with $(p^*_{it})_t \in \text{argmax}_{p_i \in P}(\sum_t p_{it} w_{it}$ s.t. $\sum_t p_{it}(\alpha^* e_{ti} + f_t(\beta^*)) \leq b)$ satisfies the capacity constraint for every treatment $t = t_1, ..., t_m$. For each treatment t, there are three cases to consider:

Case 1: $\beta^*_t < 0$. Then $f_t(\beta^*) \equiv \max\{0, \min\{\beta^*_t, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}\}\} = 0$ and hence $\beta^*_t = f_t(\beta^*) + z^*_t$ implies $\beta^*_t = z^*_t \equiv \sum_i p^*_{it} - c_t < 0$, implying $\sum_i p^*_{it} < c_t$, i.e., the capacity constraint holds.

Case 2: $\beta^*_t \in [0, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}]$. By the definition of f, I have $f_t(\beta^*) = \beta^*_t$. Then $\beta^*_t = f_t(\beta^*) + z^*_t$ implies $z^*_t = 0$, i.e., the capacity constraint holds with equality.

Case 3: $\beta^*_t > nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}$. Then $f_t(\beta^*) = nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}$ and hence $\beta^*_t = f_t(\beta^*) + z^*_t$ implies that $z^*_t = \beta^*_t - nb + \alpha^* \bar{e} 1\{\bar{e} > 0\} > 0$, i.e., treatment t is in excess demand at price $\pi_{te} \equiv \alpha^* e + f_t(\beta^*)$. However, for any possible predicted effect level $e \leq \bar{e}$, I have

$$\pi_{te} \equiv \alpha^* e + f_t(\beta^*) = \alpha^* e + nb - \alpha^* \bar{e} 1\{\bar{e} > 0\} = \begin{cases} nb + \alpha^*(e - \bar{e}) \geq nb & \text{if } \bar{e} > 0 \\ nb + \alpha^* e \geq nb & \text{otherwise,} \end{cases}$$

where the last inequality is by $\alpha^* < 0$ and $e \leq \bar{e} \leq 0$. Therefore, for each subject i, $p^*_{it} \leq b/\pi_{te_{ti}} \leq 1/n$. This implies that $\sum_i p^*_{it} \leq 1 \leq c_t$, a contradiction.

Finally, the construction of β^* as $\beta^* = \lim_{\gamma_n \to 0} \beta^*_{\gamma_n}$ guarantees that every subject i's p_i solves the utility maximization problem with the minimum expenditure $\sum_t p_{it}\pi_{te_{ti}}$ This completes the proof of Step 4 and Lemma 1. □

□

I use Lemma 1 to show there exists (p^{**}_{it}) that satisfies the conditions in Definition 2. Let (p^*_{it}) be the assignment probability profile found in Lemma 1. Define $I(w,e) \equiv \{i \in \{i_1, ..., i_n\} | w_i = w, e_i = e\}$ be the set of subjects whose WTP and predicted effect vectors are w and e, respectively. For each w, e, and $i \in I(w,e)$, let

$$p^{**}_i \equiv \frac{\sum_{i \in I(w,e)} p^*_i}{|I(w,e)|}.$$

p^{**}_i solves the utility maximization problem in Step (1) of Definition 2 with the minimum expenditure since $\sum_t p^{**}_i w_{it} = \sum_t p^*_i w_{it}$ and $\sum_t p^{**}_i \pi_{te_{ti}} = \sum_t p^*_i \pi_{te_{ti}} \leq b$. The above construction guarantees that $p^{**}_i = p^{**}_j$ for any i and j with $w_i = w_j$ and $e_i = e_j$. p^{**}_i also satisfies the capacity constraints by $\sum_i p^{**}_i = \sum_i p^*_i \leq c_t$ for all t, where the last inequality is by Lemma 1. p^{**}_i thus satisfies the conditions in Definition 2.

EXAM is ex ante Pareto efficient subject to the randomization and capacity constraint: Suppose to the contrary that there exists $\epsilon \in [0, \bar{\epsilon})$ such that $p^*_{it}(\epsilon)$ is ex ante Pareto dominated by another feasible treatment assignment probabilities $(p_{it}(\epsilon))_{i,t} \in P^n$ with $p_{it}(\epsilon) \in [\epsilon, 1-\epsilon]$ for all i and t and $\sum_i p_{it} \leq c_t$ for all $t = t_1, ..., t_m,$, i.e.,

- $\sum_t p_{it}(\epsilon) e_{ti} \geq \sum_t p^*_{it}(\epsilon) e_{ti}$ for all i and
- $\sum_t p_{it}(\epsilon) w_{it} \geq \sum_t p^*_{it}(\epsilon) w_{it}$ for all i

with at least one strict inequality. Let me use $p_{it}(\epsilon)$ to define the following treatment assignment probabilities:

$$p_{it} \equiv [p_{it}(\epsilon) - q p^{RCT}_t]/(1-q),$$

where $q \equiv \inf\{q' \in [0,1] | (1-q')p^*_{it} + q' p^{RCT}_t \in [\epsilon, 1-\epsilon] \text{ for all } i \text{ and } t\}$ is the mixing weight used for defining and computing $p^*_{it}(\epsilon)$ in Definition 2. In other words, p_{it} are the treatment assignment probabilities such that the following holds:

$$p_{it}(\epsilon) = (1-q)p_{it} + q p^{RCT}_t.$$

Since both $p_{it}(\epsilon)$ and p_t^{RCT} are in convex set P^n, p_{it} is also in P^n (note that $\sum_t p_{it} = \sum_t [p_{it}(\epsilon) - qp_t^{RCT}]/(1-q)) = (1-q)/(1-q) = 1$ for every i). For each i, I have

$$\begin{aligned}\sum_t p_{it}(\epsilon)e_{ti} \geq \sum_t p_{it}^*(\epsilon)e_{ti} &\Leftrightarrow \sum_t ((1-q)p_{it} + qp_t^{RCT})e_{ti} \geq \sum_t ((1-q)p_{it}^* + qp_t^{RCT})e_{ti} \\ &\Leftrightarrow \sum_t (1-q)p_{it}e_{ti} \geq \sum_t (1-q)p_{it}^* e_{ti} \\ &\Leftrightarrow \sum_t p_{it}e_{ti} \geq \sum_t p_{it}^* e_{ti}.\end{aligned}$$

Similarly, for each i, I have

$$\begin{aligned}\sum_t p_{it}(\epsilon)w_{it} \geq \sum_t p_{it}^*(\epsilon)w_{it} &\Leftrightarrow \sum_t ((1-q)p_{it} + qp_t^{RCT})w_{it} \geq \sum_t ((1-q)p_{it}^* + qp_t^{RCT})w_{it} \\ &\Leftrightarrow \sum_t (1-q)p_{it}w_{it} \geq \sum_t (1-q)p_{it}^* w_{it} \\ &\Leftrightarrow \sum_t p_{it}w_{it} \geq \sum_t p_{it}^* w_{it}.\end{aligned}$$

Therefore, the assumption that $p_{it}(\epsilon)$ ex ante Pareto dominates $p_{it}^*(\epsilon)$ implies that p_{it} ex ante Pareto dominates p_{it}^*, i.e.,

- $\sum_t p_{it}e_{ti} \geq \sum_t p_{it}^* e_{ti}$ for all i and
- $\sum_t p_{it}w_{it} \geq \sum_t p_{it}^* w_{it}$ for all i

with at least one strict inequality. There are two cases to consider.

Case 1: $\sum_t p_{\tilde{i}t}e_{t\tilde{i}} > \sum_t p_{\tilde{i}t}^* e_{t\tilde{i}}$ for some $\tilde{i}$. This implies

$$\begin{aligned}\sum_t \sum_i p_{it}e_{ti} > \sum_t \sum_i p_{it}^* e_{ti} &\Leftrightarrow \sum_t \sum_i p_{it}(\pi_{te_{ti}} - \beta_t)/\alpha > \sum_t \sum_i p_{it}^*(\pi_{te_{ti}} - \beta_t)/\alpha \\ &\quad \text{(by the definition of } \pi_{te} \equiv \alpha e + \beta_t \text{ with } \alpha \neq 0) \\ &\Leftrightarrow \sum_t \sum_i p_{it}\pi_{te_{ti}}/\alpha > \sum_t \sum_i p_{it}^* \pi_{te_{ti}}/\alpha \\ &\quad \text{(since } \sum_i p_{it} = \sum_i p_{it}^* = c_t) \\ &\Leftrightarrow \sum_t \sum_i p_{it}\pi_{te_{ti}} < \sum_t \sum_i p_{it}^* \pi_{te_{ti}}. \\ &\quad \text{(since } \alpha < 0 \text{ by Definition 2)}\end{aligned}$$

I thus have

$$\sum_t \sum_i p_{it}\pi_{te_{ti}} < \sum_t \sum_i p^*_{it}\pi_{te_{ti}}. \tag{11}$$

However, it has also to be the case that $\sum_t p_{it}\pi_{te_{ti}} \geq \sum_t p^*_{it}\pi_{te_{ti}}$ for any i since (a) $\sum_t p_{it}w_{it} \geq \sum_t p^*_{it}w_{it}$ by assumption and (b) $(p^*_{it})_t$ is (a mixture of) the cheapest among all feasible assignment probability vectors that i most prefers under prices $(\pi_{te})_{t,e}$ and budget b. Thus $\sum_t \sum_i p_{it}\pi_{te_{ti}} \geq \sum_t \sum_i p^*_{it}\pi_{te_{ti}}$, a contradiction to inequality (11).

Case 2: $\sum_t p_{\tilde{i}t}w_{\tilde{i}t} > \sum_t p^*_{\tilde{i}t}w_{\tilde{i}t}$ for some $\tilde{i}$. Since $\tilde{i}$ most prefers $(p^*_{\tilde{i}t})_t$ among all assignment probability vectors in P^n that satisfies the budget constraint under prices $(\pi_{te})_{t,e}$, the strictly more preferred treatment assignment probability vector $(p_{\tilde{i}t})_t$ must violate the budget constraint, i.e., $\sum_t p_{\tilde{i}t}\pi_{te_{\tilde{i}i}} > b \geq \sum_t p^*_{\tilde{i}t}\pi_{te_{\tilde{i}i}}$, where the second weak inequality comes from the assumption that $(p^*_{\tilde{i}t})_t$ satisfies the budget constraint under prices (π_{te}). Moreover, for any other subject $i \neq \tilde{i}$, $\sum_t p_{it}\pi_{te_{ti}} \geq \sum_t p^*_{it}\pi_{te_{ti}}$ since $(p^*_{it})_t$ is (a mixture of) the cheapest among all assignment probability vectors in P that i most prefers under prices $(\pi_{te})_{t,e}$ and budget b. I thus have

$$\sum_t p_{\tilde{i}t}\pi_{te_{\tilde{i}i}} + \sum_{i\neq\tilde{i}} \sum_t p_{it}\pi_{te_{ti}} > \sum_t p^*_{\tilde{i}t}\pi_{te_{\tilde{i}i}} + \sum_{i\neq\tilde{i}} \sum_t p^*_{it}\pi_{te_{ti}} \Leftrightarrow \sum_i \sum_t p_{it}\pi_{te_{ti}} > \sum_i \sum_t p^*_{it}\pi_{te_{ti}}.$$

However, by the logic described in Case 1, the assumption ($\sum_t p_{it}e_{ti} \geq \sum_t p^*_{it}e_{ti}$ for all i) implies that $\sum_i \sum_t p_{it}\pi_{te_{ti}} \leq \sum_i \sum_t p^*_{it}\pi_{te_{ti}}$, a contradiction.

Therefore, $p^*_{it}(\epsilon)$ with any $\epsilon \in [0, \bar{\epsilon})$ is never ex ante Pareto dominated by another treatment assignment probabilities $(p_{it}(\epsilon))_{i,t} \in P^n$ with $p_{it}(\epsilon) \in [\epsilon, 1-\epsilon]$ for all i and t.

Proof of Proposition 3

The proof uses intermediate observations.

Lemma 2. *EXAM is "envy-free," i.e., for any experimental design problem, any $\epsilon \in [0, \bar{\epsilon})$, any subjects i and j with $e_{ti} = e_{tj}$ for all t,*

$$\sum_t p^*_{it}(\epsilon)w_{it} \geq \sum_t p^*_{jt}(\epsilon)w_{it}.$$

Proof of Lemma 2. In Definition 2, all subjects have the same budget and any subjects i and j with $e_{ti} = e_{tj}$ face the same price π_{te} of treatment t. For any subjects i and j with $e_{ti} = e_{tj}$ for all t, therefore, $(p^*_{jt})_t$ satisfies i's budget constraint and $\sum_t p^*_{it}w_{it} \geq \sum_t p^*_{jt}w_{it}$.

This implies the desired conclusion since

$$\sum_t p^*_{it} w_{it} \geq \sum_t p^*_{jt} w_{it} \Leftrightarrow (1-q)\sum_t p^*_{it} w_{it} + q\sum_t p^{RCT}_t w_{it} \geq (1-q)\sum_t p^*_{jt} w_{it} + q\sum_t p^{RCT}_t w_{it}$$
$$\Leftrightarrow \sum_t p^*_{it}(\epsilon) w_{it} \geq \sum_t p^*_{jt}(\epsilon) w_{it},$$

where the first equivalence is by $p^{RCT}_t = p^{RCT}_t \equiv c_t/n$. □

Lemma 3. *EXAM with WTP reporting is "semi-anonymous." That is, for any sequence of experimental design problems, any n with any $\epsilon^n \in [0, \bar{\epsilon}^n)$, any subjects i and j with $e_{ti} = e_{tj}$ for all t, let $(w_i, w_j, w_{-\{i,j\}})$ be a permutation of $(w_j, w_i, w_{-\{i,j\}})$ obtained by permuting i and j's WTP reports w_i and w_j. Semi-anonymity means that*

$$p^{*n}_i(w_i, w_j, w_{-\{i,j\}}; \epsilon^n) = p^{*n}_j(w_j, w_i, w_{-\{i,j\}}; \epsilon^n),$$
$$p^{*n}_j(w_i, w_j, w_{-\{i,j\}}; \epsilon^n) = p^{*n}_i(w_j, w_i, w_{-\{i,j\}}; \epsilon^n), \text{ and}$$
$$p^{*n}_k(w_i, w_j, w_{-\{i,j\}}; \epsilon^n) = p^{*n}_k(w_j, w_i, w_{-\{i,j\}}; \epsilon^n), \text{ for all } k \neq i, j$$

Proof of Lemma 3. In Definition 2 of EXAM, all subjects have the same budget and any subjects i and j with $e_{ti} = e_{tj}$ face the same price π_{te} of treatment t. For any subjects i and j with $e_{ti} = e_{tj}$ for all t, therefore, given any $w_{-\{i,j\}}$, subject i with WTP report w_j solves the same constrained utility maximization problem as subject j with WTP report w_j does. Therefore, $p^{*n}_i(w_i, w_j, w_{-\{i,j\}}; 0) = p^{*n}_j(w_j, w_i, w_{-\{i,j\}}; 0)$ and $p^{*n}_j(w_i, w_j, w_{-\{i,j\}}; 0) = p^{*n}_i(w_j, w_i, w_{-\{i,j\}}; 0)$. This implies semi-anonymity since

$$\begin{aligned} p^{*n}_i(w_i, w_j, w_{-\{i,j\}}; \epsilon^n) &\equiv (1-q^n) p^{*n}_i(w_i, w_j, w_{-\{i,j\}}; 0) + q^n p^{RCTn}_i \\ &= (1-q^n) p^{*n}_j(w_j, w_i, w_{-\{i,j\}}; 0) + q^n p^{RCTn}_j \\ &\equiv p^{*n}_j(w_j, w_i, w_{-\{i,j\}}; \epsilon^n), \end{aligned}$$

where q^n is the mixing probability q for the n-th problem in the sequence of experimental design problems while $p^{RCTn}_i = p^{RCTn}_j \equiv c^n_t/n$. The last line follows from the fact that the switch of w_i and w_j have no effect on the utility maximization problem for any other $k \neq i, j$. □

Lemmas 2 and 3 imply Proposition 3 by using Theorem 1 of Azevedo and Budish (2017) (precisely, a generalization of their Theorem 1 in their Supplementary Appendix B).

A Statistical Lemma and Its Proof

Lemma 4. *Assume a sample of m subjects is uniformly randomly drawn (i.e., every combination of m subjects occurs with equal probability) from the fixed finite population of n subjects with a fixed vector of a variable $(X_1, ..., X_n)$. Denote the random sample by I. Let $\mu \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$, $\sigma^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$, $\hat{\mu} \equiv \frac{1}{m}\sum_{i\in I} X_i$ and $\hat{\sigma}^2 \equiv \frac{1}{m-1}\sum_{i\in I}(X_i - \hat{\mu})^2$. Then,*

$$V(\hat{\mu}) = \frac{n-m}{nm}\sigma^2 \text{ and } E(\hat{\sigma}^2) = \sigma^2.$$

Proof of Lemma 4. Let $W_i = 1\{i \in I\}$ so that $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{n} X_i W_i$ and $\hat{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 W_i$. Then $E(W_i) = E(W_i^2) = \frac{m}{n}$ for all i, implying $V(W_i) = \frac{m}{n} - (\frac{m}{n})^2 = \frac{m(n-m)}{n^2}$ for all i. Since $E(W_i W_j) = \frac{m(m-1)}{n(n-1)}$ for any $i \neq j$, it is the case that for any $i \neq j$,

$$\begin{aligned} Cov(W_i, W_j) &= E[(W_i - \frac{m}{n})(W_j - \frac{m}{n})] \\ &= E(W_i W_j) - \frac{m}{n}E(W_j) - \frac{m}{n}E(W_i) + (\frac{m}{n})^2 \\ &= \frac{m(m-1)}{n(n-1)} - (\frac{m}{n})^2 \\ &= -\frac{m(n-m)}{n^2(n-1)}. \end{aligned} \tag{12}$$

It follows that

$$\begin{aligned} V(\hat{\mu}) &= \frac{1}{m^2}V(\sum_{i=1}^{n} X_i W_i) \\ &= \frac{1}{m^2}(\sum_{i=1}^{n} X_i^2 V(W_i) + \sum_{i=1}^{n}\sum_{j\neq i} X_i X_j Cov(W_i, W_j)) \\ &= \frac{1}{m^2}(\frac{m(n-m)}{n^2}\sum_{i=1}^{n} X_i^2 - \frac{m(n-m)}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} X_i X_j) \\ &= \frac{n-m}{n^2 m}(\sum_{i=1}^{n} X_i^2 - \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} X_i X_j) \\ &= \frac{n-m}{n^2 m}(\sum_{i=1}^{n} X_i^2 - \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j + \frac{1}{n-1}\sum_{i=1}^{n} X_i^2) \\ &= \frac{n-m}{n^2 m}(\frac{n}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j) \end{aligned}$$

$$
\begin{aligned}
&= \frac{n-m}{nm(n-1)}(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j) \\
&= \frac{n-m}{nm(n-1)}(\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j + \frac{1}{n}(\sum_{i=1}^{n} X_i)^2) \\
&= \frac{n-m}{nm(n-1)}\sum_{i=1}^{n}(X_i^2 - \frac{2}{n} X_i \sum_{j=1}^{n} X_j + \frac{1}{n^2}(\sum_{j=1}^{n} X_j)^2) \\
&= \frac{n-m}{nm(n-1)}\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j)^2 \\
&= \frac{n-m}{nm}\sigma^2.
\end{aligned}
$$

For the other part,

$$
\begin{aligned}
E(\hat{\sigma}^2) &= \frac{1}{m-1}E(\sum_{i=1}^{n}(X_i - \hat{\mu})^2 W_i) \\
&= \frac{1}{m-1}E(\sum_{i=1}^{n}(X_i^2 W_i - 2X_i W_i \hat{\mu} + \hat{\mu}^2 W_i)) \\
&= \frac{1}{m-1}E(\sum_{i=1}^{n} X_i^2 W_i - 2m\hat{\mu}^2 + m\hat{\mu}^2) \\
&= \frac{1}{m-1}E(\sum_{i=1}^{n} X_i^2 W_i - m\hat{\mu}^2) \\
&= \frac{1}{m-1}(\frac{m}{n}\sum_{i=1}^{n} X_i^2 - m(V(\hat{\mu}) + [E(\hat{\mu})]^2)) \\
&= \frac{1}{m-1}(\frac{m}{n}\sum_{i=1}^{n} X_i^2 - \frac{n-m}{n}\sigma^2 - m\mu^2) \\
&= \frac{1}{m-1}(\frac{m(n-1)}{n}\sigma^2 - \frac{n-m}{n}\sigma^2) \\
&= \sigma^2,
\end{aligned}
$$

where the third last equality is by $E(\hat{\mu}) = \mu$ while the second last equality is by the definition of σ^2. □

Proof of Proposition 4

The proof uses the following lemma.

Lemma 5. *There exists estimator $\hat{\theta}_{EXAM,t}$ such that $E(\hat{\theta}_{EXAM,t}|p^*(\epsilon)) = E(\hat{\mu}_{RCT}(t)^2|p^{RCT})$*

where $\hat{\mu}_{RCT}(t) \equiv \dfrac{\sum_i D_{it} Y_i}{c_t}$ *with* D_{it} *being the treatment assignment indicator under RCT.*

Proof of Lemma 5. Let $\mu_t \equiv \dfrac{1}{n}\sum_{i=1}^n Y_i(t)$ and $S_t^2 \equiv \dfrac{1}{n-1}\sum_{i=1}^n (Y_i(t)-\mu_t)^2$. I have

$$\begin{aligned} E(\hat{\mu}_{RCT}(t)^2|p^{RCT}) &= Var(\hat{\mu}_{RCT}(t)|p^{RCT}) + E(\hat{\mu}_{RCT}(t)|p^{RCT})^2 \\ &= \frac{n-c_t}{nc_t} S_t^2 + \mu_t^2 \\ &= \frac{n-c_t}{nc_t}\frac{1}{n-1}(\sum_{i=1}^n Y_i(t)^2 - n\mu_t^2) + \mu_t^2 \\ &= \frac{n-c_t}{(n-1)c_t}\frac{1}{n}\sum_{i=1}^n Y_i(t)^2 + \frac{n(c_t-1)}{(n-1)c_t}\mu_t^2, \end{aligned} \tag{13}$$

where the second equality holds by the first part of Lemma 4 and the fact that $E(\hat{\mu}_{RCT}(t)|p^{RCT}) = \mu_t$.

Below I construct unbiased estimators with EXAM's data for each term in the right-hand side of equation (13), that is, $\dfrac{1}{n}\sum_{i=1}^n Y_i(t)^2$ and μ_t^2. I then combine these estimators into developing an unbiased estimator for $E(\hat{\mu}_{RCT}(t)^2|p^{RCT})$ (which I interpret as a parameter) with EXAM. Under EXAM $p^*(\epsilon)$, $\hat{\theta}_{1t} = \dfrac{1}{n}\sum_p \dfrac{1}{p_t}\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it}$ unbiasedly estimates $\dfrac{1}{n}\sum_{i=1}^n Y_i(t)^2$ because

$$\begin{aligned} E(\hat{\theta}_{1t}|p^*(\epsilon)) &= \frac{1}{n}\sum_p \frac{1}{p_t}\sum_{i:p_i^*(\epsilon)=p} Y_i(t)^2 E(D_{it}|p^*(\epsilon)) \\ &= \frac{1}{n}\sum_p \frac{1}{p_t}\sum_{i:p_i^*(\epsilon)=p} Y_i(t)^2 p_t \\ &= \frac{1}{n}\sum_p \sum_{i:p_i^*(\epsilon)=p} Y_i(t)^2 \\ &= \frac{1}{n}\sum_{i=1}^n Y_i(t)^2. \end{aligned}$$

Next I obtain an unbiased estimator for μ_t^2 under EXAM $p^*(\epsilon)$. Recall $n_p \equiv \sum_{i=1}^n 1\{p_i^*(\epsilon) = p\}$ is the number of subjects with assignment probability vector p and $N_{pt} \equiv \sum_{i:p_i^*(\epsilon)=p} D_{it}$ is a random variable that stands for the number of subjects with assignment probability vector p and assigned to treatment t. Let $\mu_{pt} \equiv \dfrac{1}{n_p}\sum_{i:p_i^*(\epsilon)=p} Y_i(t)$, $\hat{\mu}_{EXAM,pt} \equiv \dfrac{1}{p_t n_p}\sum_{i:p_i^*(\epsilon)=p} Y_i D_{it}$ and $\hat{\mu}_{EXAM,t} \equiv \sum_p \dfrac{n_p}{n}\dfrac{p_t n_p}{N_{pt}}\hat{\mu}_{EXAM,pt}$. Note that $\mu_t = \sum_p \dfrac{n_p}{n}\mu_{pt}$. By Definition 2 (3), conditional on (N_{pt}), every deterministic treatment assignment consistent with (N_{pt}) happens

with equal probability. This implies that

$$E(D_{it}|(N_{pt}) = (n_{pt}), p_i^*(\epsilon) = p) = \frac{n_{pt}}{n_p}.$$

In addition, conditional on $(N_{pt}) = (n_{pt})$, the set $\{i : p_i^*(\epsilon) = p, D_{it} = 1\}$ can be regarded as a random sample of n_{pt} subjects from the subpopulation of n_p subjects with propensity vector p. Applying the first part of Lemma 4, I have

$$Var(\frac{1}{N_{pt}} \sum_{i:p_i^*(\epsilon)=p} Y_i D_{it}|(N_{pt}) = (n_{pt}), p^*(\epsilon)) = \frac{n_p - n_{pt}}{n_p n_{pt}} S_{pt}^2,$$

where $S_{pt}^2 = \frac{1}{n_p - 1} \sum_{i:p_i^*(\epsilon)=p} (Y_i(t) - \mu_{pt})^2$. It follows from the above expressions that

$$\begin{aligned} E(\hat{\mu}_{EXAM,pt}|(N_{pt}) = (n_{pt}), p^*(\epsilon)) &= \frac{1}{p_t n_p} \sum_{i:p_i^*(\epsilon)=p} Y_i E(D_{it}|(N_{pt}) = (n_{pt}), p^*(\epsilon)) \\ &= \frac{1}{p_t n_p} \sum_{i:p_i^*(\epsilon)=p} Y_i \frac{n_{pt}}{n_p} = \frac{n_{pt}}{p_t n_p} \mu_{pt}, \end{aligned} \tag{14}$$

and

$$\begin{aligned} Var(\hat{\mu}_{EXAM,pt}|(N_{pt}) = (n_{pt}), p^*(\epsilon)) &= \Big(\frac{n_{pt}}{p_t n_p}\Big)^2 Var(\frac{1}{N_{pt}} \sum_{i:p_i^*(\epsilon)=p} Y_i D_{it}|(N_{pt}) = (n_{pt}), p^*(\epsilon)) \\ &= \Big(\frac{n_{pt}}{p_t n_p}\Big)^2 \frac{n_p - n_{pt}}{n_p n_{pt}} S_{pt}^2. \end{aligned} \tag{15}$$

Definition 2 (3) also implies that treatment assignments are independent across subpopulations with different propensities conditional on (N_{pt}). Hence, $\hat{\mu}_{EXAM,pt}$ is independent across p conditional on (N_{pt}). I have

$$\begin{aligned} E(\hat{\mu}_{EXAM,t}^2|(N_{pt}), p^*(\epsilon)) &= E((\sum_p \frac{n_p}{n} \frac{p_t n_p}{N_{pt}} \hat{\mu}_{EXAM,pt})^2|(N_{pt}) = (n_{pt}), p^*(\epsilon)) \\ &= \sum_p \sum_{p' \neq p} \frac{n_p n_{p'}}{n^2} \frac{p_t n_p}{n_{pt}} E(\hat{\mu}_{EXAM,pt}|(N_{pt}), p^*(\epsilon)) \frac{p_t' n_{p'}}{n_{p't}} E(\hat{\mu}_{EXAM,p't}|(N_{pt}), p^*(\epsilon)) \\ &\quad + \sum_p \frac{n_p^2}{n^2} \Big(\frac{p_t n_p}{n_{pt}}\Big)^2 E(\hat{\mu}_{EXAM,pt}^2|(N_{pt}) = (n_{pt}), p^*(\epsilon)) \\ &= \sum_p \sum_{p' \neq p} \frac{n_p n_{p'}}{n^2} \mu_{pt} \mu_{p't} + \sum_p \frac{n_p^2}{n^2} \Big(\frac{p_t n_p}{n_{pt}}\Big)^2 (Var(\hat{\mu}_{EXAM,pt}|(N_{pt}) = (n_{pt}), p^*(\epsilon)) \end{aligned}$$

$$
\begin{aligned}
&+ E(\hat{\mu}_{EXAM,pt}|(N_{pt}) = (n_{pt}), p^*(\epsilon))^2) \\
&= \sum_p \sum_{p' \neq p} \frac{n_p n_{p'}}{n^2} \mu_{pt} \mu_{p't} + \sum_p \frac{n_p^2}{n^2} \Big(\frac{p_t n_p}{n_{pt}} \Big)^2 \Big\{ \Big(\frac{n_{pt}}{p_t n_p} \Big)^2 \frac{n_p - n_{pt}}{n_p n_{pt}} S_{pt}^2 + \Big(\frac{n_{pt}}{p_t n_p} \Big)^2 \mu_{pt}^2 \Big\} \\
&= \sum_p \sum_{p' \neq p} \frac{n_p n_{p'}}{n^2} \mu_{pt} \mu_{p't} + \sum_p \frac{n_p^2}{n^2} (\frac{n_p - n_{pt}}{n_p n_{pt}} S_{pt}^2 + \mu_{pt}^2) \\
&= (\sum_p \frac{n_p}{n} \mu_{pt})^2 + \sum_p \frac{n_p^2}{n^2} \frac{n_p - n_{pt}}{n_p n_{pt}} S_{pt}^2 \\
&= \mu_t^2 + \sum_p \frac{n_p^2}{n^2} \frac{n_p - n_{pt}}{n_p n_{pt}} S_{pt}^2,
\end{aligned}
$$

where I use the independence of $\hat{\mu}_{EXAM,pt}$ across p conditional on (N_{pt}) for the second equality, equation (14) for the third and the fourth equalities and equation (15) for the fourth equality. By the law of iterated expectations,

$$
\begin{aligned}
E(\hat{\mu}_{EXAM,t}^2 | p^*(\epsilon)) &= E(E(\hat{\mu}_{EXAM,t}^2 | (N_{pt}), p^*(\epsilon)) | p^*(\epsilon)) \\
&= \mu_t^2 + \sum_p \frac{n_p^2}{n^2} (E(\frac{1}{N_{pt}} | p^*(\epsilon)) - \frac{1}{n_p}) S_{pt}^2 \\
&= \mu_t^2 + \sum_p \frac{n_p^2}{n^2} (\frac{1}{\underline{n}_{pt}} (1 - p_t n_p + \underline{n}_{pt}) + \frac{1}{\underline{n}_{pt} + 1} (p_t n_p - \underline{n}_{pt}) - \frac{1}{n_p}) S_{pt}^2 \\
&= \mu_t^2 + \sum_p \frac{n_p^2}{n^2} (\frac{1 - p_t n_p + 2\underline{n}_{pt}}{\underline{n}_{pt}(\underline{n}_{pt} + 1)} - \frac{1}{n_p}) S_{pt}^2, \qquad (16)
\end{aligned}
$$

where the third equality holds because Definition 2 (3) implies

$$
\Pr(N_{pt} = n_{pt} | p^*(\epsilon)) = \begin{cases} 1 - p_t n_p + \underline{n}_{pt} & \text{if } n_{pt} = \underline{n}_{pt} \\ p_t n_p - \underline{n}_{pt} & \text{if } n_{pt} = \underline{n}_{pt} + 1 \\ 0 & \text{otherwise.} \end{cases}
$$

Now consider an unbiased estimator for S_{pt}^2. Let $\hat{S}_{pt}^2 \equiv \frac{1}{p_t n_p - 1} \sum_{i: p_i^*(\epsilon) = p} (Y_i - \frac{p_t n_p}{N_{pt}} \hat{\mu}_{EXAM,pt})^2 D_{it}$. This $\hat{S}_{pt}^2$ is unbiased for S_{pt}^2, because

$$
\begin{aligned}
E(\hat{S}_{pt}^2 | p^*(\epsilon)) &= E(E\{\hat{S}_{pt}^2 | (N_{pt}), p^*(\epsilon)\} | p^*(\epsilon)) \\
&= E\big[\frac{N_{pt} - 1}{p_t n_p - 1} E\big\{ \frac{1}{N_{pt} - 1} \sum_{i: p_i^*(\epsilon) = p} (Y_i - \frac{1}{N_{pt}} \sum_{j: p_j^*(\epsilon) = p} Y_j D_{jt})^2 D_{it} | (N_{pt}), p^*(\epsilon) \big\} | p^*(\epsilon) \big]
\end{aligned}
$$

$$
\begin{aligned}
&= E(\frac{N_{pt}-1}{p_t n_p - 1} S_{pt}^2 | p^*(\epsilon)) \\
&= \frac{p_t n_p - 1}{p_t n_p - 1} S_{pt}^2 \\
&= S_{pt}^2,
\end{aligned}
$$

where the first equality holds by the law of expected iterations, I use the second part of Lemma 4 for the third equality and the fact that $E(N_{pt}|p^*(\epsilon)) = p_t n_p$ for the fourth equality. Combining this with equation (16), I obtain an unbiased estimator for μ_t^2: $\hat{\theta}_{2t} = \hat{\mu}_{EXAM,t}^2 - \sum_p \frac{n_p^2}{n^2}(\frac{1 - p_t n_p + 2\underline{n}_{pt}}{\underline{n}_{pt}(\underline{n}_{pt}+1)} - \frac{1}{n_p})\hat{S}_{pt}^2$. By equation (13), the following is an unbiased estimator for $E(\hat{\mu}_{RCT}(t)^2 | p^{RCT})$:

$$
\hat{\theta}_{EXAM,t} = \frac{n - c_t}{(n-1)c_t}\hat{\theta}_{1t} + \frac{n(c_t - 1)}{(n-1)c_t}\hat{\theta}_{2t}.
$$

□

Let D_i be the set of all feasible deterministic treatment assignments for subject i, i.e.,

$$
D_i \equiv \{d_i \equiv (d_{it})_t \in \{0,1\}^{m+1} | \sum_t d_{it} = 1\}.
$$

Let D_i^{EXAM} and D_i^{RCT} be the sets of deterministic treatment assignments that happen with a positive probability under EXAM and RCT, respectively. That is, $D_i^{EXAM} \equiv \{d_i \in D_i | \Pr(d_i|p^*(\epsilon)) > 0\}$ and $D_i^{RCT} \equiv \{d_i \in D_i | \Pr(d_i|p^{RCT}) > 0\}$, where $\Pr(d_i|(p_{it}))$ is the probability that d_i occurs under experimental design (p_{it}). With Definition 2, $D_{it} = 1$ holds with a positive probability for every t and i both under EXAM and RCT, implying

$$
D_i^{RCT} = D_i^{EXAM} = D_i. \tag{17}
$$

With the support equivalence property (17), I am ready to show the proposition. Recall that given any experimental design (p_{it}), I say an estimator $\hat{\theta}(Y, D)$ is *simple* if $\hat{\theta}(Y, D)$ can be written as

$$
\hat{\theta}(Y, D) = \sum_i f(Y_i, D_i, p_i) + \sum_t \sum_p \sum_{p'} g_{tpp'} \hat{\mu}_p(t) \hat{\mu}_{p'}(t)
$$

for some function f, weights $g_{tpp'}$'s, and $\hat{\mu}_p(t) = \frac{\sum_{i:p_i=p} D_{it} Y_i}{p_t \sum_{i=1}^n 1\{p_i = p\}}$. Suppose that parameter θ is unbiasedly estimable with RCT p^{RCT} and a simple estimator $\hat{\theta}^{RCT}(Y, D) =$

$\sum_i f(Y_i, D_i, p_i^{RCT}) + \sum_t g_t \hat{\mu}^2_{RCT}(t)$:

$$E(\hat{\theta}^{RCT}(Y, D)|p^{RCT}) = \theta. \tag{18}$$

Note that g_t is constant since for RCT, the only potential randomness in g_t comes from $(\sum_i D_{it})_t$, which is the same as the constant pseudo-capacity vector $(c_t)_t$. Now consider another estimator for EXAM:

$$\hat{\theta}^{EXAM}(Y, D) \equiv \sum_i \frac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i, p_i^{RCT}) + \sum_t g_t \hat{\theta}_{EXAM,t}.$$

With the knowledge of the original estimator $\hat{\theta}^{RCT}(Y, D)$, it is possible to compute $\hat{\theta}^{EXAM}(Y, D)$ since $\Pr(D_i|p^{RCT})$ and $\Pr(D_i|p^*(\epsilon))$ are known to the experimenter. This $\hat{\theta}^{EXAM}(Y, D)$ is unbiased for θ under EXAM:

$$\begin{aligned}
&E(\hat{\theta}^{EXAM}(Y, D)|p^*(\epsilon)) \\
&= E(\sum_i \frac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i, p_i^{RCT}) + \sum_t g_t \hat{\theta}_{EXAM,t}|p^*(\epsilon)) \\
&= \sum_i \sum_{d_i \in D_i^{EXAM}} \Pr(d_i|p^*(\epsilon)) \frac{\Pr(d_i|p^{RCT})}{\Pr(d_i|p^*(\epsilon))} f(Y_i(d_i), d_i, p_i^{RCT}) + \sum_t g_t E(\hat{\mu}^2_{RCT}(t)|p^{RCT}) \\
&= \sum_i \sum_{d_i \in D_i^{RCT}} \Pr(d_i|p^{RCT}) f(Y_i(d_i), d_i, p_i^{RCT}) + \sum_t g_t E(\hat{\mu}^2_{RCT}(t)|p^{RCT}) \\
&= E(\hat{\theta}^{RCT}(Y, D)|p^{RCT}) \\
&= \theta,
\end{aligned}$$

where $Y_i(d_i) \equiv \sum_t d_{it} Y_i(t)$ is the value of observed outcome Y_i when $D_i = d_i$, the second equality is by Lemma 5, the third equality is by the support equivalence property (17), and the last equality is by the unbiasedness assumption (18). This means that $\hat{\theta}^{EXAM}(Y, D)$ is an unbiased estimator for θ under EXAM $p^*(\epsilon)$. To complete the proof of Proposition 4, it only remains to show $\hat{\theta}^{EXAM}(Y, D)$ is a simple estimator under EXAM.

Lemma 6. *$\hat{\theta}^{EXAM}(Y, D)$ is a simple estimator under EXAM $p^*(\epsilon)$.*

Proof of Lemma 6. First note that

$$\hat{S}^2_{pt} = \frac{1}{p_t n_p - 1} \Big(\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it} - 2 \sum_{i:p_i^*(\epsilon)=p} Y_i D_{it} \frac{p_t n_p}{N_{pt}} \hat{\mu}_{EXAM,pt} + \sum_{i:p_i^*(\epsilon)=p} \big(\frac{p_t n_p}{N_{pt}}\big)^2 \hat{\mu}^2_{EXAM,pt} D_{it} \Big)$$

$$= \frac{1}{p_t n_p - 1}\Big(\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it} - \frac{(p_t n_p)^2}{N_{pt}}\hat{\mu}^2_{EXAM,pt}\Big).$$

I therefore have

$$\begin{aligned}
&\hat{\theta}_{EXAM,t} \\
&\equiv \frac{n-c_t}{(n-1)c_t}\hat{\theta}_{1t} + \frac{n(c_t-1)}{(n-1)c_t}\hat{\theta}_{2t} \\
&= \frac{n-c_t}{(n-1)c_t}\frac{1}{n}\textstyle\sum_p \frac{1}{p_t}\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it} + \frac{n(c_t-1)}{(n-1)c_t}\Big(\hat{\mu}^2_{EXAM,t} - \sum_p \frac{n_p^2}{n^2}\Big(\frac{1-p_t n_p + 2\underline{n}_{pt}}{\underline{n}_{pt}(\underline{n}_{pt}+1)} - \frac{1}{n_p}\Big)\hat{S}^2_{pt}\Big) \\
&= \sum_p \sum_{i:p_i^*(\epsilon)=p} \frac{n-c_t}{(n-1)c_t n p_t} Y_i^2 D_{it} + \frac{n(c_t-1)}{(n-1)c_t}\Big\{\Big(\sum_p \frac{n_p}{n}\frac{p_t n_p}{N_{pt}}\hat{\mu}_{EXAM,pt}\Big)^2 \\
&\quad - \sum_p \frac{n_p^2}{n^2}\Big(\frac{1-p_t n_p + 2\underline{n}_{pt}}{\underline{n}_{pt}(\underline{n}_{pt}+1)} - \frac{1}{n_p}\Big)\frac{1}{p_t n_p - 1}\Big(\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it} - \frac{(p_t n_p)^2}{N_{pt}}\hat{\mu}^2_{EXAM,pt}\Big)\Big\} \\
&= \sum_p \sum_{i:p_i^*(\epsilon)=p}\Big(\frac{n-c_t}{(n-1)c_t n p_t} - \frac{(c_t-1)n_p^2}{(n-1)c_t n(p_t n_p - 1)}\Big(\frac{1-p_t n_p + 2\underline{n}_{pt}}{\underline{n}_{pt}(\underline{n}_{pt}+1)} - \frac{1}{n_p}\Big)\Big) Y_i^2 D_{it} \\
&\quad + \frac{n(c_t-1)}{(n-1)c_t}\Big\{\sum_p \sum_{p'} \frac{n_p n_{p'}}{n^2}\frac{p_t n_p}{N_{pt}}\frac{p_t' n_{p'}}{N_{p't}}\hat{\mu}_{EXAM,pt}\hat{\mu}_{EXAM,p't} \\
&\quad + \sum_p \frac{n_p^2}{n^2}\Big(\frac{1-p_t n_p + 2\underline{n}_{pt}}{\underline{n}_{pt}(\underline{n}_{pt}+1)} - \frac{1}{n_p}\Big)\frac{1}{p_t n_p - 1}\frac{(p_t n_p)^2}{N_{pt}}\hat{\mu}^2_{EXAM,pt}\Big\} \\
&= \sum_p \sum_{i:p_i^*(\epsilon)=p} a_{1pt} Y_i^2 D_{it} + \sum_p \sum_{p'\neq p} \frac{(c_t-1)n_p^2 n_{p'}^2 p_t p_t'}{(n-1)c_t n N_{pt} N_{p't}}\hat{\mu}_{EXAM,pt}\hat{\mu}_{EXAM,p't} \\
&\quad + \sum_p \frac{(c_t-1)n_p^4 p_t^2}{(n-1)c_t n N_{pt}}\Big(\frac{1}{N_{pt}} + \Big(\frac{1-p_t n_p + 2\underline{n}_{pt}}{\underline{n}_{pt}(\underline{n}_{pt}+1)} - \frac{1}{n_p}\Big)\frac{1}{p_t n_p - 1}\Big)\hat{\mu}^2_{EXAM,pt} \\
&= \sum_i a_{1p_i^*(\epsilon)t} Y_i^2 D_{it} + \sum_p \sum_{p'} a_{2pp't}\hat{\mu}_{EXAM,pt}\hat{\mu}_{EXAM,p't},
\end{aligned}$$

where $\underline{n}_{pt}$ is the greatest integer less than or equal to $p_t n_p$ and

$$\begin{aligned}
a_{1pt} &= \frac{n-c_t}{(n-1)c_t n p_t} - \frac{(c_t-1)n_p^2}{(n-1)c_t n(p_t n_p - 1)}\Big(\frac{1-p_t n_p + 2\underline{n}_{pt}}{\underline{n}_{pt}(\underline{n}_{pt}+1)} - \frac{1}{n_p}\Big) \\
a_{2pp't} &= \begin{cases} \dfrac{(c_t-1)n_p^2 n_{p'}^2 p_t p_t'}{(n-1)c_t n N_{pt} N_{p't}} & \text{if } p \neq p' \\ \dfrac{(c_t-1)n_p^4 p_t^2}{(n-1)c_t n N_{pt}}\Big(\dfrac{1}{N_{pt}} + \Big(\dfrac{1-p_t n_p + 2\underline{n}_{pt}}{\underline{n}_{pt}(\underline{n}_{pt}+1)} - \dfrac{1}{n_p}\Big)\dfrac{1}{p_t n_p - 1}\Big) & \text{if } p = p'. \end{cases}
\end{aligned}$$

It follows that

$$
\begin{aligned}
&\hat{\theta}^{EXAM}(Y,D) \\
&\equiv \textstyle\sum_i \frac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i, p_i^{RCT}) + \sum_t g_t \hat{\theta}_{EXAM,t} \\
&= \textstyle\sum_i [\frac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i, p_i^{RCT}) + \sum_t g_t a_{1p_i^*(\epsilon)t} Y_i^2 D_{it}] + \sum_t g_t \sum_p \sum_{p'} a_{2pp't} \hat{\mu}_{EXAM,pt} \hat{\mu}_{EXAM,p't} \\
&= \textstyle\sum_i f^*(Y_i, D_i, p_i) + \sum_t \sum_p \sum_{p'} g_{tpp'} \hat{\mu}_{EXAM,pt} \hat{\mu}_{EXAM,p't},
\end{aligned}
$$

where $f^*(Y_i, D_i, p_i) = \frac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i, p_i^{RCT}) + \sum_t g_t a_{1p_it} Y_i^2 D_{it}$ and $g_{tpp'} = g_t a_{2pp't}$. Therefore, $\hat{\theta}^{EXAM}(Y,D)$ is a simple estimator under EXAM $p^*(\epsilon)$. □

Proof of Corollary 1

The mean of potential outcomes for treatment t is unbiasedly estimable with RCT and a simple estimator by the following reason. Let $\hat{\theta}(Y,D) = \sum_{i=1}^n \frac{D_{it} Y_i}{c_t}$. $\hat{\theta}(Y,D)$ is a simple unbiased estimator by $E(\hat{\theta}(Y,D)|p^{RCT}) = \sum_{i=1}^n \frac{p_t^{RCT} Y_i(t)}{c_t} = \frac{1}{n} \sum_{i=1}^n Y_i(t)$.

ATE of treatment t over control t_0 is also unbiasedly estimable with RCT and a simple estimator. To see this, let $\hat{\theta}(Y,D) = \sum_{i=1}^n (\frac{D_{it} Y_i}{c_t} - \frac{D_{it_0} Y_i}{c_{t_0}})$. This is a simple estimator, and it follows from the above argument for the mean potential outcome that $E(\hat{\theta}(Y,D)|p^{RCT}) = \frac{1}{n} \sum_{i=1}^n Y_i(t) - \frac{1}{n} \sum_{i=1}^n Y_i(t_0) = \frac{\sum_{i=1}^n (Y_i(t) - Y_i(t_0))}{n}$.

The variance of potential outcomes for treatment t is unbiasedly estimable with RCT and a simple estimator. Consider two possible definitions of the variance of potential outcomes: $S_t^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i(t) - \frac{1}{n} \sum_{j=1}^n Y_j(t))^2$ and $\sum_t^2 = \frac{1}{n} \sum_{i=1}^n (Y_i(t) - \frac{1}{n} \sum_{j=1}^n Y_j(t))^2$. To see that S_t^2 is unbiasedly estimable with RCT and a simple estimator, let $\hat{\theta}_1(Y,D) = \frac{1}{c_t - 1} \sum_{i=1}^n D_{it} (Y_i - \hat{\mu}_{RCT}(t))^2$, where $\hat{\mu}(t) = \frac{1}{c_t} \sum_{i=1}^n D_{it} Y_i$. Since I can write $\hat{\theta}_1(Y,D)$ as $\hat{\theta}_1(Y,D) = \sum_{i=1}^n \frac{D_{it} Y_i^2}{c_t - 1} - \frac{c_t}{c_t - 1} \hat{\mu}^2(t)$, this is a simple estimator with

$$
f(Y_i, D_i, p_i) = \frac{D_{it} Y_i^2}{c_t - 1}, g_t = -\frac{c_t}{c_t - 1} \text{ and } g_{t'} = 0 \text{ for all } t' \neq t.
$$

For RCT, $\{i : D_{it} = 1\}$ can be seen as a random sample of c_t subjects from the population of n subjects. Using Lemma 4, I obtain $E(\hat{\theta}_1(Y,D)|p^{RCT}) = S_t^2$.

To see that $\sum_t^2$ is also unbiasedly estimable with RCT and a simple estimator, let $\hat{\theta}_2(Y,D) = \frac{n-1}{n} \hat{\theta}_1(Y,D)$. Since I can write $\hat{\theta}_2(Y,D)$ as $\hat{\theta}_2(Y,D) = \sum_{i=1}^n \frac{n-1}{n} \frac{D_{it} Y_i^2}{c_t - 1} -$

$\frac{n-1}{n}\frac{c_t}{c_t-1}\hat{\mu}^2(t)$, this is a simple estimator with

$$f(Y_i, D_i, p_i) = \frac{n-1}{n}\frac{D_{it}Y_i^2}{c_t-1}, g_t = -\frac{n-1}{n}\frac{c_t}{c_t-1} \text{ and } g_{t'} = 0 \text{ for all } t' \neq t.$$

It follows that $E(\hat{\theta}_2(Y,D)|p^{RCT}) = \frac{n-1}{n}E(\hat{\theta}_1(Y,D)|p^{RCT}) = \frac{n-1}{n}S_t^2 = \sum_t^2$.

Finally, I consider the unbiased estimability of the average treatment effect on the treated (ATT) with RCT and EXAM. I first define ATT of t over t_0 conditional on the treatment assignment being d as

$$ATT(t|d) \equiv \frac{\sum_{i=1}^n (Y_i(t) - Y_i(t_0))d_{it}}{\sum_{i=1}^n d_{it}}.$$

I define ATT of t over t_0 for experimental design (p_{it}) as

$$ATT(t|(p_{it})) \equiv E(ATT(t|D)|(p_{it})) = E\Big(\frac{\sum_{i=1}^n (Y_i(t) - Y_i(t_0))D_{it}}{\sum_{i=1}^n D_{it}}|(p_{it})\Big).$$

For RCT,

$$ATT(t|p^{RCT}) = \frac{1}{c_t}\sum_{i=1}^n (Y_i(t) - Y_i(t_0))p_t^{RCT} = \frac{1}{n}\sum_{i=1}^n (Y_i(t) - Y_i(t_0)) = ATE.$$

Since ATE is unbiasedly estimable with RCT and a simple estimator, ATT is also unbiasedly estimable with RCT. For EXAM,

$$\begin{aligned}
ATT(t|p^*(\epsilon)) &= \frac{1}{\sum_i p_{it}^*(\epsilon)}\sum_{i=1}^n (Y_i(t) - Y_i(t_0))p_{it}^*(\epsilon) \\
&= \frac{1}{\sum_i p_{it}^*(\epsilon)}\sum_p \sum_{i:p_i^*(\epsilon)=p} (Y_i(t) - Y_i(t_0))p_t \\
&= \sum_p \frac{p_t n_p}{\sum_i p_{it}^*(\epsilon)}\frac{1}{n_p}\sum_{i:p_i^*(\epsilon)=p} (Y_i(t) - Y_i(t_0)) \\
&= \sum_p \frac{p_t n_p}{\sum_i p_{it}^*(\epsilon)} CATE_{pt}.
\end{aligned}$$

Since $\frac{p_t n_p}{\sum_i p_{it}^*(\epsilon)}$ is known to the experimenter and $CATE_{pt}$ is unbiasedly estimable with EXAM and $\hat{\beta}_{pt}$, ATT is unbiasedly estimable with EXAM and $\sum_p \frac{p_t n_p}{\sum_i p_{it}^*(\epsilon)}\hat{\beta}_{pt}$.

Proof of Proposition 5

I define $N_{pt} \equiv \sum_i 1\{p_i^*(\epsilon) = p\}D_{it}$ as a random variable that stands for the number of subjects with propensity vector p and assigned to treatment t. Denote the realization of N_{pt} by $n_{pt} \equiv \sum_i 1\{p_i^*(\epsilon) = p\}d_{it}$. Recall that $\underline{n}_{pt}$ is defined as the greatest integer less than or equal to $p_t n_p$ (the expected number of subjects with propensity vector p and assigned to treatment t). Define $\mathcal{N}$ as the set of all (n_{pt}) that satisfy the following:

- $n_{pt} = \underline{n}_{pt}$ for all p and t such that $p_t n_p \in \mathbb{N}$.
- $n_{pt} \in \{\underline{n}_{pt}, \underline{n}_{pt} + 1\}$ for all p and t such that $p_t n_p \notin \mathbb{N}$.
- $\sum_t n_{pt} = n_p$ for all p.
- $\sum_p n_{pt} = \sum_i p_{it}^*(\epsilon)$ for all t.

I also define $D(n_{pt})$ as the set of deterministic treatment assignments where the realization of (N_{pt}) is (n_{pt}):

$$D(n_{pt}) \equiv \{d \in \{0,1\}^{n\times(m+1)} | \sum_t d_{it} = 1 \text{ for every } i \text{ and } \sum_i 1\{p_i^*(\epsilon) = p\}d_{it} = n_{pt} \text{ for every } p \text{ and } t\}.$$

The method of drawing deterministic treatment assignments in Definition 2 in Section 5 and Appendix A.1.1 satisfies the following properties.

Lemma 7 (Small Support)**.** *The support of* (N_{pt}) *is included by* $\mathcal{N}$*.*

Lemma 8 (Conditional Uniformity)**.** *Conditional on any* (n_{pt}) *in the support of* (N_{pt})*, every deterministic treatment assignment consistent with* (n_{pt}) *happens with equal probability:*

$$Pr(D = d|(N_{pt}) = (n_{pt}), p^*(\epsilon)) = \begin{cases} |D(n_{pt})|^{-1} & \text{if } d \in D(n_{pt}) \\ 0 & \text{otherwise.} \end{cases}$$

To show the mean part of Proposition 5, note that by Lemma 8, every feasible treatment assignment occurs equally likely conditional on (N_{pt}) so that for every p, t and i with $p_i^*(\epsilon) = p$,

$$E(D_{it}|(N_{pt}) = (n_{pt}), p_i^*(\epsilon) = p) = \frac{n_{pt}}{n_p}. \tag{19}$$

I therefore have

$$E(\hat{\beta}_t^*|(N_{pt}) = (n_{pt}), p_i^*(\epsilon) = p)$$

$$
\begin{aligned}
&= E(\sum_p \delta_p \hat{\beta}_{pt} | (N_{pt}) = (n_{pt}), p_i^*(\epsilon) = p) \\
&= \sum_p \delta_p E(\hat{\beta}_{pt} | (N_{pt}) = (n_{pt}), p_i^*(\epsilon) = p) \\
&= \sum_p \delta_p E\big[\sum_i 1\{p_i^*(\epsilon) = p\}\big(\frac{D_{it} Y_i(t)}{N_{pt}} - \frac{D_{it_0} Y_i(t_0)}{N_{pt_0}}\big) | (N_{pt}) = (n_{pt}), p_i^*(\epsilon) = p\big] \\
&= \sum_p \delta_p \sum_i 1\{p_i^*(\epsilon) = p\}\big(\frac{E(D_{it} | (N_{pt}) = (n_{pt}), p_i^*(\epsilon) = p) Y_i(t)}{n_{pt}} - \frac{E(D_{it_0} | (N_{pt}) = (n_{pt}), p_i^*(\epsilon) = p) Y_i(t_0)}{n_{pt_0}}\big) \\
&= \sum_p \delta_p \sum_i 1\{p_i^*(\epsilon) = p\}\big(\frac{(n_{pt}/n_p) Y_i(t)}{n_{pt}} - \frac{(n_{pt_0}/n_p) Y_i(t_0)}{n_{pt_0}}\big) \\
&= \sum_p \delta_p \frac{1}{n_p} \sum_i 1\{p_i^*(\epsilon) = p\}(Y_i(t) - Y_i(t_0)) \\
&= \sum_p \delta_p CATE_{pt},
\end{aligned}
$$

where I use equation (19) for the fifth equality. By the law of iterated expectation, I conclude

$$
\begin{aligned}
E(\hat{\beta}_t^* | p^*(\epsilon)) &= E[E(\hat{\beta}_t^* | (N_{pt}) = (n_{pt}), p^*(\epsilon)) | p^*(\epsilon)] \\
&= E[\sum_p \delta_p CATE_{pt} | p^*(\epsilon)] \\
&= \sum_p \delta_p CATE_{pt}.
\end{aligned}
$$

For the variance part of Proposition 5, I prove the general version given in Appendix A.1.1. For notational simplicity, I make conditioning on $p^*(\epsilon)$ implicit. By the law of total variance, $V(\hat{\beta}_t^*)$ can be written as:

$$
V(\hat{\beta}_t^*) = E(V(\hat{\beta}_t^* | (N_{pt}))) + V(E(\hat{\beta}_t^* | (N_{pt}))).
$$

As I show above, $E(\hat{\beta}_t^* | (N_{pt})) = \sum_p \delta_p CATE_{pt}$, implying $V(E(\hat{\beta}_t^* | (N_{pt}))) = 0$. Thus

$$
V(\hat{\beta}_t^*) = E(V(\hat{\beta}_t^* | (N_{pt}))). \tag{20}
$$

To show that $E(V(\hat{\beta}_t^* | (N_{pt})))$ is equal to the expression in Proposition 5, I introduce a lemma.

Lemma 9. *For all (n_{pt}) in the support of (N_{pt}),*

$$
V(\hat{\beta}_t^* | (N_{pt}) = (n_{pt})) = \sum_p \delta_p^2 \Big(\frac{S_{pt}^2}{n_{pt}} + \frac{S_{pt_0}^2}{n_{pt_0}} - \frac{S_{ptt_0}^2}{n_p}\Big).
$$

Proof of Lemma 9. By Lemma 8, treatment assignments are independent across subpopulations with different propensities conditional on (N_{pt}). $\hat{\beta}_{pt}$ is therefore independent across p conditional on (N_{pt}). Hence,

$$\begin{aligned} V(\hat{\beta}_t^*|(N_{pt}) = (n_{pt})) &= V(\sum_p \delta_p \hat{\beta}_{pt}|(N_{pt}) = (n_{pt})) \\ &= \sum_p \delta_p^2 V(\hat{\beta}_{pt}|(N_{pt}) = (n_{pt})). \end{aligned}$$

It is therefore enough to show that $V(\hat{\beta}_{pt}|(N_{pt}) = (n_{pt})) = \frac{S_{pt}^2}{n_{pt}} + \frac{S_{pt_0}^2}{n_{pt_0}} - \frac{S_{ptt_0}^2}{n_p}$. For notational simplicity, I make conditioning on $(N_{pt}) = (n_{pt})$ implicit. Let I^{ptt_0} be a random set of subjects with propensity vector p and assigned to either treatment t or t_0, i.e.,

$$I^{ptt_0} \equiv \{i|p_i^*(\epsilon) = p \text{ and } D_{it} + D_{it_0} = 1\}.$$

I^{ptt_0} takes on $\binom{n_p}{n_{pt} + n_{pt_0}}$ values equally likely, a consequence of Lemma 8. By the law of total variance, $V(\hat{\beta}_{pt})$ can be written as:

$$V(\hat{\beta}_{pt}) = E(V(\hat{\beta}_{pt}|I^{ptt_0})) + V(E(\hat{\beta}_{pt}|I^{ptt_0})). \tag{21}$$

Conditional on $I^{ptt_0} = I$, the randomness in $\hat{\beta}_{pt}$ comes from the randomness in choosing n_{pt} subjects assigned to treatment t and n_{pt_0} subjects assigned to treatment t_0 from the set I of $n_{pt} + n_{pt_0}$ subjects. Every combination occurs with equal probability, so the standard results of binary-treatment RCT (Theorems 6.1 and 6.2 in Imbens and Rubin (2015)) apply:

$$\begin{aligned} E(\hat{\beta}_{pt}|I^{ptt_0} = I) &= \frac{1}{n_{pt} + n_{pt_0}} \sum_{i \in I} (Y_i(t) - Y_i(t_0)), \\ V(\hat{\beta}_{pt}|I^{ptt_0} = I) &= \frac{S_{pt|I}^2}{n_{pt}} + \frac{S_{pt_0|I}^2}{n_{pt_0}} - \frac{S_{ptt_0|I}^2}{n_{pt} + n_{pt_0}}, \end{aligned}$$

where $S_{pt|I}^2$, $S_{pt_0|I}^2$ and $S_{ptt_0|I}^2$ are the variances of $Y_i(t)$, $Y_i(t_0)$ and $Y_i(t) - Y_i(t_0)$, respectively, conditional on the set of subjects I. Regarding $n_p, n_{pt} + n_{pt_0}$, and $Y_i(t) - Y_i(t_0)$ as performing the roles of n, m, and X_i in Lemma 4, respectively, I use Lemma 4 to get

$$\begin{aligned} V(E(\hat{\beta}_{pt}|I^{ptt_0} = I)) &= V(\frac{1}{n_{pt} + n_{pt_0}} \sum_{i \in I} (Y_i(t) - Y_i(t_0))) = \frac{n_p - n_{pt} - n_{pt_0}}{n_p(n_{pt} + n_{pt_0})} S_{ptt_0}^2, \\ E(V(\hat{\beta}_{pt}|I^{ptt_0} = I)) &= \frac{E(S_{pt|I}^2)}{n_{pt}} + \frac{E(S_{pt_0|I}^2)}{n_{pt_0}} - \frac{E(S_{ptt_0|I}^2)}{n_{pt} + n_{pt_0}} = \frac{S_{pt}^2}{n_{pt}} + \frac{S_{pt_0}^2}{n_{pt_0}} - \frac{S_{ptt_0}^2}{n_{pt} + n_{pt_0}}, \end{aligned}$$

where the last equality is by the second part of Lemma 4. Combining these with equation (21), I have $V(\hat{\beta}_{pt}) = \frac{S^2_{pt}}{n_{pt}} + \frac{S^2_{pt_0}}{n_{pt_0}} - \frac{S^2_{ptt_0}}{n_p}$. □

By Lemma 7, N_{pt} can take on either $\underline{n}_{pt}$ or $\underline{n}_{pt} + 1$. Since N_{pt} has expectation $p_t n_p$, the marginal distribution for each N_{pt} must be

$$\Pr(N_{pt} = n_{pt}) = \begin{cases} 1 - p_t n_p + \underline{n}_{pt} & \text{if } n_{pt} = \underline{n}_{pt} \\ p_t n_p - \underline{n}_{pt} & \text{if } n_{pt} = \underline{n}_{pt} + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Using equation (20), Lemma 9, and equation (22), I have

$$\begin{aligned} V(\hat{\beta}^*_t) &= E\Big\{\sum_p \delta_p^2 \Big(\frac{S^2_{pt}}{N_{pt}} + \frac{S^2_{pt_0}}{N_{pt_0}} - \frac{S^2_{ptt_0}}{n_p}\Big)\Big\} \\ &= \sum_p \delta_p^2 \Big\{ \sum_{t' \in \{t_0, t\}} E\Big(\frac{S^2_{pt'}}{N_{pt'}}\Big) - \frac{S^2_{ptt_0}}{n_p}\Big\} \\ &= \sum_p \delta_p^2 \Big\{ \sum_{t' \in \{t_0, t\}} \Big[\Big(\frac{S^2_{pt'}}{\underline{n}_{pt'}}\Big)(1 - p_{t'} n_p + \underline{n}_{pt'}) + \Big(\frac{S^2_{pt'}}{\underline{n}_{pt'} + 1}\Big)(p_{t'} n_p - \underline{n}_{pt'})\Big] - \frac{S^2_{ptt_0}}{n_p}\Big\}. \end{aligned}$$

Proof of Equation (4)

I prove equation (4) with two lemmas below.

Lemma 10. *$E(\hat{B}_t | p^*(\epsilon)) = \sum_p \lambda_{pt} CATE_{pt}$ for all t where $\hat{B}_t$ is the OLS estimate of B_t in this regression:*

$$Y_i = \sum_{t=t_1}^{t_m} B_t D_{it} + \sum_p C_p 1\{p^*_i(\epsilon) = p\} + E_i. \tag{23}$$

Proof of Lemma 10. I reparametrize the regression as follows with (B_t, D_p), where $D_p \equiv C_p + \sum_{t=t_1}^{t_m} B_t p_t$.

$$Y_i = \sum_{t=t_1}^{t_m} B_t (D_{it} - p^*_{it}(\epsilon)) + \sum_p D_p 1\{p^*_i(\epsilon) = p\} + E_i. \tag{24}$$

This reparametrization does not change $\hat{B}_t$. Note also that Y_i can be written as follows.

$$Y_i = \sum_p 1\{p_i^*(\epsilon) = p\}Y_p(0) + \sum_p \sum_{t=t_1}^{t_m} 1\{p_i^*(\epsilon) = p\}CATE_{pt}D_{it} + \mu_i,$$

where $Y_p(0) \equiv \frac{\sum_i 1\{p_i^*(\epsilon) = p\}Y_i(0)}{\sum_i 1\{p_i^*(\epsilon) = p\}}$ and $\sum_i 1\{p_i^*(\epsilon) = p\}\mu_i = 0$ for every p. Therefore, the OLS estimates $(\hat{B}_t, \hat{D}_p)$ of (B_t, D_p) in regression (24) can be written as follows.

$$\begin{aligned}
(\hat{B}_t, \hat{D}_p) = \arg\min_{(B_t, D_p)} \sum_i [\sum_p 1\{p_i^*(\epsilon) = p\}Y_p(0) + \sum_p \sum_{t=t_1}^{t_m} 1\{p_i^*(\epsilon) = p\}CATE_{pt}D_{it} \\
- \sum_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon)) - \sum_p D_p 1\{p_i^*(\epsilon) = p\}]^2 \\
= \arg\min_{(B_t, D_p)} \sum_i [\sum_p 1\{p_i^*(\epsilon) = p\}(Y_p(0) - D_p + \sum_{t=t_1}^{t_m} CATE_{pt}D_{it}) - \sum_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon))]^2 \\
= \arg\min_{(B_t, D_p)} \sum_i [\{\sum_p 1\{p_i^*(\epsilon) = p\}(Y_p(0) - D_p + \sum_{t=t_1}^{t_m} CATE_{pt}D_{it}\}^2 \\
- 2\sum_p 1\{p_i^*(\epsilon) = p\}(Y_p(0) - D_p + \sum_{t=t_1}^{t_m} CATE_{pt}D_{it}) \sum_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon)) \\
+ \{\sum_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon)\}^2] \\
= \arg\min_{(B_t, D_p)} \sum_i [\{\sum_p 1\{p_i^*(\epsilon) = p\}(Y_p(0) - D_p + \sum_{t=t_1}^{t_m} CATE_{pt}D_{it})\}^2 \\
- 2\sum_p 1\{p_i^*(\epsilon) = p\} \sum_{t=t_1}^{t_m} CATE_{pt}D_{it} \sum_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon)) + \{\sum_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon))\}^2]
\end{aligned}$$

because $\sum_i (D_{it} - p_{it}^*(\epsilon)) = 0$. Minimizing this over B_t leads to

$$\hat{B}_t = \frac{\sum_i \sum_p 1\{p_i^*(\epsilon) = p\}CATE_{pt}D_{it}(D_{it} - p_{it}^*(\epsilon))}{\sum_i (D_{it} - p_{it}^*(\epsilon))^2}.$$

Because $P(D_{it} = 1) = \frac{\sum_p \sum_i 1\{p_i^*(\epsilon) = p\}p_{it}^*(\epsilon)}{n}$ and

$$P(p_i^*(\epsilon) = p | D_{it} = 1) = \frac{\sum_i 1\{p_i^*(\epsilon) = p\}p_{it}^*(\epsilon)}{\sum_q \sum_i 1\{(p_i^*(\epsilon) = q\}p_{it}^*(\epsilon)},$$

it follows that the numerator is equal to $\sum_p p_t(1-p_t)\delta_p CATE_{pt}$ and that the denominator is equal to $\sum_p p_t(1-p_t)\delta_p$. This implies that $E(\hat{B}_t|p^*(\epsilon)) = \sum_p \lambda_{pt} CATE_{pt}$. □

Lemma 11. *$\hat{B}_t = \hat{b}_t^*$ for any t and any realization of treatment assignment D_{it}.*

Proof of Lemma 11. By the Frisch-Waugh-Lovell theorem, the OLS estimates of (23) can be obtained by regressing each of Y_i and D_{it} on the fully-saturated propensity score controls and then using the residuals from these regressions as the dependent and independent variables for a bivariate regression that omits the propensity score controls. Consider the auxiliary regressions that produce these residualized variables: they have D_{it} on the left hand side, with a saturated control for $p_i^*(\epsilon)$ on the right. By the law of iterated expectations, the conditional expectation function associated with this auxiliary regression is

$$E[D_{it}|p_i^*(\epsilon)] = p_{it}^*(\epsilon).$$

In other words, the conditional expectation function $E[D_{it}|p_i^*(\epsilon)]$ is linear in regressors $p_{it}^*(\epsilon)$, so it and the associated auxiliary regression function coincide (note that I use a saturated model for $p_i^*(\epsilon)$). Therefore, regression (3), which additively separably and linearly controls for $p_{it}^*(\epsilon)$'s, produces the same estimate as regression (23). □

Proof of Proposition 6

The proof uses the following lemma.

Lemma 12. *Suppose Assumptions 1, 2, 3, 4, 5, 6, and 7 hold. For all g, as $N \to \infty$,*

$$\sqrt{n_g}(\hat{\beta}_{N,g}^* - \beta_{N,g}^{pop}) \xrightarrow{d} \mathcal{N}(0, H_g^{-1}(\rho\Delta_g^{cond} + (1-\rho)\Delta_g^{ehw})H_g^{-1}).$$
$$\sqrt{n_g}(\hat{\beta}_{N,g}^* - \beta_{N,g}^{sample}) \xrightarrow{d} \mathcal{N}(0, H_g^{-1}\Delta_g^{cond}H_g^{-1}).$$

Proof of Lemma 12. This result is a consequence of Theorem 3 of Abadie et al. (2017). To verify their assumptions hold, fix any g, and regard subpopulation $P_{N,g}$ as the entire population. Note that $N_g \to \infty$ as $N \to \infty$ by Assumption 2 (ii). $D_{N,i}$ and 1 in my notation correspond to $U_{N,i}$ and $Z_{N,i}$ in Abadie et al. (2017). Their Assumption 3 holds by the following reason: For all g and $i \in P_{N,g}$,

$$\begin{aligned}\Pr((D_{N,it})_{t=t_0,\ldots,t_m} = d_i|R_N = r) &= E[\Pr((D_{N,it})_{t=t_0,\ldots,t_m} = d_i|R_N = r, (c_{N,t}))|R_N = r] \\ &= E[\sum_{t=t_0}^{t_m} p_{N,it}^*(\epsilon)1\{d_{it}=1\}|R_N = r]\end{aligned}$$

$$= \sum_{t=t_0}^{t_m} q_{N,g,t} 1\{d_{it} = 1\},$$

where the third equality holds by my Assumption 3. Then, $\Pr((D_{N,it})_{t=t_0,...,t_m} = d_i) = E[\Pr((D_{N,it})_{t=t_0,...,t_m} = d_i | R_N = r)] = \Pr((D_{N,it})_{t=t_0,...,t_m} = d_i | R_N = r)$. Under my Assumption 4,

$$\Pr((D_{N,it}) = (d_i) | R_N = r) = \prod_{i=1}^{N} \Pr((D_{N,it})_{t=t_0,...,t_m} = d_i | R_N = r) = \prod_{i=1}^{N} \Pr((D_{N,it})_{t=t_0,...,t_m} = d_i),$$

and

$$\Pr((D_{N,it}) = (d_i)) = E[\Pr((D_{N,it}) = (d_i) | R_N = r)] = \prod_{i=1}^{N} \Pr((D_{N,it})_{t=t_0,...,t_m} = d_i).$$

Therefore, $(D_{N,1,t})_{t=t_0,...,t_m}, ..., (D_{N,N,t})_{t=t_0,...,t_m}$ are jointly independent from each other, and independent of R_N. Assumption 3 in Abadie et al. (2017) therefore holds. Since $E[D_{N,i}] = (q_{N,g,t_1}, ..., q_{N,g,t_m})'$ for $i \in P_{N,g}$, Assumption 7 in Abadie et al. (2017) holds. Note that

$$Y_{N,i} = \sum_{t=t_0}^{t_m} D_{N,it} Y_{N,i}(t) = \sum_{t=t_1}^{t_m} D_{N,it}(Y_{N,i}(t) - Y_{N,i}(t_0)) + \sum_{t=t_0}^{t_m} D_{N,it} Y_{N,i}(t_0) = D'_{N,i}\beta_{N,i} + Y_{N,i}(t_0),$$

implying Assumption 8 in Abadie et al. (2017) holds. I next show that my $\beta_{N,g}^{pop}$ and $\beta_{N,g}^{sample}$ are equal to their θ_N^{causal} and $\theta_N^{causal,sample}$, respectively. To make it explicit that θ_N^{causal} and $\theta_N^{causal,sample}$ vary across g in our setting, denote them by $\theta_{N,g}^{causal}$ and $\theta_{N,g}^{causal,sample}$. Since $E[X_{N,i}] = 0$ and $E[X_{N,i}X'_{N,i}]$ is constant across $i \in P_{N,g}$,

$$\begin{aligned}
\theta_{N,g}^{causal} &\equiv \Big(\frac{1}{N_g} \sum_{i \in P_{N,g}} E[X_{N,i}X'_{N,i}]\Big)^{-1} \frac{1}{N_g} \sum_{i \in P_{N,g}} E[X_{N,i}Y_{N,i}] \\
&= (E[X_{N,1}X'_{N,1}])^{-1} \frac{1}{N_g} \sum_{i \in P_{N,g}} E[X_{N,i}(D'_{N,i}\beta_{N,i} + Y_{N,i}(t_0))] \\
&= (E[X_{N,1}X'_{N,1}])^{-1} \frac{1}{N_g} \sum_{i \in P_{N,g}} E[X_{N,i}((X_{N,i} + E[D_{N,i}])'\beta_{N,i} + Y_{N,i}(t_0))] \\
&= (E[X_{N,1}X'_{N,1}])^{-1} E[X_{N,1}X'_{N,1}] \frac{1}{N_g} \sum_{i \in P_{N,g}} \beta_{N,i} \\
&= \beta_{N,g}^{pop},
\end{aligned}$$

and

$$\begin{aligned}
\theta_{N,g}^{causal,sample} &\equiv \big(\frac{1}{n_g}\sum_{i\in P_{N,g}} R_{N,i}E[X_{N,i}X_{N,i}']\big)^{-1}\frac{1}{n_g}\sum_{i\in P_{N,g}} R_{N,i}E[X_{N,i}Y_{N,i}] \\
&= (E[X_{N,1}X_{N,1}'])^{-1}\frac{1}{n_g}\sum_{i\in P_{N,g}} R_{N,i}E[X_{N,i}((X_{N,i}+E[D_{N,i}])'\beta_{N,i}+Y_{N,i}(t_0))] \\
&= (E[X_{N,1}X_{N,1}'])^{-1}E[X_{N,1}X_{N,1}']\frac{1}{n_g}\sum_{i\in P_{N,g}} R_{N,i}\beta_{N,i} \\
&= \beta_{N,g}^{sample}.
\end{aligned}$$

Note that $\gamma_{N,g}^{causal}$ in Abadie et al. (2017)'s notation is

$$\gamma_{N,g}^{causal} = \frac{1}{N_g}\sum_{i\in P_{N,g}} E[Y_{N,i}] = \frac{1}{N_g}\sum_{i\in P_{N,g}} E[D_{N,i}'\beta_{N,i}+Y_{N,i}(t_0)] = E[D_{N,1}']\beta_{N,g}^{pop} + \frac{1}{N_g}\sum_{i\in P_{N,g}} Y_{N,i}(t_0).$$

For $i \in P_{N,g}$,

$$\begin{aligned}
\epsilon_{N,i} &= D_{N,i}'(\beta_{N,i}-\beta_{N,g}^{pop}) + Y_{N,i}(t_0) - \frac{1}{N_g}\sum_{i\in P_{N,g}} Y_{N,i}(t_0) \\
&= \sum_{t=t_1}^{t_m} D_{N,it}(Y_{N,i}(t)-Y_{N,i}(t_0)) - D_{N,i}'\theta_{N,g}^{causal} + Y_{N,i}(t_0) - \frac{1}{N_g}\sum_{i\in P_{N,g}} Y_{N,i}(t_0) \\
&= Y_{N,i} - D_{N,i}'\theta_{N,g}^{causal} - \frac{1}{N_g}\sum_{i\in P_{N,g}} Y_{N,i}(t_0) \\
&= Y_{N,i} - X_{N,i}'\theta_{N,g}^{causal} - E[D_{N,1}']\beta_{N,g}^{pop} - \frac{1}{N_g}\sum_{i\in P_{N,g}} Y_{N,i}(t_0) \\
&= Y_{N,i} - X_{N,i}'\theta_{N,g}^{causal} - \gamma_{N,g}^{causal},
\end{aligned}$$

where the last equality is by $\gamma_{N,g}^{causal} = E[D_{N,1}']\beta_{N,g}^{pop} + \frac{1}{N_g}\sum_{i\in P_{N,g}} Y_{N,i}(t_0)$ shown above. It only remains to check Abadie et al. (2017)'s Assumption 5 holds. Since $\|X_{N,i}\| \le \sum_{t=t_1}^{t_m}|X_{N,it}| \le \sum_{t=t_1}^{t_m}(|D_{N,it}| + |p_{N,it}^*(\epsilon)|) \le 2$ with probability one for all i, $\frac{1}{N_g}\sum_{i\in P_{N,g}} E[\|X_{N,i}\|^{4+\delta}] \le \frac{1}{N_g}\sum_{i\in P_{N,g}} 1 = 1$ for all N and for any $\delta > 0$. Hence, the sequences $\frac{1}{N_g}\sum_{i\in P_{N,g}} E[|Y_{N,i}|^{4+\delta}]$, $\frac{1}{N_g}\sum_{i\in P_{N,g}} E[\|X_{N,i}\|^{4+\delta}]$, and $\frac{1}{N_g}\sum_{i\in P_{N,g}} E[\|1\|^{4+\delta}]$ are uniformly bounded under my Assumption 5. Applying Abadie et al. (2017)'s Theorem 3 gives me Lemma 12. □

$\hat{\beta}_{N,g}^*$ is independent across g since each subject is sampled independently, and $(D_{N,1t})_{t=t_0,\ldots,t_m}$

$, ..., (D_{N,Nt})_{t=t_0,...,t_m}$ are jointly independent from each other and independent of R_N (a consequence of my Assumption 4). Notice that $E[n/N] = E[\sum_{i=1}^N R_{N,i}]/N = \rho_N$ and $Var(n/N) = Var(\sum_{i=1}^N R_{N,i})/N^2 = \rho_N(1-\rho_N)/N \to 0$. Thus, $n/N \xrightarrow{p} \rho_N$. Similarly, $n_g/N_g \xrightarrow{p} \rho_N$. The continuous mapping theorem implies $\sqrt{n_g/n} = \sqrt{\delta_{N,g}(n_g/N_g)/(n/N)} \xrightarrow{p} \sqrt{\delta_g}$. I have

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_N^* - \beta_N^{pop}) &= \sqrt{n}(\sum_{g=1}^G \frac{n_g}{n}\hat{\beta}_{N,g}^* - \sum_{g=1}^G \frac{N_g}{N}\beta_{N,g}^{pop}) \\
&= \sqrt{n}[\sum_{g=1}^G \frac{n_g}{n}(\hat{\beta}_{N,g}^* - \beta_{N,g}^{pop}) + \sum_{g=1}^G (\frac{n_g}{n} - \frac{N_g}{N})\beta_{N,g}^{pop}] \\
&= \sum_{g=1}^G [\sqrt{\frac{n_g}{n}}\sqrt{n_g}(\hat{\beta}_{N,g}^* - \beta_{N,g}^{pop}) + \sqrt{n}(\frac{n_g}{n} - \frac{N_g}{N})\beta_{N,g}^{pop}] \\
&\xrightarrow{d} \mathcal{N}(0, \sum_{g=1}^G \delta_g H_g^{-1}(\rho\Delta_g^{cond} + (1-\rho)\Delta_g^{ehw})H_g^{-1}).
\end{aligned}$$

where the last convergence is by $\sqrt{n_g/n} \xrightarrow{p} \sqrt{\delta_g}$, Lemma 12, and Assumption 8. Similarly,

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_N^* - \beta_N^{sample}) &= \sqrt{n}(\sum_{g=1}^G \frac{n_g}{n}\hat{\beta}_{N,g}^* - \sum_{g=1}^G \frac{n_g}{n}\beta_{N,g}^{sample}) \\
&= \sqrt{n}\sum_{g=1}^G \frac{n_g}{n}(\hat{\beta}_{N,g}^* - \beta_{N,g}^{sample}) \\
&= \sum_{g=1}^G \sqrt{\frac{n_g}{n}}\sqrt{n_g}(\hat{\beta}_{N,g}^* - \beta_{N,g}^{sample}) \\
&\xrightarrow{d} \mathcal{N}(0, \sum_{g=1}^G \delta_g H_g^{-1}\Delta_g^{cond}H_g^{-1}),
\end{aligned}$$

where the last convergence is by $\sqrt{n_g/n} \xrightarrow{p} \sqrt{\delta_g}$ and Lemma 12.

Proof of Corollary 4

Note that

$$\epsilon_{N,i} = D_{N,it_1}(Y_{N,i}(t_1) - \bar{Y}_N(t_1)) + (1 - D_{N,it_1})(Y_{N,i}(t_0) - \bar{Y}_N(t_0)).$$

I have

$$E[X_{N,i}\epsilon_{N,i}] = \Pr(D_{N,i} = 1)(1 - q_{N,t_1})(Y_{N,i}(t_1) - \bar{Y}_N(t_1)) + \Pr(D_{N,i} = 0)(-q_{N,t_1})(Y_{N,i}(t_0) - \bar{Y}_N(t_0))$$

$$= q_{N,t_1}(1-q_{N,t_1})(Y_{N,i}(t_1) - \bar{Y}_N(t_1) - (Y_{N,i}(t_0) - \bar{Y}_N(t_0))),$$

and

$$\begin{aligned} E[X_{N,i}^2\epsilon_{N,i}^2] &= \Pr(D_{N,i}=1)(1-q_{N,t_1})^2(Y_{N,i}(t_1) - \bar{Y}_N(t_1))^2 + \Pr(D_{N,i}=0)(-q_{N,t_1})^2(Y_{N,i}(t_0) - \bar{Y}_N(t_0))^2 \\ &= q_{N,t_1}(1-q_{N,t_1})^2(Y_{N,i}(t_1) - \bar{Y}_N(t_1))^2 + (1-q_{N,t_1})q_{N,t_1}^2(Y_{N,i}(t_0) - \bar{Y}_N(t_0))^2. \end{aligned}$$

I therefore get $\Delta^{ehw} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N E(X_{N,i}^2\epsilon_{N,i}^2) = q_{t_1}(1-q_{t_1})^2 S_{t_1}^2 + (1-q_{t_1})q_{t_1}^2 S_{t_0}^2$ and $\Delta^{cond} = \Delta^{ehw} - \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N E[X_{N,i}\epsilon_{N,i}]^2 = \Delta^{ehw} - q_{t_1}^2(1-q_{t_1})^2 S_{t_1 t_0}^2$. I also have

$$\begin{aligned} E[X_{N,i}^2] &= \Pr(D_{N,i}=1)(1-q_{N,t_1})^2 + \Pr(D_{N,i}=0)(-q_{N,t_1})^2 \\ &= q_{N,t_1}(1-q_{N,t_1}), \end{aligned}$$

leading to $H = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N E[X_{N,i}^2] = q_{t_1}(1-q_{t_1})$. (7) and (8) follow from substituting the above observations into Proposition 6. Analogously, (9) and (10) follow from Corollary 3.

Proof of Proposition 7

By Proposition 2, there is no other experimental design (p_{it}) with $p_{it} \in [\epsilon, 1-\epsilon]$ for all subject i and treatment t and such that $\sum_t p_{it}w'_{it} \geq \sum_t p_{it}^{*o}(\epsilon)w'_{it}$ for all i and $\sum_t p_{it}e'_{ti} \geq \sum_t p_{it}^{*o}(\epsilon)e'_{ti}$ for all i with at least one strict inequality. w'_{it} and e'_{ti} are consistent with ordinal $\succsim_i$ and $\succsim_t$, respectively. Therefore, there is no other experimental design (p_{it}) such that for all cardinal WTP w_{it} consistent with ordinal $\succsim_i$ and all cardinal predicted effects e_{ti} consistent with ordinal $\succsim_t$, I have $\sum_t p_{it}w_{it} \geq \sum_t p_{it}^{*o}(\epsilon)w_{it}$ for all i and $\sum_t p_{it}e_{ti} \geq \sum_t p_{it}^{*o}(\epsilon)e_{ti}$ for all i with at least one strict inequality.

A.3 Empirical Details

A.3.1 Why Subject Welfare? Data

Table 1 Panel a is based on data I assemble from the WHO International Clinical Trials Registry Platform (ICTRP) at `http://www.who.int/ictrp/en/`, retrieved in May 2019. I first use the "date of registration" variable to define the year associated with each trial. Starting from the universe of trials registered between January 1st 2007 to May 31st 2017, I exclude outlier trials with registered sample size greater than 5 millions. Some trials come with sample size classified as "Not Specified." I set their sample size as zero. This makes

my total sample size calculation conservative. For a trial that does not have a well-defined trial phase, I classify its trial phase as "Not Specified." Finally, for each trial, I define its "Geographical Region" according to which country runs the registry including that trial. Many registries like ClinicalTrial.gov recruit subjects in multiple countries under the same trial ID, making it challenging to pin down the physical location of each trial.

Table 1 Panel b is based on data I assemble from the American Economic Association's registry (AEA registry) for randomized controlled trials at `https://www.socialscienceregistry.org`, retrieved on May 27th, 2017. From the AEA registry, I obtain information about each experiment such as the sample size, the year when the experiment was conducted, the country where the experiment was conducted, registered keywords, and the randomization unit. When some information is missing, I manually enter it by referring to accompanying documents such as experimental design descriptions and abstracts. I classify an item as "Not specified" when I cannot specify it even after the manual procedure. When the sample size of an experiment is unspecified, I set the sample size as zero. This makes my total sample size calculation conservative. I use the "starting date of experiment" to define the year associated with each trial. Finally, for each trial, I define its "Geographical Region" according to the country in which the experiment was conducted. I include all registered experiments conducted during 2007-2017 period.

A.3.2 Do Clinical Trials Use Simple Randomization?

Do clinical trials randomize treatment as in Definition 1 of RCT? I provide an answer to this question using the Clinical Trial Registry India (CTRI). To my knowledge, CTRI is the only major clinical trial registry that provides data about randomization methods in clinical trials. I assembled data about individual clinical trials including the date the trial was conducted and the method used to randomize subjects into control or treatment groups. The data includes trials spanning from October 9th, 2007 to October 9, 2017. I removed trials with sample size 0 and trials that have been classified as "NA" for randomization method. According to the CTRI description manual, the relevant variable ("method of generating random sequence") takes some of the following categories:

- *Computer generated randomization*: A machine randomly assigns the subject or subject group to a study or treatment group.
- *Permuted block randomization (fixed)*: Participants are randomly allocated in a way that maintains a covariate balance across treatment groups. Allocation occurs by assigning a specified number of participants to a block that has a specified number of treatment assignments. In this case, the block size is fixed.

- *Permuted block randomization (variable)*: Same method as permuted block randomization (fixed), but with varying block sizes.
- *Random number table*: Each subject or subject group is assigned a number, and a random number table determines if the subject is assigned to a control or treatment group.
- *Coin toss, lottery, toss of dice, shuffling cards etc.*: Based on a coin toss, the subject, or subject group is placed in either a control or treatment group.
- *Stratified randomization*: In order to control for covariates (patient characteristics which might affect the outcomes), a stratum is generated for each combination of covariates, and subjects are assigned to the appropriate strata of covariates. After all subjects are assigned to strata, simple randomization is performed within each stratum to assign subjects to a treatment or control group.
- *Stratified block randomization*: Same method as stratified randomization, but once patients are assigned to their strata, permuted block randomization is performed within each stratum.
- *Adaptive randomization*: Adaptive randomization, like stratified randomization, takes covariates into account. In the "minimization method," for example, a new patient is sequentially assigned to the group with the fewest number of existing patients with the same covariates, making covariates balanced across groups.
- *Other*: 179 trials listed "Other" as the method of randomization

The popularity of each randomization method is described in Appendix Table A.1. The table shows that 85% of all trials use one of the impersonal simple randomization methods that do not take patient covariate or past data into account. This suggests that Definition 1 of RCT is a reasonable approximation to most clinical trials.

A.3.3 Treatment Effects and Preferences: Details

Sample Restriction in Treatment Effect Estimation (Table 3)

For the OLS regressions in Table 3, I impose the same sample restriction as Kremer et al. and exclude the following children: children not at Intent-to-Treat springs, i.e., springs found to be nonviable after treatment random assignment, children in households that receive water guards in 2007, children not in representative households (defined as households that are

named at least twice by all users of a given spring when survey enumerators ask spring users at a spring to name households that also use the same spring), children above age 3 at baseline and children above age 3 when they join the sample in later rounds, children whose anthropometric (weight, height, BMI) and age data are flagged as having serious error, and children in households with missing data on whether they use the identified spring exclusively or use multiple springs.

Estimation of the Mixed Logit WTP Model (Table 4)

With the random utility function (6), choice likelihoods take the following form (Train (2003), chapter 6):

$$P(o_{ijt} = 1|\theta, \gamma_1, \delta_j) = \int_{(\beta_i, c_i)} \frac{\exp\left((\beta_i + \gamma_1 X_i)T_{jt} - c_i D_{ij} + \delta_j\right)}{\sum_{h \in H} \exp\left((\beta_i + \gamma_1 X_i)T_{ht} - c_i D_{ih} + \delta_h\right)} f(\beta_i, c_i|\theta) d(\beta_i, c_i)$$

where $o_{ijt} \in \{0, 1\}$ is the indicator that household i chooses source j in trip t among alternatives $h \in H$ and $f(\beta_i, c_i|\theta)$ is the mixing distribution parametrized by θ. $f(\beta_i, c_i|\theta)$ is taken to be the normal distribution with unknown mean and variance for the spring protection treatment coefficient β_i and the triangular distribution (restricted to be nonnegative) for the distance coefficient c_i. I use the quasi Newton method to maximize a simulation approximation of the joint likelihood $\sum_{ijt} P(o_{ijt} = 1|\theta, \gamma_1, \delta_j)$ with respect to θ, γ_1, and δ_j, producing maximum simulated likelihood estimates $\hat{\theta}, \hat{\gamma}_1$, and $\hat{\delta}_j$. I compute standard errors using the information matrix with the Hessian being estimated by the outer product of the gradient of the simulated likelihood at the estimated parameter value.

Simulation of WTP (Figure 1 Panel b and Subsequent Figures)

I create simulated WTP data for Figure 1 Panel b and subsequent figures with parametric bootstrap below.

(1) Simulate a value of the distance coefficient $c \sim Triangular(\widehat{\theta^D})$ for each household group sharing the same characteristics where $\widehat{\theta^D}$ is the point estimate of the parameter of the distance coefficient distribution, i.e., the estimated mean and standard deviation. To correct for potential measurement error in distance, follow Kremer et al. (2011)'s method and multiply the distance coefficient by -1/0.38, where 0.38 is the correlation across survey rounds in the reported walking distance to the reference spring and is taken to be the size of measurement error from recall error. See Kremer et al. (2011)'s Section IV.B for details.

(2) Draw a value of the treatment coefficient $w \sim N(\hat{\mu}, \hat{\sigma})$ for each household group sharing the same characteristics where $\hat{\mu}$ and $\hat{\sigma}$ are the point estimates of mean μ and standard deviation σ of the treatment coefficient distribution.

(3) Compute the ratio w/c as WTP for the treatment in terms of minutes of walking time. Follow Kremer et al. (2011)'s method to get WTP in terms of the number of workdays taken to walk to the spring in a year. Specifically, multiply the ratio by $(32 \times 52)/(60 \times 8)$ where 32×52 is the average number of water trips taken by a household per year and 60×8 is the number of minutes per workday. See Kremer et al. (2011)'s Section IV.B for details.

A.3.4 EXAM vs RCT: Algorithm Details

In this section, I describe the details of the algorithm I use for computing EXAM's treatment assignment probabilities $p^*_{it}(\epsilon)$ in my empirical application in Section 6.3. I first define subroutines and then call them together at the end to perform the main computation. Though simple, this algorithm works well in my application: The market clearing error, defined as $\sqrt{\sum_t (\sum_i p^*_{it} - c_t)^2} / \sum_t c_t$, is smaller than 0.005 in all simulation runs.

Algorithm 1 Experimental as Market (EXAM)

Input: n the number of subjects, m the number of treatments, $(c_t) \in \mathbb{N}$ treatment t's pseudo capacity with $\sum_t c_t = n$, (w_{it}) subject i's WTP for treatment t, (e_{ti}) treatment t's predicted treatment effect for subject i, b the budget constraint

Output: (p^*_{it}) treatment t's assignment probability for subject i, (α^*, β^*_t) parameters determining treatment t's equilibrium price of the form $\pi^*_{te} = \alpha^* e + \beta^*_t$, error_{min} minimized market clearing error relative the total capacity of treatments

function InitialAlpha( )
  $\alpha \leftarrow$ generate random number $\sim$ Uniform$(-b, 0)$ ▷ set the value of α
  return α

function InitBeta( )
  $\beta_{t_0} \leftarrow 0$
  for each $t = t_1, ..., t_m$ **do**
    $\beta_t \leftarrow$ generate random number $\sim$ Uniform$(-b, b)$ ▷ set the initial value of β_t
  return (β_t) ▷ return an m-dimensional vector

function Price(α, (β_t)) ▷ get the price of treatment t
  for each $i, t = t_1, ..., t_m$ **do**
    $\pi_{te_{ti}} = \alpha e_{ti} + \beta_t$
  return $(\pi_{te_{ti}})$ ▷ return the $n \times m$ price matrix

function Demand($(\pi_{te_{ti}})$) ▷ get subject i's demand for treatment t
  for each i **do** ▷ perform utility maximization for each subject i
    $(p_{it})_t \leftarrow \arg\max_{(p_{it})_t \in P} \sum_t w_{it} p_{it}$ s.t. $\sum_t \pi_{te_{ti}} p_{it} \leq b$
  return (p_{it}) ▷ return the $n \times m$ demand matrix

function ExcessDemand((p_{it})) ▷ get the excess demand for treatment t
  for each $t = t_1, ..., t_m$ **do**
    $d_t \leftarrow \sum_i p_{it} - c_t$
  return (d_t) ▷ return the m-dimensional excess demand vector

function ClearingError((d_t)) ▷ get the market clearing error
  if $d_t < 0$ for all t **then**
    return 0
  else
    error $\leftarrow \sqrt{\sum_t d_t^2} / \sum_t c_t$
    return error ▷ return the market clearing error

$\delta_\beta \leftarrow b/50$ ▷ scaling factor for β_t's to set new prices

function BetaNew((β_t, d_t)) ▷ recalibrate β_t's to set new prices
    for each $t = t_1, ..., t_m$ **do**
        $\beta_t^{new} \leftarrow \beta_t + d_t \delta_\beta$
    return (β_t^{new})

function ClearMarket( ) ▷ the main function
    $\alpha \leftarrow$ InitialAlpha( )
    $(\beta_t) \leftarrow$ InitBeta( )
    $(\pi_{te_{ti}}) \leftarrow$ Price$(\alpha, (\beta_t))$
    $(p_{it}) \leftarrow$ Demand$((\pi_{te_{ti}}))$
    $(d_t) \leftarrow$ ExcessDemand$((p_{it}))$
    error $\leftarrow$ ClearingError$((d_t))$
    $\text{error}_{min} \leftarrow$ error ▷ initialize the minimum of clearing error
    ClearingThreshold $\leftarrow 0.01$ ▷ threshold for the market clearing error
    IterationThreshold $\leftarrow 10$ ▷ threshold for iteration times
    iterations $\leftarrow 0$ ▷ initialize iteration time count
    while True **do**
    if iterations > IterationThreshold **then**
        $\alpha \leftarrow$ InitialAlpha( ) ▷ start new equilibrium research
        $(\beta_t) \leftarrow$ InitBeta( )
        iterations $\leftarrow 0$
    else
        $(\beta_t) \leftarrow$ BetaNew$((\beta_t), (d_t))$
    $(\pi_{te_{ti}}) \leftarrow$ Price$(\alpha, (\beta_t))$
    $(p_{it}) \leftarrow$ Demand$((\pi_{te_{ti}}))$
    $(d_t) \leftarrow$ ExcessDemand$((p_{it}))$
    error $\leftarrow$ ClearingError$((d_t))$
    if error $<$ error_{min} **then**
        $\text{error}_{min} \leftarrow$ error
        $\alpha^* \leftarrow \alpha$ ▷ the new prices reduce the error
        $(\beta_t^*) \leftarrow (\beta_t)$
        $(p_{it}^*) \leftarrow (p_{it})$
    if $\text{error}_{min} <$ ClearingThreshold **then**
        break
    iterations += 1
    return $((p_{it}^*), \alpha^*, (\beta_t^*), \text{error}_{min})$ ▷ return the outputs

A.3.5 Additional Tables and Figures

Table A.1: Do Clinical Trials Use Simple Randomization?

Frequency of Randomization Methods in Clinical Trial Registry India (CTRI)		
Randomization Methods	%	Subgroup %
Simple Randomization		
Computer generated randomization	63%	
Permuted block randomization, fixed	5%	
Random number table	6%	85%
Coin toss	8%	
Permuted block randomization, variable	2%	
Adaptive or Stratified Randomization		
Stratified block randomization	5%	
Stratified randomization	3%	9%
Adaptive randomization	1%	
Other	6%	6%
Total interventional trials	5733	

Notes: This table shows summary statistics of the popularity of different randomization methods in clinical trials, based on the Clinical Trial Registry India (CTRI). The data includes trials spanning from October 9th, 2007 to October 9, 2017. I removed trials with sample size 0 and trials that have been classified as "NA" for randomization method. See Appendix A.3.2 for discussions about this table.

Table A.2: A Selection of High-stakes RCTs (Continued from Table 2)

(a) Medical Clinical Trials

	Subjects	Sample Size
i	Coronary Heart Disease Patients	4444 Individuals
ii	Patients with Elevated Intraocular Pressure	1636 Individuals
iii	HIV Negative Gay Men and Transgender Women	2499 Individuals
iv	Serodiscordant Couples	1763 Couples
v	Postmenopausal Women	16608 Individuals

(b) Social and Economic Experiments

	Subjects	Sample Size
I	Poor Households in Kenya	940 Households
II	Crime Hot Spots in Minneapolis	110 Spots
III	Unmarried Women in Malawi	1007 Individuals
IV	Uninsured Individuals in Oregon	12229 Individuals
V	Public Sector Job Applicants in Mexico	350 Job Vacancies

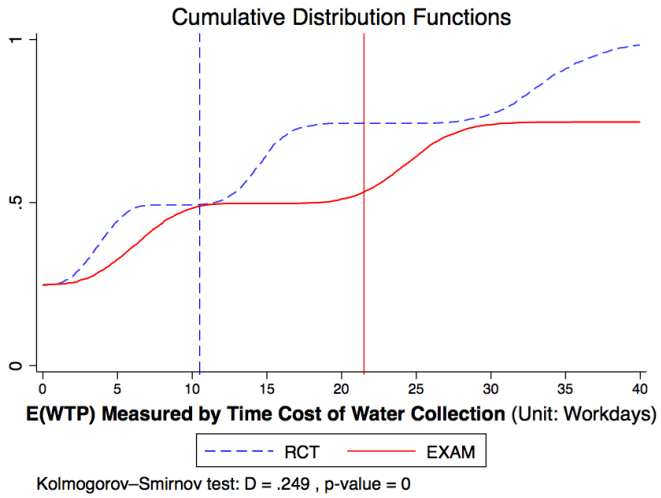
Notes: This is a continuation of Table 2. This table lists examples illustrating the high-stakes nature of certain RCTs. See the following references for the details of each RCT:

Panel a Study i: Scandinavian Simvastatin Survival Study Group and Others (1994)
Panel a Study ii: Kass et al. (2002)
Panel a Study iii: Grant et al. (2010)
Panel a Study iv: Cohen et al. (2011)
Panel a Study v: Writing Group for the Women's Health Initiative Investigators and Others (2002)
Panel b Study I: Haushofer and Shapiro (2016)
Panel b Study II: Sherman and Weisburd (1995)
Panel b Study III: Angelucci and Bennett (2017)
Panel b Study IV: Baicker et al. (2013)
Panel b Study V: Dal Bó et al. (2013), where the control is a lower wage job offer.

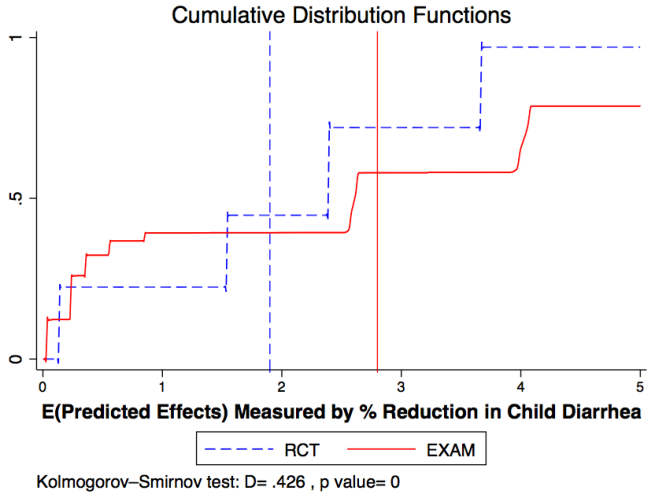
See Section 2 for discussions about this table.

Figure A.1: EXAM vs RCT: Welfare (Robustness Check with $\epsilon = 0.1$)

(a) Average WTP for Assigned Treatments w_i^*



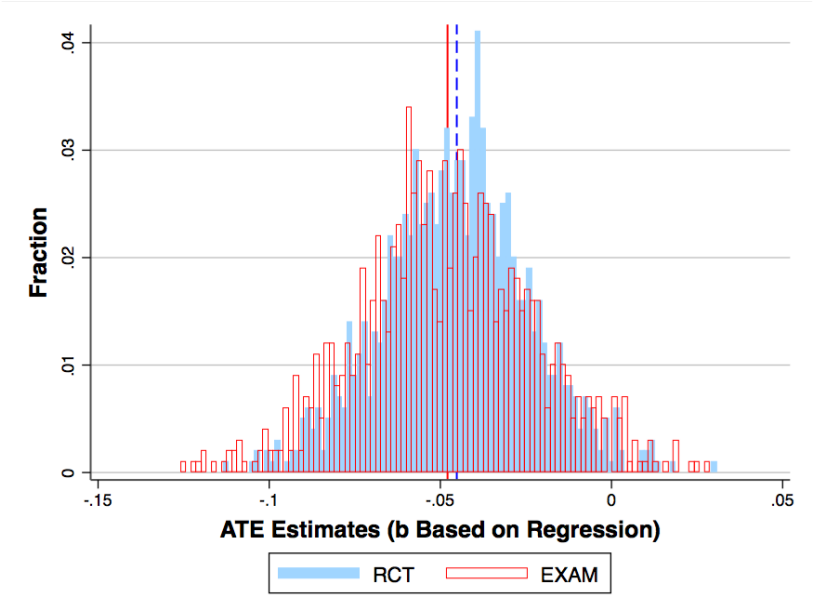
(b) Avg Predicted Effects of Assigned Treatments e_i^*



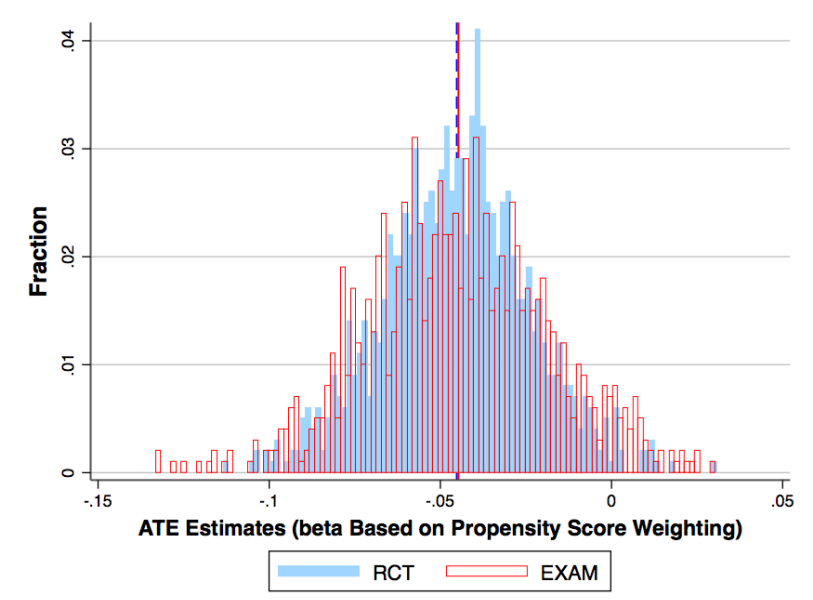
Notes: This figure reports the same results as Figure 2 except that this figure sets ϵ to 0.1. To compare EXAM and RCT's welfare performance, this figure shows the distribution of average subject welfare over 1000 bootstrap simulations under each experimental design. Panel a measures welfare with respect to average WTP w_i^* for assigned treatments while Panel b with respect to average predicted effects e_i^* of assigned treatments. A dotted line indicates the distribution of each welfare measure for RCT while a solid line indicates that for EXAM. Each vertical line represents mean. Kolmogorov-Smirnov tests find the EXAM and RCT distributions to be significantly different both for w_i^* and e_i^*. Both predicted effects $\hat{e}_{t_1 i}$ and WTP $\hat{w}_{it_1}$ are based on the main statistical specifications including all of the interactions between the treatment indicator and household characteristics (baseline latrine density, diarrhea prevention knowledge score, and mother's years of education). See Section 6.3 for discussions about this figure.

Figure A.2: EXAM vs RCT: ATE Estimates (Robustness Check with $\epsilon = 0.1$)

(a) Distribution of Average Treatment Effect Estimates $\hat{b}^*$



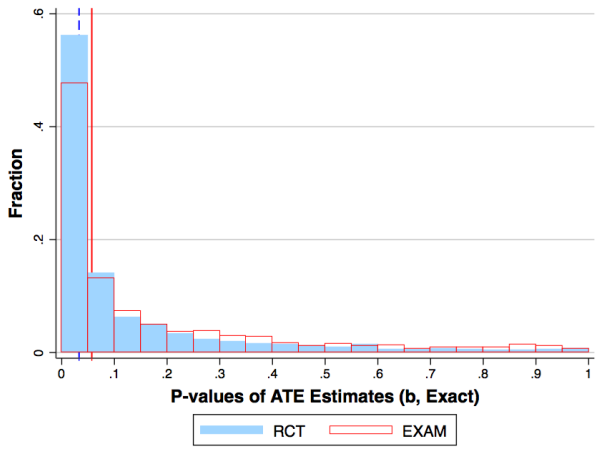
(b) Distribution of Average Treatment Effect Estimates $\hat{\beta}^*$



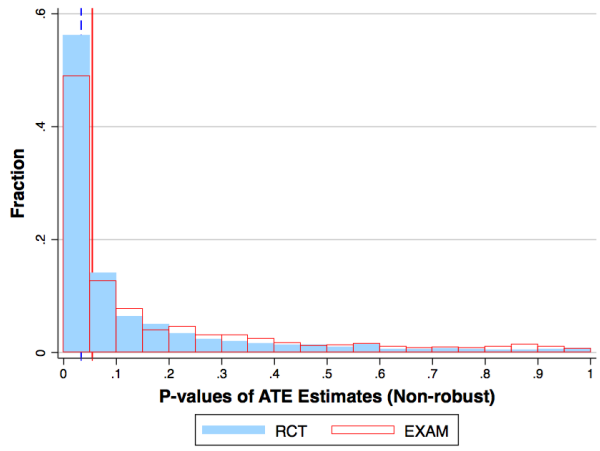
Notes: This figure reports the same results as Figure 3 except that this figure sets ϵ to 0.1. This figure compares EXAM and RCT's causal inference performance by showing the distribution of average treatment effect estimates under each experimental design. Grey bins indicate average treatment effect estimates for RCT while transparent bins with black outlines indicate those for EXAM. The solid vertical line indicates mean for EXAM while the dashed vertical line indicates that for RCT. See Section 6.3 for discussions about this figure.

Figure A.3: EXAM vs RCT: p Values (Robustness Check with $\epsilon = 0.1$)

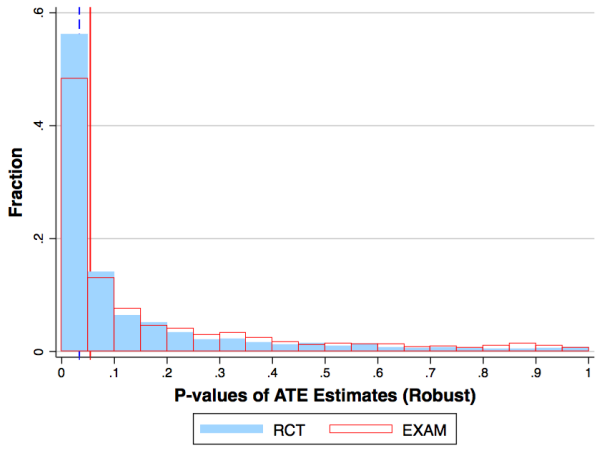
(a) p Values for $\hat{b}^*$ (Exact, Finite Sample)



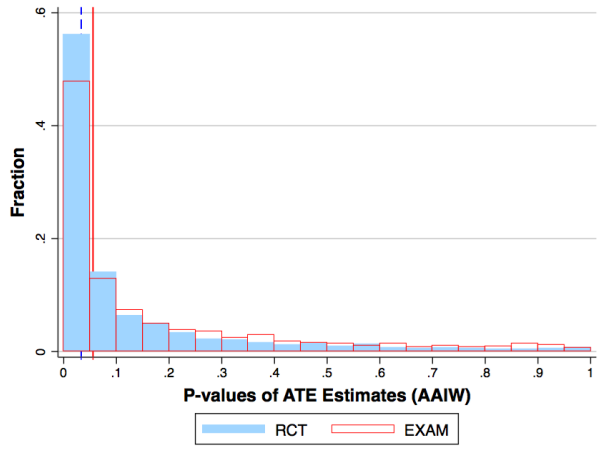
(b) p Values for $\hat{b}^*$ (Non-robust)



(c) p Values for $\hat{b}^*$ (Robust)



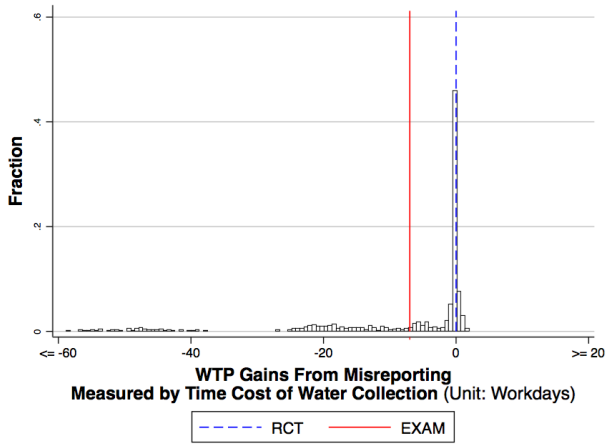
(d) p Values for $\hat{b}^*$ (Abadie et al., 2017)



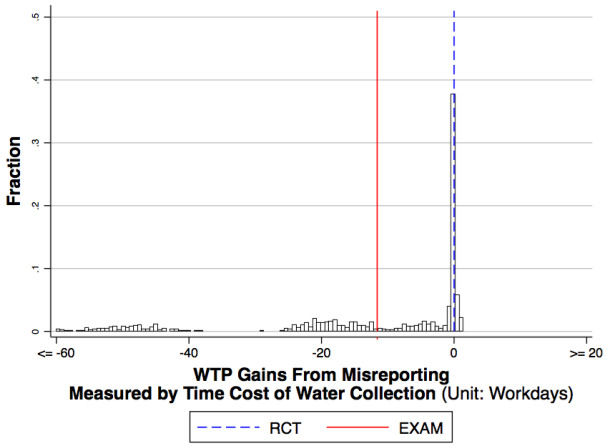
Notes: This figure reports the same results as Figure 4 except that this figure sets ϵ to 0.1. This figure compares EXAM and RCT's causal inference performance by showing the distribution of p values accompanying average treatment effect estimates $\hat{b}^*$ under each experimental design. The p values are based on exact, non-robust, robust, or Abadie et al. (2017)'s standard errors. Grey bins indicate p values for RCT while transparent bins with black outlines indicate those for EXAM. The solid vertical line indicates median for EXAM while the dashed vertical line indicates that for RCT. See Section 6.3 for discussions about this figure.

Figure A.4: EXAM vs RCT: Incentive (Robustness Check with $\epsilon = 0.1$)

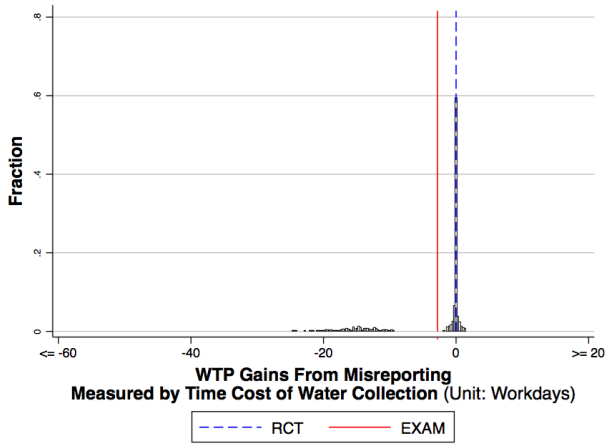
(a) WTP manipulation $\sim$ true WTP+$N(0, 100)$



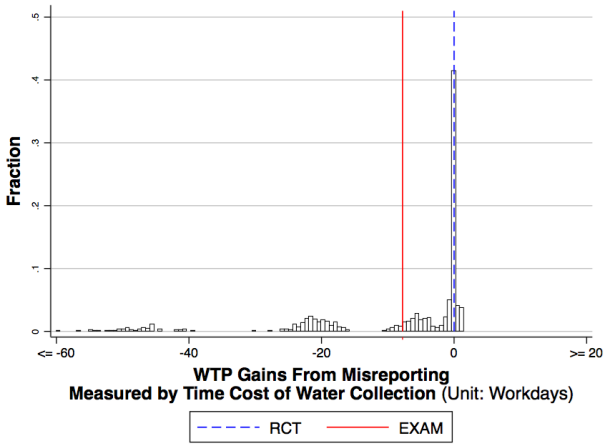
(b) WTP manipulation $\sim$ true WTP+$N(0, 1000)$



(c) WTP manipulation $\sim$ true WTP+$U(0, 100)$



(d) WTP manipulation $\sim$ true WTP+$U(-100, 0)$



Notes: This figure reports the same results as Figure 5 except that this figure sets ϵ to 0.1. This figure shows the histogram of true WTP gains from potential WTP misreporting to EXAM, quantifying the incentive compatibility of EXAM. Different panels use different ways of drawing WTP manipulations indicated by the panel titles. Each solid vertical line represents the mean WTP gain from potential WTP misreporting to EXAM. The dash vertical line is for RCT, where the true WTP gain from any WTP misreport is zero. See Section 6.3 for discussions about this figure.

Table A.3: EXAM vs RCT: Incentive Details

	$\Delta w/w_{it_1}$ True WTP Gains from Misreporting Relative to True WTP for Treatment t_1					
	95 Percentile	96 Percentile	97 Percentile	98 Percentile	99 Percentile	Max
Figure A.4 Panel a $\epsilon = 0.1$ & manipulation $\sim N(0, 100)$	1.37%	1.47%	1.53%	1.64%	1.76%	2.32%
Figure A.4 Panel b $\epsilon = 0.1$ & manipulation $\sim N(0, 1000)$	1.08%	1.18%	1.32%	1.52%	1.76%	2.38%
Figure A.4 Panel c $\epsilon = 0.1$ & manipulation $\sim U(0, 100)$	1.21%	1.33%	1.39%	1.54%	1.64%	2.06%
Figure A.4 Panel d $\epsilon = 0.1$ & manipulation $\sim U(-100, 0)$	1.11%	1.24%	1.40%	1.52%	1.64%	1.93%

Notes: This table provides additional details about Figure A.4. In particular, for each scenario in Figure A.4, this table shows the most profitable true WTP gains from potential WTP misreporting to EXAM, quantifying the incentive compatibility of EXAM. See Section 6.3 for discussions about this table.