EXAMining Welfare in Randomized Experiments

Yusuke Narita[a,1]

[a]Department of Economics and Cowles Foundation, Yale University, New Haven, CT 06511.

This manuscript was compiled on June 30, 2020

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 welfare-maximizing allocation of treatment assignment probabilities, is almost 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 (1). In this empirical setting, compared to standard RCTs, EXAM improves subjects' predicted well-being while reaching similar treatment effect estimates with similar precision.

social experiment | clinical trial | A/B test | mechanism design | experimental design | development economics

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 involve large numbers of participants. Between 2007 and 2017, over 360 million patients and 22 million individuals participated in registered clinical trials and social RCTs, respectively (see the Supplementary Information). Moreover, these experiments often 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 (2). In expectation, therefore, the lives of up to 9% of the study's 573 participants depend on who receives treatments. Social RCTs also often randomize critical treatments such as basic income, high wage job offers, and HIV testing. RCTs thus influence the fate of many people around the world, raising a widely recognized concern with the randomness of RCT treatment assignment (see also the Significance Statement).

To alleviate the concern, this paper develops an experimental design that optimally incorporates subject welfare. I define welfare by two measures, (a) the predicted effect of each treatment on each subject's outcome and (b) each subject's preference or willingness-to-pay (WTP) for each treatment. My experimental design improves welfare compared to RCTs, while also always unbiasedly estimating treatment effects. The proposed design thereby expands prior pioneering designs that incorporate only parts of the welfare measures (3–9). This proposal also complements clinical-trial regulations that safeguard patients from excessive experimentation (10), as well as adaptive experimental designs to most statistically efficiently estimate treatment effects (11).

I start by defining experimental designs as procedures that determine each subject's treatment assignment probabilities based on data about the two welfare measures. In practice, the experimenter may estimate the welfare measures 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 experiment based on an imaginary centralized market and its competitive equilibrium (12, 13). 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.

This virtual-market construction gives EXAM nice welfare and incentive properties. EXAM is Pareto optimal, 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 for each treatment to be used by EXAM, every subject's optimal choice is to report her true WTP, at least for large experiments.

Importantly, EXAM also allows the experimenter to unbiasedly estimate the same treatment effects as standard RCTs do. Intuitively, this is because EXAM is an experiment stratified on observable predicted effects and WTP, in which the experimenter observes each subject's assignment probabilities (propensity scores). As a result, EXAM's treatment assignment is random (independent from anything else) conditional on the observables. The conditionally independent treatment assignment allows the experimenter to unbiasedly estimate

Significance Statement

RCTs determine the fate of numerous people, giving rise to a long-standing ethical 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)

The goal of this paper is to alleviate this dilemma. To do so, this paper proposes and empirically implements an experimental design that improves subjects' welfare while producing similar experimental information as typical RCTs.

[1]To whom correspondence should be addressed. E-mail: yusuke.narita@yale.edu

the average treatment effects conditional on observables. By integrating such conditional effects, EXAM can unbiasedly estimate the (unconditional) average treatment effect and other effects, as is the case with any stratified experiment.

Power is also a concern. I characterize the statistical efficiency in EXAM's average treatment effect estimation. In general, the standard error comparison of EXAM and a typical RCT is ambiguous, as is often the case with comparing RCT and stratified experiments. This motivates an empirical comparison of the two designs to confirm and quantify the power, unbiasedness, welfare, and incentive properties.

I apply EXAM to data from a water cleaning experiment in Kenya (1). 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.

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. Experimental design is a potentially life-saving application of economic market design (14, 15).

Experiment-as-Market (EXAM)

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 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'.
- 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. When multiple outcomes matter, e_{ti} can be set to the predicted effect on a known function of these outcomes.

I assume e_{ti} and w_{it} to be deterministic for simplicity. Without loss of generality, I normalize e_{ti} and w_{it} by assuming $e_{t_0 i} = w_{it_0} = 0$ for every subject i.

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 (RCT)*, which assigns each subject i to each treatment t with the impersonal treatment assignment probability p_t^{RCT}, 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 (see the Supplementary Information).

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

Definition 1 (*Experiment-as-Market* a.k.a. *EXAM*)**.** In the experimenter's computer, distribute any common artificial budget $b > 0$ to every subject. It is possible to let different subjects have different budgets. I make b the same for every subject, so that EXAM has the equality property that no subject strictly prefers anybody else's treatment assignment probabilities over her own. Find any price-discriminated competitive equilibrium, i.e., any treatment assignment probabilities (p_{it}^*) and their prices π_{te} with the following properties:

- 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
$$\pi_{te} = \alpha e + \beta_t.$$
This price is decreasing in treatment effect prediction e since I would like to more likely assign each treatment to subjects who are more likely to benefit from it.

- 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 $w_i \equiv (w_{it})_t$ and $e_i \equiv (e_{ti})$.

- 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.

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 $\bar{\epsilon} \equiv \min_t p_t^{RCT}$ is the largest possible value of ϵ. I define EXAM's treatment assignment probabilities as
$$p_{it}^*(\epsilon) \equiv (1-q) p_{it}^* + q p_t^{RCT},$$
where $q \equiv \inf\{q' \in [0,1] | (1-q') p_{it}^* + q' p_t^{RCT} \in [\epsilon, 1-\epsilon] \text{ for all } i \text{ and } t\}$.

The steps for implementing EXAM are as follows.

(1) Obtain predicted effects e_{ti} if possible and relevant.

(2) Obtain WTP w_{it} if possible and relevant.

(3) Apply Definition 1 of EXAM to the data from steps 1 and 2, producing assignment probabilities $p_{it}^*(\epsilon)$. α, β_t, and the resulting $p_{it}^*(\epsilon)$ are the equilibrium objects to be found by the experimenter's computer so as to satisfy the equilibrium constraints (see the Supplementary Information for an algorithm to do it).

Among these steps, subjects themselves only need to report WTP w_{it}. The remaining parts are run by the experimenter.

Welfare and Incentive

EXAM is an enrichment of RCT, as formalized below.

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*

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

However, in cases where the experimenter is concerned about WTP or predicted effects, EXAM differs from RCT and is welfare optimal.

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

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

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). 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 below.

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. Unfortunately, no experimental design satisfies the welfare property in Proposition 2 and exact incentive compatibility for general problems (12). This compels me to investigate approximate incentive compatibility in large experimental design problems. Only for this section, consider a sequence of experimental design problems $(i_1, ..., i_n, t_0, t_1, ..., t_m, (c^n_t))_{n \in \mathbb{N}}$ 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.

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 $p^{*n}_i(w_i, e_i, w_{-i}, e_{-i}; \epsilon^n)$ 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 the case where other subjects' WTP reports and predicted effects are random:

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

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 finite 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.

EXAM approximately incentivizes every subject to report her true WTP, at least for large enough experimental design problems.

Proposition 3 (Incentive)**.** *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*

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

The experimenter using EXAM can therefore ask subjects to report their true WTP without any deception. I also provide empirical support for incentive compatibility below.

Information

Despite the welfare merit, EXAM also lets the experimenter estimate treatment effects as unbiasedly as they would do in RCTs. To spell it out, here I switch back to any given finite problem with fixed WTP and predicted effects.

Suppose the experimenter is interested in the causal effect of each treatment on an outcome Y_i. 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 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, $\frac{\sum_{i=1}^n (Y_i(t) - Y_i(t_0))}{n}$. The experimenter estimates θ with an estimator $\hat{\theta}(Y, D)$, a function only of observed outcomes and treatment assignments. I say parameter θ is *unbiasedly estimable with experimental design* $p \equiv (p_{it})$ if there exists a "simple" estimator $\hat{\theta}(Y, D)$ (in the sense defined in the Supplementary Information) 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}).

EXAM turns out to be as informative as RCT in terms of the set of parameters unbiasedly estimable with each experimental design.

Proposition 4 (Unbiased Estimability)**.** *Under regularity conditions in the Supplementary Information, if parameter θ is unbiasedly estimable with RCT p^{RCT}_t, then θ is also unbiasedly estimable with EXAM $p^*_{it}(\epsilon)$ with any $\epsilon > 0$.*

Many key parameters, such as the average treatment effect and the treatment effect on the treated are known to be unbiasedly estimable with RCT and a simple estimator (see the Supplementary Information). Proposition 4 implies that these parameters are also unbiasedly estimable with EXAM.

Corollary 1. *The average treatment effect and the treatment effect on the treated are unbiasedly estimable with EXAM.*

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? It is because once it is constructed, EXAM is a particular stratified experiment stratified on observable WTP and predicted effects. 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:

$$(Y_i(t))_t \perp\!\!\!\perp D_i | (e_i, w_i). \quad [1]$$

Conditional independence (1) implies that the same conditional independence holds conditional on the assignment probability $p_i^*(\epsilon) \equiv (p_{it}^*(\epsilon))_t$ ((16) section 12.3), which is again known to the econometrician:

$$(Y_i(t))_t \perp\!\!\!\perp D_i | p_i^*(\epsilon). \quad [2]$$

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 $CATE_{pt}$.

By summing up such 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 $\delta_p \equiv \sum_{i=1}^n 1\{p_i^*(\epsilon) = p\}/n$, I use $CATE_{pt}$'s to get ATE as follows: $\sum_p \delta_p CATE_{pt} = ATE_t$.

The above estimability argument motivates a strategy to estimate ATE with EXAM's data. As a warm-up, fo[illegible]n $\{i|p_i^*(\epsilon) = p\}$, 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 (2), [illegible]s 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 estimator $\hat{\beta}_t^*$ unbiasedly estimates the average treatment effect with its variance in an analytical form, as shown in the Supplementary Information. Alternatively, empirical researchers may prefer a single regression:

$$Y_i = a + \sum_{t=t_1}^{t_m} b_t D_{it} + \sum_{t=t_1}^{t_m} c_t p_{it}^*(\epsilon) + e_i, \quad [3]$$

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

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

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

Empirical Application

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). 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. Yet there is controversy about its health impacts. Experts argue that improving source water quality may only have limited effects, since, for example, water is likely recontaminated in transport and storage.

This controversy motivated researchers to analyze randomized spring protection in Kenya (1). 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 [illegible]erimental subjects. The researchers found that diarrhea among children in treatment households fell by about a quarter of the baseline level. I call this real experiment the "original experiment" and distinguish it from EXAM and RCT as formal concepts in my model.

I consolidate the original experimental data and my methodological framework to empirically evaluate EXAM. Applying the language and notation of my model, experimental subjects are households in the original experiment's sample. 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 household i's WTP for better water access t_1 is denoted by w_{it_1}, which I estimate in the Supplementary Information. I also estimate the heterogeneous treatment effect $e_{t_1 i}$ of spring protection t_1 on household i's child diarrhea outcome.

Given the estimates, imagine somebody is planning a new experiment to further investigate the same spring protection treatment. What experimental design should she use? Specifically, which is better between RCT and EXAM? 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. I fix the set of subjects and treatments as in the original experiment. That is, there are 1540 households as subjects to be assigned either to the 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 the original experiment. I set the bound parameter ϵ to be 0.2. I fix predicted effects $e_{t_1 i}$ to their point estimate $\hat{e}_{t_1 i}$.

I start with evaluating EXAM's welfare performance. I 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 1. The mean of average WTP w_i^* for assigned treatments is about 89% 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 in child diarrhea or 42% reduction relative to RCT's level. This predicted effect benefit amounts to about 17% of the average treatment effect of the spring protection found by the original experiment.

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 treatment effect model I estimate in the Supplementary Information. 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). I also implement the other propensity-score-weighting estimator $\hat{\beta}^*$. The procedure for RCT is analogous except that the treatment assignment probability is fixed at p_t^{RCT}.

Program evaluation with EXAM is as unbiased and precise as that with RCT. Figures 2 Panel (a) and S2 plot the distribution of the resulting treatment effect estimates $\hat{b}^*$ and $\hat{\beta}^*$

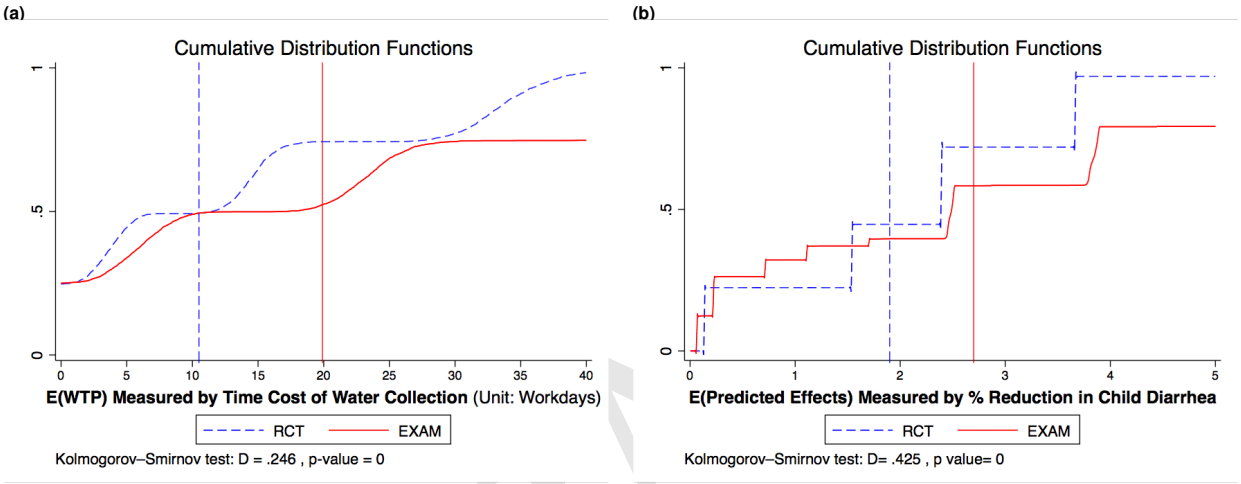

Fig. 1. EXAM vs RCT: Welfare. This figure shows the distribution of average subject welfare over 1000 bootstrap simulations under each experimental design. Panel (a) shows the average WTP for assigned treatments w_i^* and panel (b) shows the average predicted effects of assigned treatments e_i^*. A dotted 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^*.

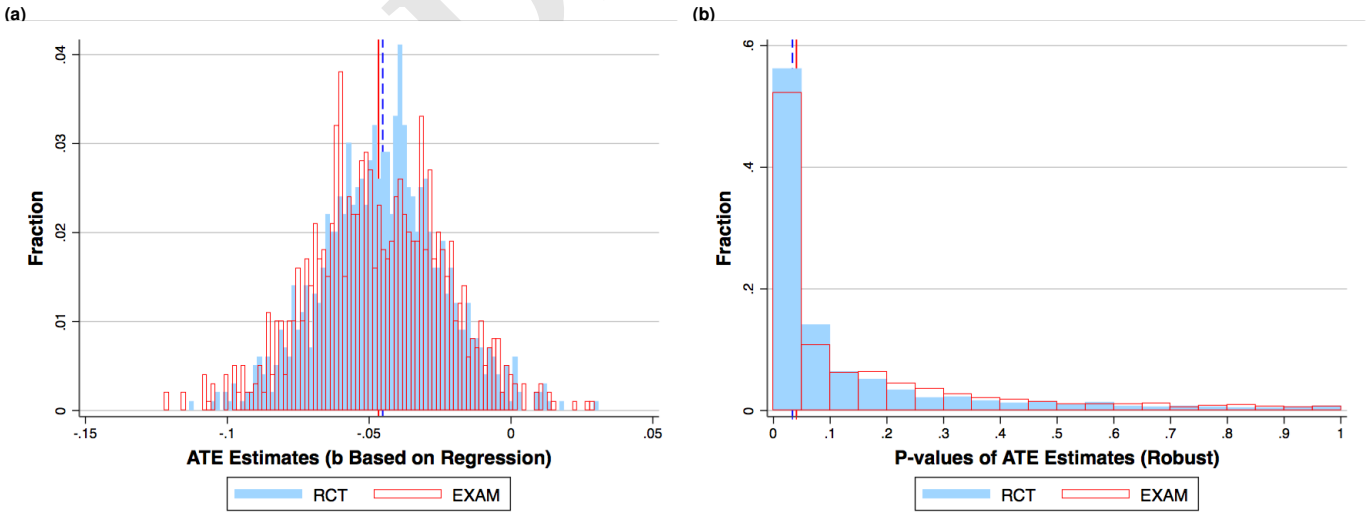

Fig. 2. EXAM vs RCT: Average Treatment Effect Estimates. This figure compares EXAM and RCT's causal inference performance by showing the distribution of average treatment effect estimates under each experimental design. Panel (a) shows the distribution of treatment effect estimates $\hat{b}^*$, and panel (b) shows robust p values for $\hat{b}^*$. Blue bins indicate average treatment effect estimates for RCT while transparent bins with red outlines indicate those for EXAM. The solid vertical line indicates the mean for EXAM while the dashed vertical line indicates that for RCT. Mean is represented by a solid line while median by a dashed line. The p values are based on robust standard errors. Blue bins indicate p values for RCT while transparent bins with red outlines indicate those for EXAM. The solid vertical line indicates median for EXAM while the dashed vertical line indicates that for RCT.

over simulations. In line with Propositions 4 and 5 (in the Supplementary Information), the means of $\hat{b}^*$ and $\hat{\beta}^*$ for EXAM are indistinguishable from those under RCT. Both designs successfully recover (1)'s average treatment effect estimate (4.5% reduction in diarrhea; replicated in Table S5).

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 Panel (B) of Figure 2, which shows the distribution of the robust p values for the estimates $\hat{b}^*$. Figure S3 additionally shows p values based on exact, non-robust, and finite population causal standard errors, where the exact standard error means the standard deviation in the distribution of $\hat{b}^*$ in Figure 2. 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.

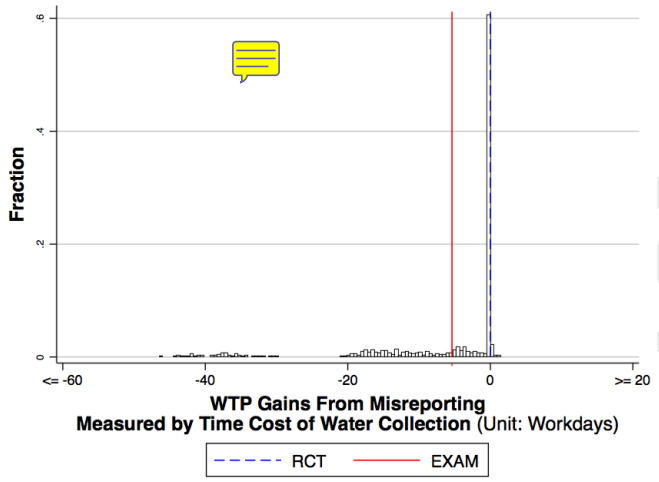

Fig. 3. EXAM vs RCT: Incentive (WTP manipulation $\sim$ true WTP$+N(0, 100)$). This figure shows the histogram of true WTP gains from potential WTP misreporting to EXAM, quantifying the incentive compatibility of EXAM. Each solid vertical line represents the mean WTP gain from potential WTP misreporting to EXAM. The dash vertical line is for RCT, where the true WTP gain from any WTP misreport is zero.

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. The manipulation w'_{jt_1} is from $N(w_{jt_1}, 100)$ where w_{jt_1} is j's true WTP. The Supplementary Information report similar results under different scenarios, covering different types of misreporting, that is, both over- 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 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. Figures 3 and S4 show this by drawing the distribution of Δw over simulations and households. Across all scenarios, the WTP gain Δw from misreporting is mostly negative and well below zero on average. Overall, in this empirical setting, EXAM provides subjects with stronger average incentives for truthful WTP reporting than RCT does (because ~~in~~ RCTs are indifferent among all possible WTP reports). EXAM may therefore be better at eliciting reliable WTP data.

Conclusion

Motivated by the high-stakes nature of many RCTs, I propose a data-driven experiment dubbed Experiment-as-Market (EXAM). EXAM is a particular stratified experiment derived from 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 2-4). 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.

1. M Kremer, J Leino, E Miguel, AP Zwane, Spring cleaning: Rural water impacts, valuation, and property rights institutions. *Q. J. Econ.* **126**, 145–205 (2011).
2. R Stupp, 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 Oncol.* **10**, 459–466 (2009).
3. M Zelen, A new design for randomized clinical trials. *New Engl. J. Medicine* **300**, 1242–1245 (1979).
4. J Angrist, G Imbens, Sources of identifying information in evaluation models. Working Paper (1991).
5. S Chassang, GP i Miquel, E Snowberg, Selective trials: A principal-agent approach to randomized controlled experiments. *Am. Econ. Rev.* **102**, 1279–1309 (2012).
6. M Zelen, Play the winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.* **64**, 131–146 (1969).
7. L Wei, S Durham, The randomized play-the-winner rule in medical trials. *J. Am. Stat. Assoc.* **73**, 840–843 (1978).
8. F Hu, WF Rosenberger, Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *J. Am. Stat. Assoc.* **98**, 671–678 (2003).
9. M Kasy, A Sautmann, Adaptive treatment assignment in experiments for policy choice. Working Paper (2019).
10. LM Friedman, C Furberg, DL DeMets, DM Reboussin, CB Granger, *Fundamentals of Clinical Trials.* (Springer) Vol. 3, (1998).
11. J Hahn, K Hirano, D Karlan, Adaptive experimental design using the propensity score. *J. Bus. Econ. Stat.* **29**, 96–108 (2011).
12. A Hylland, RJ Zeckhauser, The efficient allocation of individuals to positions. *J. Polit. Econ.* **87(2)**, 293–314 (1979).
13. E Budish, YK Che, F Kojima, P Milgrom, Designing random allocation mechanisms: Theory and applications. *Am. Econ. Rev.* **103**, 585–623 (2013).
14. PR Milgrom, *Putting Auction Theory to Work.* (Cambridge University Press, Cambridge), (2004).
15. AE Roth, *Who Gets What and Why: The New Economics of Matchmaking and Market Design.* (Houghton Mifflin Harcourt), (2015).
16. GW Imbens, DB Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences.* (Cambridge University Press), (2015).