Graduate School of Economics, Hitotsubashi University Discussion Paper No.2020-07

Ethics of randomized field experiments: Evidence from a randomized survey experiment*

Hide-Fumi Yokoo†

November 4, 2020

Abstract

To conduct randomized field experiments while easing the disutility of subjects and the concerns of practitioners, I empirically study the ethical concerns held by potential subjects. Two types of online surveys are implemented, targeting approximately 2,000 respondents each. In the first survey, respondents are asked whether they recognize ethical issues in six existing experiments conducted by economists. Among these six experiments, an early childhood intervention is recognized as the most acceptable, while a charitable fund-raising experiment using lotteries is recognized as the least acceptable from an ethical perspective. To investigate methods to ease such ethical concerns, I conduct the second survey in which respondents are randomly assigned to four groups and shown different descriptions of the studies, which adopt different research designs. From this randomized survey, I find a nonsignificant impact of changing the research methodology from a randomized field experiment to an uncontrolled before–after study. Changing the topic of the study from charitable giving to other behaviors decreases respondents' unethical feelings. However, ethical concerns significantly increase when informed consent is not enough or when subjects are randomly sampled. These findings support a randomized experiment with agreed-upon participants, although it may limit the external validity of the experiment.

Keywords: Ethical issues, Field experiments, Online surveys, Randomized controlled trials
JEL classification: C93, D63, O22

* I thank Keitaro Aoyagi, Yuki Higuchi, Yohei Kobayashi, Koichi Kuriyama, Aya Suzuki, Kenji Takeuchi, Fumio Ohtake, Junichi Yamasaki, and participants at a seminar at the Research Institute of Economy, Trade and Industry (RIETI), Hitotsubashi, as well as conference participants at SEEPS 2017 and JEA (Spring) 2019 for their helpful comments. I also thank Aya Tamori (INTAGE Inc.), Kimiko Kaneyama, Hidemi Motojima, and Nagisa Yoshioka for their excellent assistance with the surveys and Kazuki Motohashi, Tomohiro Suzuki, Ikuya Yamada, and Kotoko Yanagisawa for their excellent research assistance. This study was conducted as part of the project "Promoting evidence-based policy in Japan," undertaken at RIETI. This study was also supported by Hitotsubashi University and JSPS KAKENHI, grant number 17K18547.

† Graduate School of Economics, Hitotsubashi University. E-mail: hidefumi.yokoo@r.hit-u.ac.jp

1 Introduction

Ethical issues often arise when we run randomized field experiments (Glennerster, 2017; Groves Williams, 2016; Ravallion, 2009). One reason behind these issues is that economists sometimes do not inform the research subjects that they are in an experiment (Levitt and List, 2009). Economists are most likely to acquire informed consent for data collection from the subjects but less likely to explain the experimental design to the subjects (Glennerster and Powers, 2016; Teele, 2014). This is quite uncommon or not acceptable for randomized controlled trials (RCTs) in medicine.[1] Possibly due to this practice in economics, implementing partners (e.g., governments and NGOs) raise ethical and reputational concerns about running randomized evaluations and sometimes hesitate to conduct them. Since randomized field experiments can benefit society by their ability to cleanly identify causal impact, excessive concerns would constitute a barrier to making effective policies.

To rigorously evaluate policies while easing the disutility of subjects and the concerns of practitioners, I empirically study the ethical concerns held by potential subjects regarding randomized field experiments. I conduct a series of online surveys on ethical concerns for existing randomized field experiments in the field of economics. I select the following six studies using the criteria described in Section 3: Allcott (2011); Fryer et al. (2015); Hanna et al. (2016); Hosono and Aoyagi (2018); Landry et al. (2006); and Thornton (2008).

In my first survey, comprising approximately 2,000 respondents in Japan, I provided brief explanations on the studies and asked if respondents recognized any ethical issues with them. In my second survey, I focused on two studies among six—the most and least concerning studies from the first survey—and explored a method to ease such ethical concerns by modifying each study. To do so, I applied a randomized online survey experiment (see Cruces et al., 2013; Kuziemko et al., 2015) in which approximately 2,000 respondents were

[1] Following Favereau (2016), I use the term *randomized field experiment* to refer to an experimental design centered on a random assignment of treatments in the field of economics, while the term *randomized controlled trials (RCTs)* is used to refer to that in medicine throughout this paper.

randomly assigned into three treatment groups and a control group and shown different descriptions. This survey design allows me to estimate the causal impact of changing an attribute of the study—for example, the treatment (from economic incentive to information provision) or research design (from a randomized experiment to a before–after study of an intervention without a control group)—on ethical concerns.

The previous studies on the ethics of randomized field experiments are controversial. For example, both Glennerster and Powers (2016) and Teele (2014) provide a framework for thinking about the ethics of randomized evaluations; however, their arguments are different.[2] Teele (2014) concludes that randomized field experiments differ fundamentally from laboratory experiments or observational studies and require informed consent, the full assessment of the risk of the experiment, and nonexploitative participant selection procedures as minimal steps. Conversely, considering the implementation of programs and other methodologies (e.g., quasi-experimental approaches) as the counterfactual, Glennerster and Powers (2016) conclude that while there are ethical issues specific to randomized evaluations, most of them are not unique to this methodology. Relatedly, List (2008) discusses informed consent associated with natural field experiments and argues that the lack of informed consent seems defensible when the research makes participants better off, benefits society, and confers anonymity and just treatment to all subjects.[3] Note that, these studies discuss the ethics of randomized field experiments from a normative point of view.

While normative analyses on economic methodologies are absolutely important, such debates tend to result in two extreme opinions. Unlike the above studies, which conceptually examine the ethics of the experiments, I conduct an empirical study. Using online surveys, I explore what kind of randomized field experiments are considered by laypeople as involv-

[2] Both papers examine various ethical questions that arise during the conduct of randomized evaluations based on three principles set out in the Belmont Report (i.e., respect for persons, beneficence, and justice). In 1974, the United States put into place a framework for medical and nonmedical research involving human subjects. The Belmont Principles are the ethical guidelines produced through this framework (Glennerster and Powers, 2016).

[3] Relatedly, O'Flynn et al. (2016) discuss the definition of ethics in the context of randomized evaluations. Groves Williams (2016) discuss ethics in international development evaluation in general.

ing ethical issues and how researchers can alleviate these concerns by modifying their own research plans. From the results of the positive analyses, I present evidence that contributes to the normative analyses of field experiments. Note that the objective of this paper is to improve the methods of economic studies but not ethically criticize individual papers.

In the first survey, respondents are shown a description of the study and asked the following question: "Do you recognize any ethical issues in this study?" Then, they indicate their concern on a five-point scale. From the results, I find that respondents' concerns vary among experiments. Relatively few respondents (24%) believe that there is an ethical issue involved in a description that summarizes the work of Fryer et al. (2015), who study the effects of a preschool using the Chicago Heights Early Childhood Center (CHECC) project. In contrast, more than 45% of the respondents recognize that there is an issue in a description that summarizes the work of Landry et al. (2006), who study the impact of a lottery incentive on charitable giving. These results suggest that randomized field experiments are ethically evaluated according to their context, outcomes, treatments, and design.

In the first half of the second survey, which focuses on Fryer et al. (2015), respondents are randomly assigned one out of four slightly different descriptions, which present slightly different experimental designs. Then, they are asked the same question as that in the first survey. I obtain several findings from this survey experiment. First, the response to "There is an ethical issue" significantly increases if parental consent is absent. Second, ethical concern increases if participants are selected at random rather than through self-selection. This result implies that randomization within the self-selected subject pool is more acceptable. These results mean tradeoffs among the Hawthorne effect, specific sample problems (Peters et al., 2016), and ethical issues.

In the second half of the second survey, which focuses on Landry et al. (2006), I find a statistically insignificant impact of changing the research methodology from a randomized field experiment to a before–after study of an intervention without a control group. This result suggests that the internal validity of the analysis can be obtained without increasing ethical issues. Conversely, if the outcome variable of the study is changed from charitable giving to garbage sorting, which both can be considered voluntary public goods provision, then ethical concern significantly decreases. These results imply that as long as the purpose of the research is not the evaluation of a specific program but rather the testing of a specific theory, then it is possible to alleviate the associated concerns by modifying the research topics.

This paper contributes to several strands of the literature. First, this paper relates to the abovementioned debates on the ethical concerns of randomized field experiments (e.g., Glennerster and Powers, 2016; Groves Williams, 2016; List, 2008; Teele, 2014). Unlike these normative analyses, two recent papers have reported the results of surveys on the perceptions of randomized field experiments. Meyer et al. (2019) and Mislavsky et al. (2019) conduct surveys asking about the appropriateness and acceptability, respectively, of the hypothetical scenarios of field experiments. The results of the above two papers are, at first glance, contradictory; Meyer et al. (2019) find that respondents are averse to being involved in experiments, while Mislavsky et al. (2019) find that respondents equally accept an experiment to test a policy and a universal implementation of the same policy. This present paper shares the motivation of the study with the above two papers and presents similar surveys. In addition to the survey of the acceptability of experiments, this present paper investigates the causal mechanisms of ethical concerns and methods to alleviate them to improve economic studies. In Section 5, I explain the above two papers in more detail and discuss how their conclusions are complemented by the insights of the present paper.

Second, this paper contributes to the nascent literature on the design of experiments that decrease ethical concerns. Duflo et al. (2007) recommend encouragement designs when evaluating programs over which randomizations of the treatment itself are not feasible for ethical reasons. Angrist and Imbens (1991) present an experimental design in which an eligible population is randomly selected, but eligible individuals are allowed to freely choose whether to participate in the program. Narita (2020) develops another experimental design in which subjects with an imaginary budget and personalized clearing price purchase treatment assignment probabilities. While these studies focus on subjects' preferences for *treatments* and propose methods to address ethical issues, this present paper further considers preferences for *studies* including research methodologies, topics, sampling methods, and informed consent. As a result, this paper contributes to the literature by proposing practical methods to address these concerns.

Third, this paper is part of a small but growing set of papers using randomized online survey experiments to study beliefs and preferences. While previous studies investigate the nature of preferences for policies, such as redistribution (e.g., Cruces et al., 2013; Kuziemko et al., 2015) or legalizing payments to kidney donors (Elías et al., 2019), this paper investigates the nature of preferences for economic studies by using the same methodology, a randomized survey experiment.

The paper proceeds as follows. Section 2 presents the motivation behind the surveys conducted in Japan. Section 3 presents the descriptions of the six experiments examined in the first survey and presents the results. Section 4 explains the design of the second survey, describes the data, and presents the main results. Section 5 presents the results of an additional analysis on the effect of implementers, discusses the implications of the findings, explains how they fill the gap between the arguments of Meyer et al. (2019) and those of Mislavsky et al. (2019), and discusses the limitations of the present study. Section 6 concludes the paper. All my online surveys, data, and programs are available in the Online Appendix.

2 Randomized field experiments and preferences in Japan

In the last ten years, there has been a rise in the use of randomized field experiments by the Japanese government. Examples can be found in several policy areas. First, the Japan International Cooperation Agency (JICA) has started to run randomized evaluations in developing countries. The JICA, mostly in collaboration with economists, has evaluated its programs in various countries, such as Burkina Faso, Cote d'Ivoire, Indonesia, Mongolia, Mozambique, and Niger.[4]

Second, a series of field experiments were conducted to curb electricity use in Japan.Ito et al. (2018) study one of those experiments that compared the impact of moral suasion and critical peak pricing on electricity demand in Kyoto Province in 2012.[5] The program was designed and jointly implemented by the authors in collaboration with the Ministry of Economy, Trade, and Industry of Japan (METI), a local government, and several private companies.[6] Subsequently, in 2015, the METI evaluated OPOWER's Home Energy Report (HER) in a northern province in Japan by referring to Allcott (2011).[7]

Third, in 2017, the Japanese government launched a so-called nudge unit, which runs several randomized field experiments (Behavioral Sciences Team, 2019).[8]

All of these movements of evidence-based policy making have brought about an increase in discussions about ethical issues, which arise when we run randomized field experiments among policy makers and researchers (see, for example, Behavioral Sciences Team, 2019). This discussion motivated me to conduct the present study in Japan.

[4] One of the early randomized intervention was started in 2010 in Burkina Faso which evaluated the effects of a school-based management (SBM) program (Sawada et al., 2019). Kozuka (2018) also evaluates the SBM program in Niger. Takahashi et al. (2019) evaluate agricultural training in Cote d'Ivoire which is provided in the project by JICA. Tanaka et al. (2018) evaluate the effects of showing leaflets to encourage participation in the public pension system by self-employed workers in Mongolia.

[5] Throughout this paper, I use the term *province* to indicate regions and local governments in Japan although the actual administrative term is *prefecture*. Yamaguchi et al. (2018) argue that *province* is more intuitive for most readers.

[6] Relatedly, Matsukawa (2018) evaluates another treatment tested in the same project which was the provision of in-home displays to show households' half-hourly electricity consumption in real time.

[7] Jyukankyo Research Institute Inc. (2016) reports the result of this randomized evaluation.

[8] The Nudge Unit Japan is named the Behavioral Sciences Team (BEST).

3 Survey on six experiments

3.1 Survey data collection

The first survey was designed by the author and implemented in Japan by the survey company INTAGE Research Inc. in March 2017.[9] This company maintains a panel of respondents and undertakes online surveys. In my study, potential respondents in the panel were randomly selected with weights to create a representative Japanese sample in terms of residential area, gender, and age group. Those who joined the survey were paid if they fully completed the survey, although the author did not share information about the exact pay for this survey.[10]

The survey request was sent by email to randomly chosen candidates. I requested the company to implement a sample size of 2,000. In response to this request, the company sent invitation emails to 6,698 candidates. The email did not mention the details of the survey but requested that respondents "please participate in a survey about everyday life." Those who decided to participate were accepted until the number of respondents reached a set number (not known by the author). As a result of this procedure, the sample size for the first survey is 2,107.[11] Prior to the survey, respondents were told that their responses would be used by research institutions, local governments, companies, etc., and they gave their consent.

[9] INTAGE Holdings which includes INTAGE Research Inc., founded in 1960 in Japan, is ranked 9th in the 2017 American Marketing Association (AMA) Gold Global Top 25 Market Research Firms.

[10] Individuals in panel are called *cue monitors*. The number of cue monitors reside in Japan as of January 2017 was 1.41 million. In general, they are paid in "cue monitor points" for participating online surveys. Those points can be exchanged for Amazon gift cards, electronic money, or vouchers at a rate of JPY 1.00 for one point. The points per survey depend on the survey.

[11] The survey requests were sent from 2 p.m. on Friday (March 10, 2017) to 6,698 candidates. From the set number of respondents, only valid respondents based on INTAGE Research's determination standard were left, resulting in 2,107 respondents, which were 31.5% of the number of candidates who received the request. The first survey finished at 9 a.m. on March 13, 2017.

3.2 The six randomized field experiments used in the first survey

[Table 1]

For the first survey, I selected six economic studies based on the following two criteria: whether the experiments seem to involve ethically sensitive issues and their relevance to current policy discussions in Japan. The selected experimental studies are shown and summarized in Table 1. Half of the selected studies relate to human capital issues: health status, disease testing, or preschools. This reflects that in general, people care more about topics involving human life and death and childhood circumstances. Three experiments are conducted in developed countries, while the other three are conducted in developing countries. This reflects a balance between the increased usage of randomized field experiments in developing countries and the focus of the present study being ethical concerns recognized in a developed country. Finally, a project conducted by the JICA is included as an example for the experiment conducted by Japanese organizations.

In principle, I attempted to introduce six experiments and to summarize the experiments described in original articles as accurately as possible. However, I made several modifications to the original experimental designs, which are mentioned below. Most of the modifications were made to simplify the descriptions to make them easy for respondents to understand. In the descriptions, I kept the authors of the six papers anonymous. Throughout the surveys in the present study, I avoided using the word "experiment," and instead, I used "study" and "project." The Japanese version of the six descriptions shown in this section and the Appendix was used. Respondents were shown three randomly assigned descriptions of studies in random order and answered questions for each.[12] The selected six studies are as follows.

Study on a preschool: Fryer et al. (2015)

[12] To keep the time for reading the survey materials and responding to questions short, I provided three randomly assigned descriptions, instead of all the six descriptions, per respondent. Note that three additional descriptions of another study (not shown in this paper) are also shown to each respondent. Furthermore, the respondents were asked two questions for each description. In this paper, survey responses to only one of the two questions is used. In total, respondents were shown six descriptions and answered 12 questions for each. See Section 3.3 for the average duration of the survey.

The first study I chose is the CHECC project. This project conducts randomized field experiments to evaluate early childhood education interventions. For example, a child and their parents are randomized into one of three groups—preschool treatment, parent academy treatment, and control (Cappelen et al., 2020)[13]—where the preschool used is established for the purpose of this experiment (see Gneezy and List, 2013).[14] Various outcomes are examined, such as cognitive and noncognitive test scores (Fryer et al., 2015), time preferences (Andreoni et al., 2019), risk preferences (Andreoni et al., 2020), and social preferences (Cappelen et al., 2020). I summarized the project and prepared a description of it with several simplifications. Specifically, respondents are shown the following:

Study on a preschool

Recent findings show that the care and education one receives in early childhood affect one's academic achievement and lifetime earnings in adulthood. Following these findings, Professor X established a preschool in a low-income area.

Overview of the preschool:

- *The preschool is free of charge.*
- *This preschool uses a curriculum called "Tools of the Mind" to foster patience and social skills.*
- *Inside the preschool is similar to a small "town," where one can experience various types of jobs.*
- *Children of this preschool are surveyed periodically.*
- *Followup surveys are planned for every few years following graduation.*

Professor X called for applicants to this preschool.

Overview of admissions:

- *Parents and children, for a total of 140 families, applied for admission.*
- *Only 70 children selected based on a lottery were admitted.*
- *The remaining 70 children were not able to enroll in the preschool.*

[13] Fryer et al. (2013) provide an outline of the project, especially in the early stage.

[14] The work of Gneezy and List (2013) was translated into Japanese and published in 2014.

- *However, the children who were not able to enroll, as well as their parents, are regularly invited to parties held on holidays.*

After the preschool was opened, Professor X invited the children who were enrolled and their parents—as well as the children who were not able to enroll and their parents—to regularly held parties and surveyed them. The surveys were periodically conducted for over 10 years, even after the children entered primary school. Finally, Professor X conducted a study comparing children who attended the preschool with those who were not able to enroll. Note that the parents of the 140 children who became subjects of the study received an explanation regarding them being the subjects of the study, and they gave their consent.

A screenshot of the survey is presented in the Online Appendix. Respondents were then asked two questions. In this study, I focus on the first question: Do you recognize any ethical issues in this study? Respondents chose one of five options (from "There is a major ethical issue" to "There is no ethical issue at all"). Note that, in the introduction of the above description, I focused on academic achievement and income, meaning that Fryer et al. (2015) study is the closest among the existing papers produced from the CHECC project. For this reason, to intuitively label the above description, I refer to it as "Fryer, Levitt, and List (2015)."[15]

Two remarks on the modifications made to the original experiments should be mentioned. First, while the project includes two treatments, I focused on the preschool and made the number of groups two to make the survey simple and short.[16] Second, I made the group size 70 following information provided in Gneezy and List (2013). This figure is similar to the size of the analytical sample of Fryer et al. (2015) and Cappelen et al. (2020). For an implementer of the CHECC project, I anonymized and framed it as "Professor X."

Study on HIV testing: Thornton (2008)

[15] Note that, however, Fryer et al. (2015) focus on the parent academy treatment instead of the preschool treatment.

[16] According to Gneezy and List (2013) and Cappelen et al. (2020), children in the preschool treatment are further randomized to either the *Literacy Express* curriculum or to the *Tools of the Mind* curriculum. Again, for simplification, I focused on only one of them (the latter).

The second study I chose is the study on HIV testing for AIDS prevention in the developing world. Thornton (2008) analyzes the dataset collected in the experiment, which randomly assigned monetary incentives to learn the results of HIV testing. The sample of the study consists of 2,812 individuals in rural Malawi who accepted an HIV test and the followup survey. Thornton (2008) evaluates the impact of incentives on the demand for learning HIV status and subsequent behaviors.[17]

In the present study, I focus on the behavior of learning HIV status to simplify the description. Note that, however, Thornton (2008) also studies other behaviors, such as the purchase of condoms.[18] In addition, while there were more than two variations in incentives provided in Thornton (2008), I reduce these variations to two groups—a control and a treatment—for simplification.[19] For this and the four studies that follow, the description used in the survey is attached in Appendix A.

Study on charitable giving: Landry et al. (2006)

The third study I chose is on voluntary contributions to public goods. Landry et al. (2006) conducted a randomized field experiment to study the impact of lotteries on charitable giving. They conducted door-to-door fundraising in North Carolina, where 44 solicitors approached 4,833 households. For households in one among four randomly assigned groups, the single-prize lottery treatment was offered; donors were provided a ticket for a raffle where the winner would receive a USD 1,000 prepaid credit card.[20]

In the present study, I chose two groups in the original experiment (a voluntary contri-

[17] Thornton (2012) and Godlonton and Thornton (2013) use the dataset collected through the same project (the Malawi Diffusion and Ideational Change Project). Dupas (2011), Delavande and Kohler (2016), and Godlonton et al. (2016) also experimentally study the beliefs and behaviors related to the risk of HIV infection.

[18] In addition, the experiment of Thornton (2008) created random variation in the distance to the HIV results center. This design is also abstracted from the description used in the present study.

[19] In Thornton (2008), subjects were given randomly assigned vouchers between zero and three dollars, redeemable upon obtaining their test results at a nearby center.

[20] Following this study, Landry et al. (2010) conducted another experiment to examine the dynamics of charitable fundraising. Various other studies conducted randomized field experiment using door-to-door fundraising, for example, Soetevent (2011), DellaVigna et al. (2012), and Edwards and List (2014). Other studies that used randomized field experiment to study charitable giving include List and Lucking-Reiley (2002), Frey and Meier (2004), and Shang and Croson (2009).

butions mechanism without seed money and the single-prize lottery) to simplify the description.[21] For the objective of the experiment, I described it as "*to obtain more donations.*" Note that in contrast to the previous two studies, the subjects of Landry et al. (2006) are not informed that such solicitation is part of a research project; thus, it is considered a natural field experiment in the parlance of Harrison and List (2004). I explicitly mentioned this feature in the description as follows: *Note that the 4,800 households that were solicited for donations were not informed of their involvement in the study.*

Study on electricity conservation: Allcott (2011)

The fourth study I chose is on the nudge to encourage electricity conservation. Allcott (2011) evaluate the program that sent Home Energy Report (HER) letters to households. The HER consists of two components: the social comparison module, which compares households' electricity use to that of their neighbors, and the action steps module, which includes energy conservation tips.[22]

Allcott (2011) pools observations from 17 experiments that include approximately 600,000 households in total. However, in the present study, I described the sample size as "*40,000 households*," which is approximately equivalent to the average sample size of the 17 experiments. This size of the experiment is to some extent comparable with the HER experiment conducted in Japan in 2015 (Jyukankyo Research Institute Inc., 2016).

Study on household air pollution from cooking: Hanna et al. (2016)

The fifth study I chose is the evaluation of a program to reduce household air pollution in a developing country. Hanna et al. (2016) evaluate a program implemented in India, where improved cooking stoves are distributed almost for free.[23] While the original study collected data on approximately 2,500 households and randomly divided them into three groups using a public lottery, I simplified my description to two groups of 1,600 households.

[21] The other two treatments are a voluntary contributions mechanism with seed money and the multiple-prize lottery.

[22] Other papers that experimentally evaluate the HER include Ayres et al. (2012), Costa and Kahn (2013), Allcott and Rogers (2014), and Allcott and Kessler (2019).

[23] Other papers that study the impact of improved cooking stoves include Mobarak et al. (2012), Bensch and Peters (2015), Miller and Mobarak (2015), Levine et al. (2018) and Jeuland et al. (2020).

Note that unlike the other five experiments, this program was designed to rollout the treatment, meaning that households in the control group also received stove construction afterward. According to Duflo et al. (2007), such an experimental design, which randomizes the order of phase-in, is considered the fairest way to implement programs. Thus, I mentioned this feature in the description as follows: *Professor X carried out a project in a developing country, whereby "improved cooking stoves" were constructed free of charge in an area with 1,600 households. Stove construction was carried out over two periods over 5 years. For the first three years, stoves were built for only 800 households selected by a lottery. The remaining 800 households waited for their turn.*

Study on recyclable waste sorting: Hosono and Aoyagi (2018)

The last study I chose is an experiment implemented by a Japanese organization. Hosono and Aoyagi (2018) analyze a dataset collected in a project conducted by the JICA in Mozambique. In the project, the JICA attempted to encourage household waste-sorting behavior.[24] A total of 1,000 households in a suburb of Maputo are randomly assigned to one of the four groups. Three treatments are evaluated to encourage the sorting of recyclable waste (e.g., plastics and aluminum) from other garbage: free distribution of buckets to store recyclables, face-to-face persuasive communication, and in-kind incentives.

In the present study, I focus on in-kind incentive treatment and control groups. Households in the treatment group can obtain a stamp on their card if they dispose of recyclable waste separately from other garbage, and they can obtain laundry detergent if they collect ten stamps.

Three descriptions from the above six are shown to 2,107 respondents. As the first description is labeled "Fryer et al. (2015)," five other descriptions are also labeled by the

[24] Murase et al. (2017) and Yokoo and Harada (2020) also analyze a dataset collected through solid waste management programs by the JICA, which involves randomized evaluation. Chong et al. (2015) evaluate recycling campaigns conducted by an NGO in Peru by using randomized field experiments. Other papers experimentally study interventions to encourage household recycling in developed countries include, for example, Hopper and Nielsen (1991); Schultz (1999); Koford et al. (2012).

representative papers. Moreover, 75.0% of the descriptions used in the first survey mention the implementer of the program as "Professor X," but the rest of them purposely mention a different implementer. I intentionally and randomly made this difference to examine another research question examined later. Throughout Section 3, I focus on the comparison of six studies and leave the discussion on the impact of the implementer to Section 5. In the regression analysis in this section, I control for this randomness to focus on the comparison of six studies, holding the difference in implementers constant.

3.3 Data and summary statistics

[Table 2]

Table 2 shows the characteristics of the sample that completed the first survey. On average, 48% of the respondents are women, 61% are married, 38% live with children, and their average age is 46.7. In addition, I collected information on the time spent on the survey. The median time is 3.4 minutes, and the average is 23 minutes.[25] Compared to another survey conducted in Japan, that of Hanaoka et al. (2018), my sample is younger, has lower income, and has fewer employed.

3.4 Descriptive results

[Figure 1]

Figure 1 shows the distribution of the responses. Panel A shows the result for Fryer et al. (2015), where approximately 32% of the respondents recognize that there is no ethical issue (Unethical Rating 1 and 2), 44% feel neutral (Unethical Rating 3), and 24% recognize that there is an issue (Unethical Rating 4 and 5). A similar but slightly worse result is obtained for Panel D of Allcott (2011), where approximately 29% recognize that there is no ethical

[25] Figure A1 in the Online Appendix shows a histogram of the time spent on the survey. Table A1 in the Online Appendix reports the result of the regression analysis on the characteristics and time spent on the survey. The time is significantly longer if respondents live with children or if they are part-time employees.

issue, while 27% recognize that there is an issue. For Thornton (2008), the result shows that approximately 24% recognize that there is no ethical issue, while 32% recognize that there is an issue, which is quite similar to the results for Hanna et al. (2016) and Hosono and Aoyagi (2018). The study that is recognized as the most unethical is Landry et al. (2006), where approximately 13% of respondents recognize that there is no ethical issue, while more than 45% recognize that there is an issue.

3.5 Results from econometric analysis

To quantitatively compare the ethical concerns among the six studies, I conduct a regression analysis. In this section, I use a dataset compiled by pooling the responses from the sample of 2,107 respondents. Consider an ordered logit model in the latent variable:

$$y^*_{ij} = \sum_{j=1}^{5} \beta_j \cdot EXP_j + x'_i \cdot \gamma + \delta \cdot z_{ij} + \varepsilon_{ij}, \tag{1}$$

where y^*_{ij} denotes the degree of ethical issues in study j recognized by respondent i. EXP_j is a dummy variable indicating study j, where $j = 1, \ldots, 5$ represents Fryer et al. (2015) to Hanna et al. (2016), respectively. x_i represents a vector of characteristics, z_{ij} represents an order when study j appears in a survey of respondent i, and ε_{ij} is the error term, which is assumed to follow a standard logistic distribution.[26] The five studies are compared to Hosono and Aoyagi (2018) by estimating β_j.

The observed, ordered dependent variable is linked to the latent variable y^*_{ij} through cut points μ_k in the following way:

$$y_{ij} = \begin{cases} 1 \text{ if } y^*_{ij} < \mu_1, \\ k \text{ if } \mu_{k-1} \leq y^*_{ij} < \mu_k \text{ where } k = \{2, 3, 4\}, \\ 5 \text{ if } \mu_4 \leq y^*_{ij}. \end{cases}$$

[26] In the first survey, the variable *Order* is an integer ranging from one to six. See Footnote 12 for more information.

Columns (1) and (2) in Table 3 report the estimation results. The estimated coefficients for Fryer et al. (2015) and Allcott (2011) are negative and significant, meaning that respondents on average recognize less ethical issues in them compared to Hosono and Aoyagi (2018). Landry et al. (2006) is significantly positive, meaning that significantly large ethical issues are recognized. The coefficients for Thornton (2008) and Hanna et al. (2016) are close to zero and not significant at the 10% level, meaning that ethical issues are almost similar to those of Hosono and Aoyagi (2018). *Order* is statistically significantly negative, meaning that the recognition of ethical issues is small for the same experiment if it is shown later in the survey.

[Table 3]

Columns (3) and (4) in Table 3 report the results from linear regressions of Equation (1). The signs and statistical significance of the coefficients are similar to the ordered logit results. The constant term in Column (3) is 3.2, meaning that Hosono and Aoyagi (2018) is, on average, recognized as "3: Neutral" or slightly worse. Since the coefficient for Fryer et al. (2015) is -0.22, its average ethical issue is 2.8 on a five-point scale. The coefficient for Landry et al. (2006) is 0.37 and significant. Columns (2) and (4) consistently show that women are more likely to recognize ethical issues than are men. Age is positively associated with ethical concerns. The coefficient for *Order* is -0.03.

4 Randomized survey experiments on two experiments

4.1 Overview of the second survey

The results of the previous section show that Fryer et al. (2015) is recognized as having the least ethical issues, while Landry et al. (2006) is recognized as having the most ethical issues among the six studies. Why do ethical concerns vary among the experiments? Can we alleviate these concerns by modifying the original studies?

To investigate the above questions, the second survey was designed. It was implemented in March 2018 by INTAGE Research Inc. In the second survey, I focus on the two studies as a contrasting example and use the design of a randomized online survey experiment. I develop three hypotheses, as described below, for each study. Using the same procedure as that of the first survey, 2,146 respondents are invited to take the second survey and are randomly shown one of four descriptions in each study.[27] A description and a question for each study is shown in random order.

4.2 Hypotheses and treatments

4.2.1 Hypotheses about small ethical concerns in Fryer et al. (2005)

Examining the description used in the first survey leads us to several hypotheses regarding why the work of Fryer et al. (2015) is ethically more acceptable than are other studies. First, informed consent may matter. The last sentence in the description mentions the following: "*the parents of the 140 children who became subjects of the study received an explanation regarding them being the subjects of the study, and they gave their consent.*" The presence of consent from subjects, which is missing in Landry et al. (2006), could have alleviated the recognition of an ethical issue. The first treatment tests this hypothesis by deleting this last sentence from the description.

Second, the sampling strategy may matter. In the CHECC project, households were recruited to the project, applied according to their decision, and were randomly assigned to control and treatment groups (Fryer et al., 2015; Cappelen et al., 2020). In contrast, the samples in Landry et al. (2006) did not request to be solicited but became targets of door-to-door fundraising. Thus, in the second treatment, I modified the sentence to mention that subjects were defined by a researcher rather than by applicants as subjects: *parents and their children from 140 families living in the area are defined as the research subjects.*

[27] Respondents for the first survey are intentionally excluded from the second survey. The second online survey was conducted from March 2 to March 5, 2018.

Appendix 2.1 provides the full description of this treatment group.

Third, the existence of a followup for the control group may matter. In the CHECC project, parents and their children in the control group are also invited to holiday parties (see, Gneezy and List, 2013). This may be recognized as compensation to the control group and alleviate the issues. In the third treatment, I deleted the sentences on invitations to holiday parties (see Appendix 2.2 for the description).

4.2.2 Hypotheses about large ethical concerns in Landry et al. (2006)

To investigate methods to ease ethical concerns by modifying Landry et al. (2006), I developed three hypotheses. First, the research design of a randomized field experiment may increase the recognition of ethical issues. To test whether it is worse than other research designs in terms of ethical concerns, I changed the program evaluation methodology to a before–after study of an intervention without a control group (see Appendix 2.3 for the description).

Second, the treatment may matter. Landry et al. (2006) use a raffle to encourage donations. As previous studies discuss a crowding out of intrinsic altruism by extrinsic incentives (e.g., Bénabou and Tirole, 2006), people may not like this treatment as a means of fostering prosocial behavior. Alternatively, in the second treatment, I change the treatment to social comparison information that is used in Allcott (2011) and others for energy conservation and Frey and Meier (2004) and Shang and Croson (2009) for charitable giving. Specifically, I mention that donations are collected with flyers where a message of "*In the neighboring town, 80% of the households donated*" is printed (see Appendix 2.4 for the description).

Third, the topic of the study may matter. Studies to encourage charitable giving may be recognized as unethical, regardless of how we encourage or evaluate such programs. Theoretically, charitable giving is modeled as the private provision of public goods (e.g., Bergstrom et al., 1986). Similarly, household waste sorting to decrease social cost to the environment is also modeled as the private provision of public goods (e.g., Fullerton and Kinnaman, 1995;

Brekke et al., 2003). Therefore, in the third treatment, I change the topic of study from charitable giving to waste sorting, which is similar to Hosono and Aoyagi (2018). Note that I keep the treatment and methodology of program evaluation unchanged. More specifically, I mention that Professor X collaborates with a city government and calls for sorting food waste from other garbage. In the campaign, households are "*asked to sort with a raffle in which one among all recyclers could win JPY 100,000*" (see Appendix 2.5 for the description).

[Table 4]

Table 4 summarizes the four groups for each study. Respondents in the control group are shown the same descriptions used in the first survey. Respondents are randomly assigned to one group among four groups for each study. This results in 2,146 respondents being randomly assigned to 16 groups. As in the first survey, the orders in which Fryer et al. (2015) and Landry et al. (2006) have been shown are randomly determined. In this second survey, I use "Professor X" as the implementer of the program for all the descriptions.

4.3 Verifying randomizations

[Table 5]

[Table 6]

Tables 5 and 6 present summary statistics for respondents of randomized survey experiments on Fryer et al. (2015) and Landry et al. (2006), respectively. The means and standard deviations are separately reported for each of the four groups. The p-values of the tests of the null hypotheses that two groups cannot be distinguished are also reported. The results show that only three and four differ at $p < 0.10$ out of 39 differences each in Fryer et al. (2015) and Landry et al. (2006), respectively. From these, I conclude that the four groups in each survey experiment are very similar.

4.4 Main results

[Figure 2]

Figure 2 shows the distribution of the responses to the survey on Fryer et al. (2015). Panel A shows the responses in a control group, where the distribution is quite similar to that in the first survey (see Panel A in Figure 1). This result shows that I succeeded in replicating the first survey one year later with different samples from the same country. Panels B, C, and D show the responses in the treatment 1, 2, and 3 groups, respectively. All the treatments decrease the recognition of no ethical issue while increasing neutral responses.

[Figure 3]

Figure 3 shows the distribution of the responses to the survey on Landry et al. (2006). Again, the response in the control group (Panel A) is similar to the response in the first survey (Panel C in Figure 1), meaning that the result is replicated. Treatments 1 (Panel B) and 2 (Panel C) are similar to the control, while treatment 3 (Panel D) slightly shifts the distribution to the left.[28]

To evaluate the causal impacts of the treatments, I estimate the models of ordered logit and OLS separately for the samples in each study:

$$y_{ij}^{*} = \beta_1 \cdot T_1 + \beta_2 \cdot T_2 + \beta_3 \cdot T_3 + \delta \cdot z_{ij} + \varepsilon_i, \qquad \text{for } j = 1 \text{ or } 3, \tag{2}$$

where T_n is a dummy variable indicating treatment n.

[Table 7]

Table 7 reports the results of the survey on Fryer et al. (2015). I compute p-values based on the randomization inference procedure of Young (2019) for individual treatment

[28] Note that the distribution of the responses to the waste-sorting version of Landry et al. (2006)(Figure 3 Panel D) is closer to that of Hosono and Aoyagi (2018)(Figure 1 Panel F) rather than that of Landry et al. (2006) in the first survey (Figure 1 Panel C).

effects. I also report the results adjusting for multiple hypothesis testing using the procedure of Westfall and Young (1993) under the null hypothesis that all treatment effects in the equation are zero.

The results show that deleting the sentence on informed consent by parents increases ethical concerns (the randomization-t p-value of 0.007, column 1). The coefficient for this treatment is 0.17 for the OLS (column 3). This magnitude of the effect is similar to the difference between Fryer et al. (2015) and Thornton (2008) in the first survey (Table 3, column 3). If the sample is selected by a researcher, irrelevant to one's willingness to participate, then ethical concerns increase (p-value 0.016), while the magnitude of the effect is slightly smaller than that of treatment 1. Deleting the sentence on holiday parties in which control groups are also invited does not increase these concerns (p-value 0.963). From the results, adjusting for multiple hypothesis testing, I can reject the null hypothesis that all treatment effects are zero (p-value 0.018). Finally, *Order* negatively affects the recognition of ethical issues, which is consistent with the first survey.

[Table 8]

Table 8 reports the results of the survey on Landry et al. (2006). Changing the methodology of program evaluation from a randomized field experiment to a before–after study does not decrease ethical concerns (the randomization-t p-value of 0.478, column 1). Changing treatments from a raffle to social comparison message slightly and weakly decreases ethical concerns (p-value 0.097). Finally, changing a topic of the study from encouraging charitable giving to waste sorting decreases ethical concerns (p-value 0.000 for the individual coefficient and 0.001 for the result, adjusting for the Westfall-Young multiple testing). The magnitude of this effect is large. The coefficient for this treatment is -0.21 for the OLS (column 3), which accounts for more than half of the difference between Landry et al. (2006) and Hosono and Aoyagi (2018) in the first survey (the coefficient of 0.37, Table 3, column 3).

4.5 Subgroup analysis

[Table 9]

[Table 10]

Tables 9 and 10 report the results of subgroup analyses to examine whether there is heterogeneity in impacts by gender. Table 9 reports that women are significantly affected by the treatments in Fryer et al. (2015).[29] The result for men shows no significant impacts for any of the three treatments. Moreover, men are not affected by the order of the survey, while women are affected significantly.

Table 10 reports that, unlike in Table 8, changing a raffle to a message in Landry et al. (2006) significantly decreases the ethical concerns of women. Furthermore, the magnitude of the effect is not small, as the coefficient for the treatment is -0.18 for the OLS (column 3). For the treatment that changes the topic from charitable giving to waste sorting, the result for women shows significant negative impacts, while men show weakly significant and nonnegligible negative impacts. Overall, there is heterogeneity in the impacts—women are more sensitive than men to the modifications of the studies in terms of ethical concerns.

5 Discussion

5.1 Additional analysis on the effect of an implementer

In the literature, it is considered that practical ethical issues of field experiments are likely associated with a question of whether the researchers who are designing programs should be regulated as researchers or as implementers (for more details, see Glennerster and Powers,

[29] I also report p-values adjusted for multiple-hypothesis testing using the procedure of Westfall and Young (1993) and Young (2019) within two regressions of a same model for women and men (e.g., columns 1 and 4 in Table 9). I can reject the null hypothesis that all treatment effects are zero for both men and women (p-value of 0.001 for columns 1 and 4). Similarly, I can reject the null hypothesis for Table 10 as well (p-value 0.003).

2016).[30] While the previous studies discuss this issue from the normative perspective, I empirically examine respondents' recognition of ethical issues by the type of implementers in this subsection.

Do respondents recognize fewer ethical issues if the experiment is run by an implementer other than researchers? To examine this question, in the three studies examined in the first survey, I randomly made small changes in the descriptions. For the respondents who are assigned Allcott (2011), Hanna et al. (2016), or Hosono and Aoyagi (2018), a randomly assigned half of them are shown a description that mentions the implementer of the program being someone other than "Professor X."

In the experiments studied in Allcott (2011), a company called OPOWER was the implementer of the program. Thus, a randomly assigned half of the respondents is shown descriptions similar to those in Appendix A1.3, but "Professor X" is replaced with "a company." In the experiments studied in Hanna et al. (2016), the program was not implemented by the authors but by an NGO. Thus, a randomly assigned half of the respondents is shown descriptions similar to those in Appendix A1.4, but "Professor X" is replaced with "a nonprofit organization." In the experiments studied in Hosono and Aoyagi (2018), the program was implemented by the JICA, as was already mentioned in Section 3. Thus, "Professor X" is replaced with "an international development agency."

This design of randomized survey experiments allows me to estimate a causal impact of changing the implementer of the experiment to a nonresearcher. I conduct regression analyses using the subsample of the dataset used in Section 3.

[Table 11]

Table 11 reports the estimation results.[31] Changing the implementer of the program from a researcher to a company does not change the concerns (columns 1 and 2). Changing the

[30] In contrast to the present study, Barnett and Camfield (2016) discuss specific ethical issues that arise in the randomized evaluation of programs by nonresearchers.

[31] Table A2 in the Online Appendix presents the summary statistics for the respondents in the randomized survey experiments of Table 11 to verify the randomizations. From this table, I conclude that two groups in each survey experiment are very similar.

implementer of the program to an international development agency does not change the concerns (columns 5 and 6). However, changing the implementer of the program to a non-profit organization (NPO) significantly decreases the concerns (columns 3 and 4). Since the coefficient (and the standard error) of the OLS estimation result is -0.11 (0.06), the magnitude of this *implementer effect* is approximately half of the effect of changing an outcome variable from charitable giving to waste sorting (see Table 8). Note that the treatment is significant even after adjusting for multiple-hypothesis testing within ordered logit and OLS regression (the randomization-t p-value of 0.064). The result suggests that although the magnitude of the effect is not large, the implementer of the program can affect respondents' recognition of ethical issues.

Several interpretations are possible. Respondents may consider that the objective of the program and its random assignment is different for researchers and NPOs. People may consider that unlike NPOs, researchers may randomize an intervention just to extend knowledge but not to improve social welfare. If they feel this way, then they may rate the experiment conducted by researchers lower than that conducted by NPOs. Another interpretation can be that respondents may trust NPOs more than social scientists. If the implementer of the program reliably explains how the findings obtained from the experiment can contribute to society, subjects and related individuals may be more likely to accept the experiment. In addition, the result implies that ethical concerns are lesser when a researcher evaluates a program implemented by NPOs rather than one implemented by herself/himself.

5.2 Robustness checks related to the time spent on the surveys

Some respondents may not carefully read the descriptions. I conduct a comparison of the six studies the same way as I did in Section 3 but drop respondents whose time spent on the survey is in the bottom 10% (see Online Appendix Table A3). The result is consistent with Table 3. Moreover, the absolute values of the estimated coefficients are larger than those in Table 3, indicating that differences in ethical concerns among studies become larger if we focus on respondents who take a long time to complete the survey.

Similarly, I analyze the two randomized survey experiments considering the time spent on the survey. For the dataset used in Section 4, I create a dummy variable that takes a value of one if time spent on the survey is longer than the median and zero otherwise (*Long time*). Table A4 in the Online Appendix shows the results of the analyses incorporating the interaction terms of treatments and *Long time*. For Fryer et al. (2015), the interaction terms are consistent with Table 7, while treatments without interaction with *Long time* are not significant. This result can be interpreted as those who read the description carefully being more sensitive to the lack of informed consent or self-selection into the experiment.[32]

The result is slightly different for Landry et al. (2006). Table A5 shows the result, which is consistent with Table 8 for treatment 3 (waste sorting rather than donations) *without* the interaction. This suggests that those who read the description quickly find fewer ethical issues when glancing a waste sorting study; however, changing the topic is not enough to alleviate the concerns of those who read the description carefully. Finally, changing the design of the study to a before–after study does not affect the concerns within each of the two groups (*Long time* = 0 or 1), suggesting no heterogeneous effects and an average effect.

5.3 Interpretations and implications of the results

Several implications are obtained from a series of surveys. From the first survey, I find that the distribution of the recognition of ethical issues widely varies among the six studies. Implementing partners frequently raise ethical concerns about randomized evaluations in general; however, whether subjects identify ethical issues depends on the experiments. Not all field experiments but some specific topics, treatments and designs involve ethical issues. In a specific worst case, researchers are required to modify research plans to improve social welfare through research activities.

[32] Interestingly, the coefficient for *Long time* is negative and significant. This correlation can be interpreted in two ways. First, those who recognize relatively large ethical issues are more likely to quickly read through the description. Second, those who spend a longer time reading the description recognize less ethical issues as a result of reading it carefully.

At first glance, experiments that may affect lifetime success, such as early childhood interventions, seem ethically contentious. However, the results reveal that the number of respondents who recognize ethical issues is the lowest for the CHECC project. Respondents may balance the risks and benefits of the experiment considering whether the findings from the experiment are beneficial and relevant to their lives. Another explanation is that a situation where only half of the applicants are admitted to attend a preschool is common and unsurprising for the respondents since the demand for subsidized childcare often exceeds supply in Japan (for more details, see Yamaguchi et al., 2018). People may be more likely to accept an experiment if the partial and random assignment of the treatment is a familiar situation for the context and culture of their lives.

The second survey partly identifies the reasons for low ethical concerns in the CHECC project. First, women recognize more ethical issues if there is no sentence on informed consent. Among the six studies, the CHECC is the only experiment that informed the subjects of the objective of the study and acquired consent. Note that for the descriptions of the other five experiments, I explicitly mentioned that the subjects were not fully informed of the objectives and designs of the studies (see Appendix A). This result empirically supports the normative discussions in the literature on the importance of informed consent (e.g., Glennerster and Powers, 2016; Teele, 2014). Second, respondents (especially women) recognize fewer ethical issues if subjects voluntarily participate in an experiment based on their decisions compared to researchers randomly selected from the population. Taken together, the random assignment of treatments over subjects who agreed to be in the experiment is recognized as being better from an ethical perspective.

These findings pose tradeoffs to randomized field experiments. Informing subjects that they are taking part in an experiment may change their behavior (Duflo et al., 2007; Harrison and List, 2004). So-called Hawthorne and John Henry effects can occur when we acquire informed consent and may limit the external validity of experiments. Similarly, self-selection into an experiment often makes the sample different from the policy population, which results in biases in the estimate and limits external validity (Deaton, 2010; Peters et al., 2016). Apparently, researchers and implementers face a difficult problem of balancing the external validity of the result and ethical concerns of subjects when running randomized evaluations.[33]

Allcott (2011) involves the second-least ethical issues. This result ethically supports the recent rise in HER experiments in Japan (Jyukankyo Research Institute Inc., 2016; Behavioral Sciences Team, 2019). For Thornton (2008), Hanna et al. (2016), and Hosono and Aoyagi (2018), the two sets of responses—those who recognize issues and those who do not recognize issues—are almost similar in amount. Ethical issues may be less salient for the Japanese if the experiments are conducted in other countries, such as less developed countries.

Among the six examined studies, Landry et al. (2006) is recognized as the least acceptable from an ethical perspective. The result of the second survey suggests that respondents do not recognize concerns because the researcher randomizes the treatment. Possibly, however, respondents are concerned with the research question itself; that is, "Can we encourage charitable giving by a raffle?" One interpretation of the result is that respondents believe that it is unethical to incentivize charitable giving. My result shows that it is less problematic if subjects are solicited using a message with a nudge. Previous studies examine the crowding-out of intrinsic motivations to donate by monetary incentives (e.g., Mellström and Johannesson, 2008). People may dislike being incentivized to make donations. This implies that the preferences for experiments are associated with the preferences for treatments. The result also suggests that preferences are associated with the type of outcome variables. Holding the treatment constant and changing the outcome variable from charitable giving to waste sorting cease such concerns. This suggests that if the motivation to use the experiment is not an evaluation of a program (e.g., a raffle to encourage charitable giving) but rather a test of a theory (e.g., a model of voluntary provision of public goods), then we can alleviate

[33] Teele (2014) also notes the tradeoff between the ethical principle in the Belmont Report and the Hawthorne effect.

ethical concerns by changing the topic and context of the study.

5.4 Comparison with Mislavsky et al. (2019) and Meyer et al. (2019)

In this subsection, I discuss how the result of this paper complements the evidence obtained from two previous studies. Motivated by the experiment by Facebook on emotional contagion,[34] which received backlash, Mislavsky et al. (2019) conducted a series of randomized online survey experiments to study the acceptability of field experiments implemented by companies.[35] In one of their surveys, they examine a hypothetical scenario in which Facebook plans to change the sort status updates users see to a new format in which the "happier" status updates are shown first.[36] Their survey randomizes respondents to either show the scenario of a randomized field experiment where only half of the customers experience the new sorting or the scenario where Facebook decides not to change the way they sort. From their survey results, they find that the acceptability of the two scenarios is not significantly different.[37] This result is consistent with the result for treatment 1 in Table 8. Note that I examine an experiment implemented by a researcher but not a company. In addition, the experimental design was compared to a before–after study without a control group.

Additionally, Mislavsky et al. (2019) compare a similar scenario to the above, but the new way of sorting has the "sadder" status updates being shown first, which is a less acceptable policy change for respondents. Respondents, on average, believe that it is significantly less acceptable if Facebook conducts an experiment to examine the impact of the "sadder" status updates compared to the scenario where Facebook examines this policy but decides not to change. However, respondents believe that this experiment is slightly more acceptable than the scenario where Facebook decides to change its policy to the "sadder" policy without an

[34] See Kramer et al. (2014) for Facebook's field experiment, which manipulated the content seen by users.

[35] Mislavsky et al. (2019) conducted the survey experiments through Amazon's Mechanical Turk (MTurk) using Qualtrics.

[36] See Study 4 (Policies: Happy/No Change) in Mislavsky et al. (2019).

[37] A main question used in Mislavsky et al. (2019) is as follows: Is it okay for Facebook to run this experiment? The question is answered on a seven-point scale (1 = It is really bad; 4 = It is okay; 7 = It is really good).

experiment. From their evidence and several other results obtained by the above authors, they argue that Facebook faced backlash probably not because it ran an experiment but because it tested unacceptable policies. From my survey experiment, I obtain a similar implication, where some studies are unacceptable due to topics and treatments, regardless of whether they involve experiments or not.

Relatedly, Meyer et al. (2019) conducted another series of randomized online survey experiments.[38] They study respondents' perceptions of the appropriateness of field experiments using hypothetical scenarios. In one of their scenarios, a hospital director attempts to reduce infections due to medical treatments by doctors and comes up with two ideas to achieve this goal. One idea is to use a badge, and the other idea is to use a poster on the wall to help doctors remember standard safety precautions. In this setting, respondents are randomly assigned to either a group that is shown the script where a director decides to "use a badge (A)," "use a poster (B)," or "run an experiment to compare these two ideas (A/B test)." Then, respondents are asked the following: "How appropriate is the director's decision?"[39] The result shows that respondents in the A/B test group choose "inappropriate" significantly more than do the other groups. They also find similar results in other domains, such as comparing two poverty alleviation programs, and conclude that people frequently rate A/B tests as inappropriate compared to universally implementing one treatment.[40]

Is the result of Meyer et al. (2019) contradictory to those of Mislavsky et al. (2019) and the present study? Several differences exist among the three studies. Meyer et al. (2019) conducted randomized field experiments that compared two *unobjectionable* interventions.[41] Note that there is no pure control group in their hypothetical experiments. The appropriateness of the experiment is compared with the universal implementation of one intervention (a

[38] Meyer et al. (2019) also conducted the survey through MTurk using the SurveyMonkey, Qualtrics, and Pollfish platforms. In their Supplementary Information, they mention that all participants accessed these platforms using American IP addresses.

[39] The question is answered on a five-point scale (1 = very inappropriate; 3 = neither inappropriate nor appropriate; 5 = very appropriate).

[40] Other topics that Meyer et al. (2019) examine are genetic testing, autonomous vehicles, retirement plans, health worker recruitment, teacher well-being, and basic income.

[41] From Study 1 of Meyer et al. (2019), all the treatments are found to be unobjectionable by respondents.

badge or a poster) but not with a scenario without any intervention. In such a setting, Meyer et al. (2019) find that people are averse to being experimented on. In contrast, Mislavsky et al. (2019) compare "universally keeping the way of sorting unchanged" and "testing happier sorting," and they find no significant difference. In addition, they find that respondents feel okay toward "testing sadder sorting" more than toward "universally implementing sadder sorting," the latter of which they consider an unacceptable policy change. Finally, the present study finds that encouraging charitable giving using a raffle is considered ethically unacceptable, and the evaluation of the treatment using the experiment and that using a before–after study of the intervention are equally unacceptable.

Based on these differences, my interpretation of the results of the three studies is as follows. People do not prefer some type of treatments and research topics. For example, people dislike being studied and incentivized regarding their decisions to donate. In those cases, people do not care whether they are studied experimentally or not. They just do not prefer the study. In contrast, if they do not have strong negative preferences toward the treatment or the topic, then they do not prefer to be experimented with, as shown by Meyer et al. (2019). Moreover, people prefer the universal implementation of one treatment if it seems beneficial to them.

5.5 Study limitations

Some limitations in the present paper are worth noting. First, I compare only six randomized field experiments among thousands implemented or analyzed by economists.[42] Second, while my randomized survey experiments partly unmask the reasons for relatively low or high ethical concerns for specific studies, the findings in the present study cannot fully explain the large difference between the CHECC project and Landry et al. (2006). Relatedly, I

[42] There are 3,511 studies registered in the AEA RCT Registry as of April 28, 2020. Peters et al. (2016) review 92 papers that used a randomized field experiment and were published between 2009 and 2014 in the American Economic Review, Econometrica, Quarterly Journal of Economics, Journal of Political Economy, Review of Economic Studies, Economic Journal, Journal of Public Economics, and the American Economic Journal: Applied Economics. Lewis and Rao (2015) review 25 randomized field experiments that measure returns to digital advertising.

show that the examined strategies can alleviate concerns; however, the magnitude of such alleviation is not large. Third, there may exist disutility other than that represented by ethical concerns. For example, subjects may feel anxiety due to being treated by untested treatments. Furthermore, subjects may find disutility from the inequality of treatment status as a result of a random assignment. These types of possible disutility are not examined in the present paper.

The present study surveys the population, which is somehow different in regard to preferences than are Americans and Europeans. Previous studies also elicit preferences of the resident population in Japan. Using the dataset of the Global Preference Survey provided by Falk et al. (2016, 2018), we can compare the preferences of Japanese individuals with those of individuals from other countries. Compared to the Western European average, Japanese individuals are on average impatient, risk-averse, equivalently altruistic, less reciprocal, and less trusting in others.[43] Kameda and Sato (2017) conduct classroom experiments in Japan, where the design of the experiments is the same as that of Engelmann and Strobel (2004). From their results, they conclude that fewer students in Japan than in Germany have concerns about efficiency.

Moreover, my sample is not even a random sample of the population in Japan. Although the company carefully aimed to create the sample where the distributions of three characteristics represent those of Japan, the samples were opted into the survey given the randomly sent invitation (see Footnote 11). More problematically, the samples are on the panel of the online survey company, which raises a concern about the external validity of the study. Respondents of this study might have more experience being surveyed compared to average residents in Japan, which is likely to be associated with ethical concerns in academic research. I do not have quantitative evidence to reject the possibility that respondents in this study are systematically different in their ethical view from the policy population. I instead

[43] Averages are calculated using the dataset of Falk et al. (2016, 2018). Compared to the United States average, Japanese individuals are on average impatient, risk-averse, less altruistic, not positively reciprocal but negatively reciprocal, and less trusting in others. Compared to the world average, however, Japanese individuals are slightly patient.

emphasize that this study shows that it is possible to alleviate such concerns by modifying the research plans of randomized field experiments.

6 Conclusions

Economists who study policies are, in general, interested in the welfare gains and losses of citizens. Randomized field experiments can improve social welfare by rigorously evaluating policies or testing economic theories. However, there is a concern that experiments may generate utility loss for subjects and implementing partners. In this study, I conduct an online survey to compare potential subjects' ethical concerns with six previous experiments in the field of economics. I find that the degree of ethical concerns varies among respondents and experiments. A certain proportion of respondents are very concerned, while others are not. Both researchers and practitioners need to take into account this heterogeneity in preferences for economic studies.

The majority of respondents find ethical issues in encouraging charitable giving by a raffle, regardless of whether the study adopts experimental designs. In contrast, many respondents find no ethical issues in evaluating preschool, especially if subjects decide to participate and are informed enough about the experiment. These contrasting results may be surprising to economists. We—economists—need to be aware that we may have less information on the preferences for experiments of laypeople. The method of this paper can be used to understand the utility or disutility of field experiments for citizens not only *ex post* but also *ex ante*. Future tasks include conducting similar surveys in other countries and examining other experiments both before and after the interventions.

From two randomized survey experiments, I find that it is possible to alleviate concerns by modifying research projects. However, the strategies to alleviate the concerns bring about tradeoffs. Easing ethical concerns results in decreasing the external validity of the randomized evaluation design.

As emphasized by Glennerster and Powers (2016), balancing the risks and benefits of research is required for economists to improve social welfare through their works. This paper reveals that randomized experiments are useful for examining a wide range of issues, even including the ethical issues involved in this method. Thus, we need to improve this beneficial method and reduce the risks involved in it to further utilize it.

References

Allcott, Hunt (2011), “Social norms and energy conservation.” *Journal of Public Economics*, 95, 1082–1095.

Allcott, Hunt and Judd B. Kessler (2019), “The welfare effects of nudges: A case study of energy use social comparisons.” *American Economic Journal: Applied Economics*, 11, 236–76.

Allcott, Hunt and Todd Rogers (2014), “The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation.” *American Economic Review*, 104, 3003–37.

Andreoni, James, Amalia Di Girolamo, John A. List, Claire Mackevicius, and Anya Samek (2020), “Risk preferences of children and adolescents in relation to gender, cognitive skills, soft skills, and executive functions.” *Journal of Economic Behavior & Organization*, 179, 729–742.

Andreoni, James, Michael A. Kuhn, John A. List, Anya Samek, Kevin Sokal, and Charles Sprenger (2019), “Toward an understanding of the development of time preferences: Evidence from field experiments.” *Journal of Public Economics*, 177, 104039.

Angrist, Joshua D and Guido W Imbens (1991), “Sources of identifying information in evaluation models.” Technical Working Paper 117, National Bureau of Economic Research. Available at `https://www.nber.org/papers/t0117`.

Ayres, Ian, Sophie Raseman, and Alice Shih (2012), “Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage.” *Journal of Law, Economics, and Organization*, 29, 992–1022.

Barnett, Chris and Laura Camfield (2016), “Ethics in evaluation.” *Journal of Development Effectiveness*, 8, 528–534.

Behavioral Sciences Team (2019), “Annual report (FY2017 and FY2018).” Report, Behavioral Sciences Team (BEST), Government of Japan. Available at `http://www.env.go.jp/earth/ondanka/nudge/report1_Eng.pdf`.

Bénabou, Roland and Jean Tirole (2006), “Incentives and prosocial behavior.” *American Economic Review*, 96, 1652–1678.

Bensch, Gunther and Jörg Peters (2015), “The intensive margin of technology adoption – experimental evidence on improved cooking stoves in rural senegal.” *Journal of Health Economics*, 42, 44–63.

Bergstrom, Theodore, Lawrence Blume, and Hal Varian (1986), “On the private provision of public goods.” *Journal of Public Economics*, 29, 25–49.

Brekke, Kjell Arne, Snorre Kverndokk, and Karine Nyborg (2003), “An economic model of moral motivation." *Journal of Public Economics*, 87, 1967–1983.

Cappelen, Alexander, John List, Anya Samek, and Bertil Tungodden (2020), "The effect of early-childhood education on social preferences." *Journal of Political Economy*, Forthcoming.

Chong, Alberto, Dean Karlan, Jeremy Shapiro, and Jonathan Zinman (2015), "(Ineffective) messages to encourage recycling: Evidence from a randomized evaluation in peru." *World Bank Economic Review*, 29, 180–206.

Costa, Dora L and Matthew E Kahn (2013), "Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment." *Journal of the European Economic Association*, 11, 680–702.

Cruces, Guillermo, Ricardo Perez-Truglia, and Martin Tetaz (2013), "Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment." *Journal of Public Economics*, 98, 100–112.

Deaton, Angus (2010), "Instruments, randomization, and learning about development." *Journal of Economic Literature*, 48, 424–455.

Delavande, Adeline and Hans-Peter Kohler (2016), "HIV/AIDS-related expectations and risky sexual behaviour in Malawi." *Review of Economic Studies*, 83, 118–164.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier (2012), "Testing for altruism and social pressure in charitable giving." *Quarterly Journal of Economics*, 127, 1–56.

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007), "Using randomization in development economics research: A toolkit." In *Handbook of Development Economics* (T. Paul Schultz and John A. Strauss, eds.), volume 4, chapter 61, 3895–3962, Elsevier.

Dupas, Pascaline (2011), "Do teenagers respond to HIV risk information? evidence from a field experiment in kenya." *American Economic Journal: Applied Economics*, 3, 1–34.

Edwards, James T. and John A. List (2014), "Toward an understanding of why suggestions work in charitable fundraising: Theory and evidence from a natural field experiment." *Journal of Public Economics*, 114, 1–13.

Elías, Julio J, Nicola Lacetera, and Mario Macis (2019), "Paying for kidneys? A randomized survey and choice experiment." *American Economic Review*, 109, 2855–88.

Engelmann, Dirk and Martin Strobel (2004), "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments." *American Economic Review*, 94, 857–869.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde (2018), "Global evidence on economic preferences." *Quarterly Journal of Economics*, 133, 1645–1692.

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde (2016), "The preference survey module: A validated instrument for measuring risk, time, and social preferences." IZA Discussion Paper 9674.

Favereau, Judith (2016), "On the analogy between field experiments in economics and clinical trials in medicine." *Journal of Economic Methodology*, 23, 203–222.

Frey, Bruno S and Stephan Meier (2004), "Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment." *American Economic Review*, 94, 1717–1722.

Fryer, Roland, Steven Levitt, John List, and Anya Samak (2013), "Chicago Heights Early Childhood Center: Early results from a field experiment on the temporal allocation of schooling." Available at `https://cpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/f/1276/files/2018/10/CHECC-Presentation-13pmhkk.pdf`.

Fryer, Roland G, Steven D Levitt, and John A List (2015), "Parental incentives and early childhood achievement: A field experiment in Chicago heights." NBER Working Paper Series 21477, National Bureau of Economic Research. Available at `http://www.nber.org/papers/w21477`.

Fullerton, Don and Thomas C. Kinnaman (1995), "Garbage, recycling, and illicit burning or dumping." *Journal of Environmental Economics and Management*, 29, 78–91.

Glennerster, Rachel (2017), "The practicalities of running randomized evaluations: Partnerships, measurement, ethics, and transparency." In *Handbook of Economic Field Experiments*, volume 1, 175–243, Elsevier.

Glennerster, Rachel and Shawn Powers (2016), "Balancing risk and benefit: Ethical tradeoffs in running randomized evaluations." In *Oxford Handbook of Professional Economic Ethics* (George DeMartino and Deirdre McCloskey, eds.), chapter 20, 367–401, Oxford University Press, Oxford.

Gneezy, Uri and John A. List (2013), *The Why Axis: Hidden Motives and the Undiscovered Economics of Everyday Life*. PublicAffairs.

Godlonton, Susan, Alister Munthali, and Rebecca Thornton (2016), "Responding to risk: Circumcision, information, and hiv prevention." *Review of Economics and Statistics*, 98, 333–349.

Godlonton, Susan and Rebecca L. Thornton (2013), "Learning from others' HIV testing: Updating beliefs and responding to risk." *American Economic Review: Papers & Proceedings*, 103, 439–44.

Groves Williams, Leslie (2016), "Ethics in international development evaluation and research: What is the problem, why does it matter and what can we do about it?" *Journal of Development Effectiveness*, 8, 535–552.

Hanaoka, Chie, Hitoshi Shigeoka, and Yasutora Watanabe (2018), "Do risk preferences change? evidence from the great east Japan earthquake." *American Economic Journal: Applied Economics*, 10, 298–330.

Hanna, Rema, Esther Duflo, and Michael Greenstone (2016), "Up in smoke: The influence of household behavior on the long-run impact of improved cooking stoves." *American Economic Journal: Economic Policy*, 8, 80–114.

Harrison, Glenn W. and John A. List (2004), "Field experiments." *Journal of Economic Literature*, 42, 1009–1055.

Hopper, Joseph R. and Joyce McCarl Nielsen (1991), "Recycling as altruistic behavior: Normative and behavioral strategies to expand participation in a community recycling program." *Environment and Behavior*, 23, 195–220.

Hosono, Tomoyuki and Keitaro Aoyagi (2018), "Effectiveness of interventions to induce waste segregation by households: Evidence from a randomized controlled trial in Mozambique." *Journal of Material Cycles and Waste Management*, 20, 1143–1153.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka (2018), "Moral suasion and economic incentives: Field experimental evidence from energy demand." *American Economic Journal: Economic Policy*, 10, 240–67.

Jeuland, Marc, Subhrendu K. Pattanayak, Jie-Sheng Tan Soo, and Faraz Usmani (2020), "Preferences and the effectiveness of behavior-change interventions: Evidence from adoption of improved cookstoves in India." *Journal of the Association of Environmental and Resource Economists*, 7, 305–343.

Jyukankyo Research Institute Inc. (2016), "Improvement of infrastructure to promote rationalization of energy use in fiscal year 2015 (Survey on the effect of providing information on energy use on the promotion of changes in household energy conservation behavior)." Technical report, Minister of Economy, Trade and Industry (METI), Japan.

Kameda, Keigo and Miho Sato (2017), "Distributional preference in Japan." *Japanese Economic Review*, 68, 394–408.

Koford, Brandon C, Glenn C Blomquist, David M Hardesty, and Kenneth R Troske (2012), "Estimating consumer willingness to supply and willingness to pay for curbside recycling." *Land Economics*, 88, 745–763.

Kozuka, Eiji (2018), "Enlightening communities and parents for improving student learning evidence from randomized experiment in Niger." JICA-RI Working Paper, JICA Research Institute.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock (2014), "Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences*, 111, 8788–8790.

Kuziemko, Ilyana, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva (2015), "How elastic are preferences for redistribution? Evidence from randomized survey experiments." *American Economic Review*, 105, 1478–1508.

Landry, Craig E, Andreas Lange, John A List, Michael K Price, and Nicholas G Rupp (2006), "Toward an understanding of the economics of charity: Evidence from a field experiment." *Quarterly Journal of Economics*, 121, 747–782.

Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp (2010), "Is a donor in hand better than two in the bush? Evidence from a natural field experiment." *American Economic Review*, 100, 958–83.

Levine, David I, Theresa Beltramo, Garrick Blalock, Carolyn Cotterman, and Andrew M Simons (2018), "What impedes efficient adoption of products? Evidence from randomized sales offers for fuel-efficient cookstoves in Uganda." *Journal of the European Economic Association*, 16, 1850–1880.

Levitt, Steven D and John A List (2009), "Field experiments in economics: The past, the present, and the future." *European Economic Review*, 53, 1–18.

Lewis, Randall A. and Justin M. Rao (2015), "The unfavorable economics of measuring the returns to advertising." *Quarterly Journal of Economics*, 130, 1941–1973.

List, John A (2008), "Informed consent in social science." *Science*, 322, 672.

List, John A and David Lucking-Reiley (2002), "The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign." *Journal of Political Economy*, 110, 215–233.

Matsukawa, Isamu (2018), "Information acquisition and residential electricity consumption: Evidence from a field experiment." *Resource and Energy Economics*, 53, 1–19.

Mellström, Carl and Magnus Johannesson (2008), "Crowding out in blood donation: Was titmuss right?" *Journal of the European Economic Association*, 6, 845–863.

Meyer, Michelle N, Patrick R Heck, Geoffrey S Holtzman, Stephen M Anderson, William Cai, Duncan J Watts, and Christopher F Chabris (2019), "Objecting to experiments that compare two unobjectionable policies or treatments." *Proceedings of the National Academy of Sciences*, 116, 10723–10728.

Miller, Grant and A. Mushfiq Mobarak (2015), "Learning about new technologies through social networks: Experimental evidence on nontraditional stoves in bangladesh." *Marketing Science*, 34, 480–499.

Mislavsky, Robert, Berkeley Dietvorst, and Uri Simonsohn (2019), "Critical condition: People don't dislike a corporate experiment more than they dislike its worst condition." *Marketing Science*, Forthcoming. Available at `https://doi.org/10.1287/mksc.2019.1166`.

Mobarak, Ahmed Mushfiq, Puneet Dwivedi, Robert Bailis, Lynn Hildemann, and Grant Miller (2012), "Low demand for nontraditional cookstove technologies." *Proceedings of the*

National Academy of Sciences, 109, 10815–10820.

Murase, Noriaki, Takehiko Murayama, Shigeo Nishikizawa, and Yuriko Sato (2017), "Quantitative analysis of impact of awareness-raising activities on organic solid waste separation behaviour in Balikpapan City, Indonesia." *Waste Management & Research*, 35, 1013–1022.

Narita, Yusuke (2020), "Toward an ethical experiment." *Proceedings of the National Academy of Sciences*, Forthcoming. Available at `http://dx.doi.org/10.2139/ssrn.3094905`.

O'Flynn, Peter, Chris Barnett, and Laura Camfield (2016), "Assessing contrasting strategies for ensuring ethical practice within evaluation: Institutional review boards and professionalisation." *Journal of Development Effectiveness*, 8, 561–568.

Peters, Jörg, Jörg Langbein, and Gareth Roberts (2016), "Policy evaluation, randomized controlled trials, and external validity—A systematic review." *Economics Letters*, 147, 51–54.

Ravallion, Martin (2009), "Evaluation in the practice of development." *World Bank Research Observer*, 24, 29–53.

Sawada, Yasuyuki, Takeshi Aida, Andrew S. Griffen, Eiji Kozuka, Haruko Noguchi, and Yasuyuki Todo (2019), "Democratic institutions and social capital: Experimental evidence on school-based management from Burkina Faso." Available at `http://www.griffen.e.u-tokyo.ac.jp/COGES.pdf`.

Schultz, P. Wesley (1999), "Changing behavior with normative feedback interventions: A field experiment on curbside recycling." *Basic and Applied Social Psychology*, 21, 25–36.

Shang, Jen and Rachel Croson (2009), "A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods." *Economic Journal*, 119, 1422–1439.

Soetevent, Adriaan R. (2011), "Payment choice, image motivation and contributions to charity: Evidence from a field experiment." *American Economic Journal: Economic Policy*, 3, 180–205.

Takahashi, Kazushi, Yukichi Mano, and Keijiro Otsuka (2019), "Learning from experts and peer farmers about rice production: Experimental evidence from Cote d'Ivoire." *World Development*, 122, 157–169.

Tanaka, Tomoaki, Junichi Yamasaki, Yasuyuki Sawada, and Khaliun Dovchinsuren (2018), "Enlightening communities and parents for improving student learning evidence from randomized experiment in Niger." JICA-RI Working Paper, JICA Research Institute.

Teele, Dawn Langan (2014), "Reflections on the ethics of field experiments." In *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences* (Dawn Langan Teele, ed.), chapter 5, 115–140, Yale University Press, New Haven & London.

Thornton, Rebecca L (2008), “The demand for, and impact of, learning HIV status.” *American Economic Review*, 98, 1829–63.

Thornton, Rebecca L. (2012), “HIV testing, subjective beliefs and economic behavior.” *Journal of Development Economics*, 99, 300–313.

Westfall, Peter H and S Stanley Young (1993), *Resampling-based multiple testing: Examples and methods for p-value adjustment.* John Wiley & Sons, New York.

Yamaguchi, Shintaro, Yukiko Asai, and Ryo Kambayashi (2018), “How does early childcare enrollment affect children, parents, and their interactions?” *Labour Economics*, 55, 56–71.

Yokoo, Hide-Fumi and Tetsuya Harada (2020), “Face-to-face communication on take-up of paid sanitation services: Experimental evidence from Indonesia.” Association of Environmental and Resource Economists (AERE) Annual Summer Conference 2020.

Young, Alwyn (2019), “Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results.” *Quarterly Journal of Economics*, 134, 557–598.

Table 1: Summary of the six experiments examined in the first survey

	(1) Label of the descriptions	(2) Outcome variables	(3) Treatments	(4) Sample size	(5) Informed	(6) Monetary incentive	(7) Human capital	(8) Developing countries	(9) Implementer
1	Fryer, Levitt, and List (2015)	Academic achievement and lifetime earnings	Preschool	140	No	No	No	No	Professor X
2	Thornton (2008)	Going to HIV testing centers to be informed	Reward	3,000	No	Yes	Yes	Yes	Professor X
3	Landry et al. (2006)	Donation	Raffle	4,800	Yes	Yes	No	No	Professor X
4	Allcott (2011)	Electricity consumption	Report	40,000	Yes	No	No	No	Professor X or a company
5	Hanna, Duflo and Greenstone (2016)	Health status	Improved cooking stove	1,600	No	No	Yes	Yes	Professor X or an NPO
6	Hosono and Aoyagi (2018)	Sorting waste	Opportunity to win a laundry detergent	500	No	Yes	No	Yes	Professor X or an IDA

Notes: This table summarizes the six experiments examined in the present study. Column 9 presents the implementer of the program in each description.

Table 2: Summary statistics of the first online survey

	Mean (1)	SD (2)
Female	0.480	0.500
Age	46.673	14.064
Married	0.609	0.488
Living with children	0.379	0.485
Household income (10 thousand JPY)	535.289	249.164
Full-time employee	0.249	0.432
Temporary/contract employee	0.052	0.222
Self-employed	0.056	0.229
Part-time employee	0.124	0.330
Housewife/househusband	0.181	0.385
Unemployed/retired	0.103	0.305
Lives in Tokyo	0.125	0.331
Lives in Osaka	0.072	0.258

Notes: This table reports the means and standard deviations from the first survey. The number of observations is 2,107, except for Household income (the number of observations is 1,645).

Table 3: Comparisons of ethical concerns in the six studies (coefficients)

	Ordered logit		OLS	
	(1)	(2)	(3)	(4)
Fryer et al. (2015)	-0.418***	-0.420***	-0.219***	-0.219***
	(0.090)	(0.090)	(0.048)	(0.047)
Thornton (2008)	0.046	0.036	0.037	0.032
	(0.088)	(0.087)	(0.047)	(0.046)
Landry et al. (2006)	0.663***	0.665***	0.367***	0.366***
	(0.087)	(0.087)	(0.047)	(0.046)
Allcott (2011)	-0.276***	-0.283***	-0.129**	-0.136***
	(0.097)	(0.096)	(0.051)	(0.051)
Hanna et al. (2016)	0.072	0.071	0.044	0.043
	(0.110)	(0.110)	(0.059)	(0.058)
Order (1-6)	-0.069***	-0.070***	-0.034***	-0.034***
	(0.013)	(0.013)	(0.007)	(0.007)
Female		0.304***		0.163***
		(0.074)		(0.039)
Age		0.011***		0.006***
		(0.003)		(0.001)
Constant			3.197***	2.830***
			(0.046)	(0.082)
Control implementers	Yes	Yes	Yes	Yes
Other control variables	No	Yes	No	Yes
Number of Observations	6321	6321	6321	6321
Pseudo R-squared / R-squared	0.013	0.019	0.037	0.054

Notes: This table reports the estimates from the regression analyses in which the dependent variable is the response to the question "Do you recognize any ethical issues in this study?" on a five-point scale (1–5), as shown in Figure 1. Five studies are compared to Hosono and Aoyagi (2018). The coefficients are reported. Standard errors, clustered at the respondent level, are in parentheses. Columns 2 and 4 include other variables in Table 2 as well as Time spent on the survey. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Table 4: The design of the randomized online survey experiments

Experiments:	Control	Treatment 1	Treatment 2	Treatment 3
Fryer et al. (2015)	Same as the first survey	Deleting the informed consent statement	Samples are selected rather than self-selection	Deleting holiday parties statement
Landry et al. (2006)	Same as the first survey	Before-after study without control	Treatment is a message rather than a raffle	Promoting waste sorting rather than donations

Table 5: Summary statistics of the randomized survey experiments by group (Fryer et al., 2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	C	T1	T2	T3	P-value		
					C vs T1	C vs T2	C vs T3
Female	0.488	0.486	0.516	0.471	0.951	0.359	0.585
	(0.500)	(0.500)	(0.500)	(0.500)			
Age	46.153	47.030	46.835	47.664	0.286	0.408	0.062
	(13.415)	(13.528)	(13.532)	(13.156)			
Married	0.603	0.615	0.612	0.633	0.708	0.782	0.323
	(0.490)	(0.487)	(0.488)	(0.483)			
Living with children	0.378	0.384	0.371	0.369	0.851	0.825	0.765
	(0.485)	(0.487)	(0.484)	(0.483)			
Household income	553.589	551.651	537.681	564.353	0.915	0.376	0.558
(10 thousand yen)	(265.456)	(261.062)	(252.768)	(268.040)			
Full-time employee	0.225	0.229	0.242	0.223	0.884	0.519	0.916
	(0.418)	(0.421)	(0.429)	(0.416)			
Part-time employee	0.130	0.151	0.129	0.148	0.335	0.965	0.393
	(0.337)	(0.358)	(0.336)	(0.356)			
Temporary/contract employee	0.061	0.047	0.058	0.072	0.281	0.821	0.475
	(0.240)	(0.211)	(0.234)	(0.259)			
Self-employed	0.039	0.069	0.049	0.056	0.031	0.441	0.202
	(0.194)	(0.254)	(0.216)	(0.229)			
Housewife/househusband	0.175	0.181	0.225	0.178	0.811	0.041	0.895
	(0.380)	(0.385)	(0.418)	(0.383)			
Unemployed/retired	0.115	0.110	0.094	0.111	0.772	0.248	0.831
	(0.320)	(0.313)	(0.292)	(0.315)			
Living in Tokyo	0.132	0.115	0.126	0.128	0.405	0.751	0.838
	(0.339)	(0.320)	(0.332)	(0.334)			
Living in Osaka	0.060	0.067	0.064	0.076	0.617	0.776	0.283
	(0.237)	(0.250)	(0.245)	(0.265)			

Notes: This table reports the means for each of the groups for Fryer et al. (2015) with standard deviations in parentheses. Columns 5–7 report p-values for the differences. The number of observations is 2,146, except for Household income (the number of observations is 1,681).

Table 6: Summary statistics of the randomized survey experiments by group (Landry et al., 2006)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	C	T1	T2	T3	P-value		
					C vs T1	C vs T2	C vs T3
Female	0.481	0.482	0.515	0.482	0.982	0.272	0.975
	(0.500)	(0.500)	(0.500)	(0.500)			
Age	47.019	47.028	46.254	47.375	0.991	0.353	0.661
	(13.411)	(13.373)	(13.532)	(13.335)			
Married	0.602	0.632	0.607	0.621	0.306	0.860	0.517
	(0.490)	(0.483)	(0.489)	(0.486)			
Living with children	0.354	0.405	0.353	0.390	0.082	0.991	0.217
	(0.479)	(0.491)	(0.478)	(0.488)			
Household income	553.800	545.192	543.112	565.366	0.633	0.554	0.526
(10 thousand yen)	(263.917)	(256.813)	(260.631)	(266.316)			
Full-time employee	0.219	0.216	0.246	0.238	0.913	0.283	0.436
	(0.414)	(0.412)	(0.431)	(0.427)			
Part-time employee	0.139	0.135	0.145	0.140	0.856	0.784	0.940
	(0.346)	(0.342)	(0.352)	(0.348)			
Temporary/contract employee	0.067	0.068	0.047	0.057	0.954	0.165	0.524
	(0.250)	(0.251)	(0.212)	(0.233)			
Self-employed	0.061	0.069	0.038	0.044	0.582	0.076	0.218
	(0.240)	(0.254)	(0.190)	(0.206)			
Housewife/househusband	0.156	0.184	0.224	0.196	0.217	0.004	0.081
	(0.363)	(0.388)	(0.417)	(0.397)			
Unemployed/retired	0.122	0.116	0.098	0.094	0.766	0.201	0.139
	(0.328)	(0.321)	(0.297)	(0.292)			
Living in Tokyo	0.135	0.105	0.135	0.126	0.130	0.994	0.643
	(0.342)	(0.307)	(0.342)	(0.332)			
Living in Osaka	0.063	0.062	0.073	0.068	0.943	0.502	0.719
	(0.243)	(0.241)	(0.261)	(0.253)			

Notes: This table reports the means for each of the groups for Landry et al. (2006) with standard deviations in parentheses. Columns 5–7 report p-values for the differences. The number of observations is 2,146, except for Household income (the number of observations is 1,681).

Table 7: Results of the randomized survey experiment (Fryer et al., 2015)

	Ordered logit		OLS	
	(1)	(2)	(3)	(4)
T1: Deleting the informed consent statement	0.312***	0.312***	0.173***	0.172***
	(0.114)	(0.114)	(0.059)	(0.059)
	[0.007]	[0.007]	[0.004]	[0.004]
T2: Samples are selected rather than self-selection	0.273**	0.271**	0.156***	0.154**
	(0.115)	(0.115)	(0.061)	(0.061)
	[0.016]	[0.016]	[0.009]	[0.010]
T3: Deleting holiday parties statement	0.005	0.006	0.008	0.006
	(0.113)	(0.113)	(0.058)	(0.058)
	[0.963]	[0.963]	[0.887]	[0.907]
Order (1/2)		-0.193**		-0.078*
		(0.080)		(0.041)
Constant			2.840***	2.957***
			(0.043)	(0.073)
Multiple-Hypothesis Testing	0.018	0.018	0.010	0.011
Number of Observations	2146	2146	2146	2146
Pseudo R-squared / R-squared	0.002	0.003	0.007	0.009

Notes: This table reports the estimates from regression analyses in which the dependent variable is the response to the question "Do you recognize any ethical issues in this study?" on a five-point scale (1–5), as shown in Figure 2. The impact of changing the description on Fryer et al. (2015) to three treatment descriptions is evaluated. The coefficients are reported. Standard errors are in parentheses in columns 1 and 2. Robust standard errors are in parentheses in columns 3 and 4. The randomization-t p-values are in brackets. Inference in each column is based on a randomization inference procedure of Young (2019). ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. The row of Multiple-Hypothesis Testing reports the randomization-t p-values for the multiple-hypothesis testing test based on a randomization inference procedure of Young (2019), which applies the procedure of Westfall and Young (1993). It tests the null hypothesis that all treatment effects in each equation (each column) are zero.

Table 8: Results of the randomized survey experiment (Landry et al., 2006)

	Ordered logit		OLS	
	(1)	(2)	(3)	(4)
T1: Before-after study without control	-0.081	-0.075	-0.058	-0.054
	(0.112)	(0.112)	(0.061)	(0.061)
	[0.478]	[0.515]	[0.350]	[0.386]
T2: Treatment is a message rather than a raffle	-0.181*	-0.176	-0.087	-0.085
	(0.111)	(0.111)	(0.059)	(0.059)
	[0.097]	[0.109]	[0.132]	[0.142]
T3: Promoting waste sorting rather than donations	-0.396***	-0.400***	-0.208***	-0.211***
	(0.111)	(0.111)	(0.060)	(0.059)
	[0.000]	[0.000]	[0.000]	[0.000]
Order (1/2)		-0.143*		-0.087**
		(0.079)		(0.043)
Constant			3.452***	3.584***
			(0.042)	(0.074)
Multiple-Hypothesis Testing	0.001	0.001	0.001	0.001
Number of Observations	2146	2146	2146	2146
Pseudo R-squared / R-squared	0.002	0.003	0.006	0.008

Notes: This table reports the estimates from regression analyses in which the dependent variable is the response to the question "Do you recognize any ethical issues in this study?" on a five-point scale (1–5), as shown in Figure 2. The impact of changing the description on Landry et al. (2006) to three treatment descriptions is evaluated. The coefficients are reported. Standard errors are in parentheses in columns 1 and 2. Robust standard errors are in parentheses in columns 3 and 4. The randomization-t p-values are in brackets. Inference in each column is based on a randomization inference procedure of Young (2019). ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. The row of Multiple-Hypothesis Testing reports the randomization-t p-values for the multiple-hypothesis testing test based on a randomization inference procedure of Young (2019), which applies the procedure of Westfall and Young (1993). It tests the null hypothesis that all treatment effects in each equation (each column) are zero.

Table 9: Results of subsample analyses (Fryer et al., 2015)

	Female			Male		
	(1) Ologit	(2) Ologit	(3) OLS	(4) Ologit	(5) Ologit	(6) OLS
T1: Deleting the informed consent statement	0.645^{***}	0.646^{***}	0.329^{***}	0.031	0.032	0.022
	(0.165)	(0.166)	(0.081)	(0.157)	(0.157)	(0.085)
	[0.000]	[0.000]	[0.000]	[0.840]	[0.839]	[0.796]
T2: Samples are selected rather than self-selection	0.448^{***}	0.433^{***}	0.230^{***}	0.100	0.101	0.069
	(0.163)	(0.162)	(0.078)	(0.163)	(0.163)	(0.093)
	[0.006]	[0.007]	[0.004]	[0.535]	[0.531]	[0.452]
T3: Deleting holiday parties statement	-0.030	-0.032	-0.007	0.030	0.031	0.015
	(0.165)	(0.164)	(0.080)	(0.155)	(0.155)	(0.084)
	[0.859]	[0.849]	[0.927]	[0.847]	[0.844]	[0.858]
Order (1/2)		-0.358^{***}	-0.163^{***}		-0.040	0.005
		(0.116)	(0.056)		(0.111)	(0.061)
Constant			3.056^{***}			2.861^{***}
			(0.097)			(0.108)
Multiple-Hypothesis Testing	0.001	0.001	0.001			
Number of Observations	1052	1052	1052	1094	1094	1094
Pseudo R-squared / R-squared	0.009	0.012	0.033	0.000	0.000	0.001

Notes: This table reports the estimates from subsample analyses of Table 7. The coefficients are reported. Standard errors are in parentheses in columns 1, 2, 4 and 5. Robust standard errors are in parentheses in columns 3 and 6. The randomization-t p-values are in brackets. Inference in each column is based on a randomization inference procedure of Young (2019). ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. The row of Multiple-Hypothesis Testing reports the randomization-t p-values for the multiple-hypothesis testing test based on a randomization inference procedure of Young (2019), which applies the procedure of Westfall and Young (1993). For example, column 1 reports the result under the null hypothesis that all treatment effects within the two regressions of the same model for women (column 1) and men (column 4) are zero, adjusting for multiple-hypothesis testing.

Table 10: Results of subsample analyses (Landry et al., 2006)

	Female			Male		
	(1) Ologit	(2) Ologit	(3) OLS	(4) Ologit	(5) Ologit	(6) OLS
T1: Before-after study without control	-0.064	-0.065	-0.033	-0.105	-0.089	-0.074
	(0.163)	(0.163)	(0.085)	(0.154)	(0.154)	(0.087)
	[0.703]	[0.697]	[0.705]	[0.492]	[0.559]	[0.394]
T2: Treatment is a message rather than a raffle	-0.367**	-0.365**	-0.180**	-0.035	-0.026	0.002
	(0.159)	(0.159)	(0.080)	(0.156)	(0.156)	(0.086)
	[0.020]	[0.021]	[0.023]	[0.825]	[0.870]	[0.979]
T3: Promoting waste sorting rather than donations	-0.568***	-0.576***	-0.295***	-0.262*	-0.262*	-0.132
	(0.161)	(0.161)	(0.081)	(0.155)	(0.155)	(0.087)
	[0.001]	[0.000]	[0.000]	[0.090]	[0.092]	[0.128]
Order (1/2)		-0.123	-0.091		-0.154	-0.081
		(0.114)	(0.057)		(0.111)	(0.062)
Constant			3.693***			3.478***
			(0.103)			(0.107)
Multiple-Hypothesis Testing	0.003	0.003	0.002			
Number of Observations	1052	1052	1052	1094	1094	1094
Pseudo R-squared / R-squared	0.006	0.006	0.018	0.001	0.002	0.005

Notes: This table reports the estimates from subsample analyses of Table 8. The coefficients are reported. Standard errors are in parentheses in columns 1, 2, 4 and 5. Robust standard errors are in parentheses in columns 3 and 6. The randomization-t p-values are in brackets. Inference in each column is based on a randomization inference procedure of Young (2019). ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. The row of Multiple-Hypothesis Testing reports the randomization-t p-values for the multiple-hypothesis testing test based on a randomization inference procedure of Young (2019), which applies the procedure of Westfall and Young (1993). For example, column 1 reports the result under the null hypothesis that all treatment effects within the two regressions of the same model for women (column 1) and men (column 4) are zero, adjusting for multiple-hypothesis testing.

Table 11: Results of the randomized survey experiment on the effect of changing the implementer of the experiment

	Allcott (2011)		Hanna et al. (2016)		Hosono and Aoyagi (2018)	
	(1) Ologit	(2) OLS	(3) Ologit	(4) OLS	(5) Ologit	(6) OLS
Company	0.097 (0.113) [0.399]	0.044 (0.061) [0.485]				
Nonprofit organization			-0.226** (0.113) [0.046]	-0.112* (0.060) [0.064]		
International development agency					-0.043 (0.114) [0.711]	-0.010 (0.057) [0.862]
Order (1-6)	-0.152*** (0.033)	-0.083*** (0.017)	-0.080** (0.033)	-0.039** (0.017)	-0.079** (0.034)	-0.038** (0.017)
Constant		3.238*** (0.076)		3.257*** (0.074)		3.212*** (0.072)
Number of Observations	1051	1051	1053	1053	1054	1054
Pseudo R-squared / R-squared	0.008	0.021	0.003	0.008	0.002	0.005

Notes: This table reports the estimates from regression analyses that use a subsample of the first survey. The dependent variable is the response to the question "Do you recognize any ethical issues in this study?" on a five-point scale (1–5). Columns 1 and 2 report the results for the survey on two types of descriptions of Allcott (2011). Randomly assigned respondents are shown a description similar to that in Appendix A1.3, but the implementer of the project is a "company" instead of Professor X. Columns 3 and 4 report the results on Hanna et al. (2016), where the randomly assigned half of the respondents are shown "NPO" instead of Professor X. Columns 5 and 6 report the results on Hosono and Aoyagi (2018), where the randomly assigned half of the respondents are shown "international development agency" instead of Professor X. The coefficients are reported. Standard errors are in parentheses in columns 1, 3 and 5. Robust standard errors are in parentheses in columns 2, 4, and 6. The randomization-t p-values are in brackets. Inference in each column is based on a randomization inference procedure of Young (2019). ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

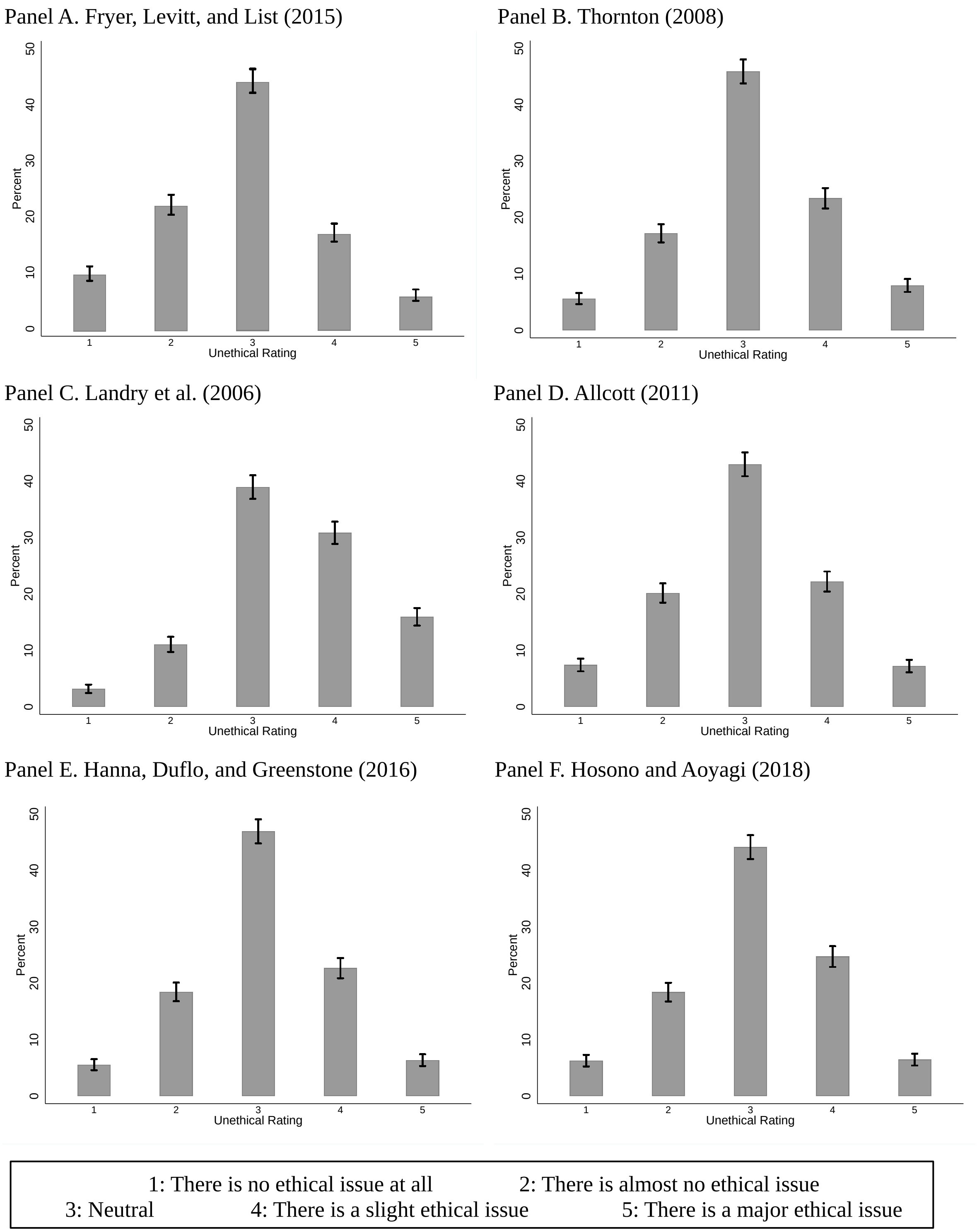
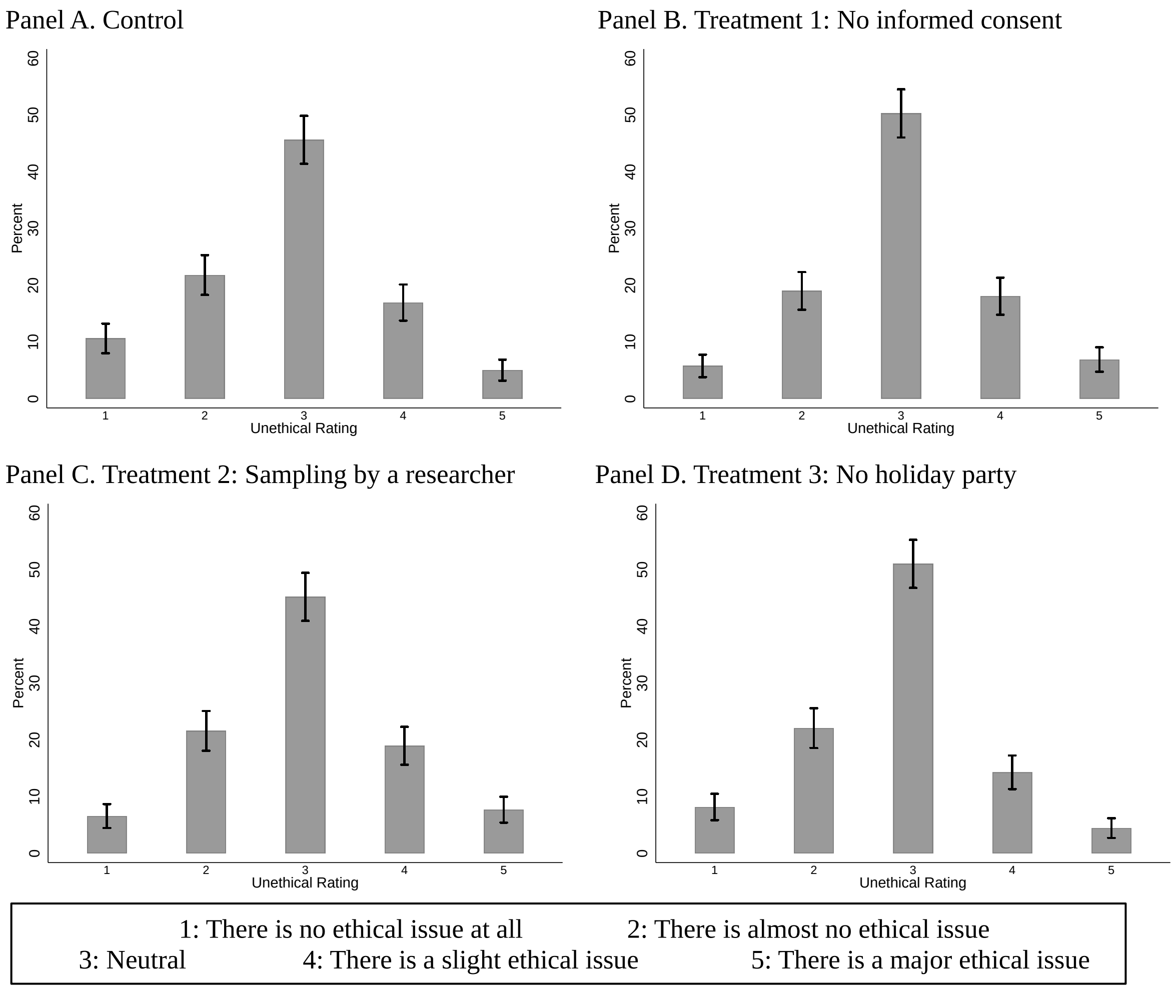


Figure 1: Response to the first survey

Notes: This figure shows the distribution (percentages) of the survey response to the question “Do you recognize any ethical issues in this study?” The vertical bars and caps are 95 % confidence intervals. The average number of observations for the six questions is 1,053.5.



Panel A. Control

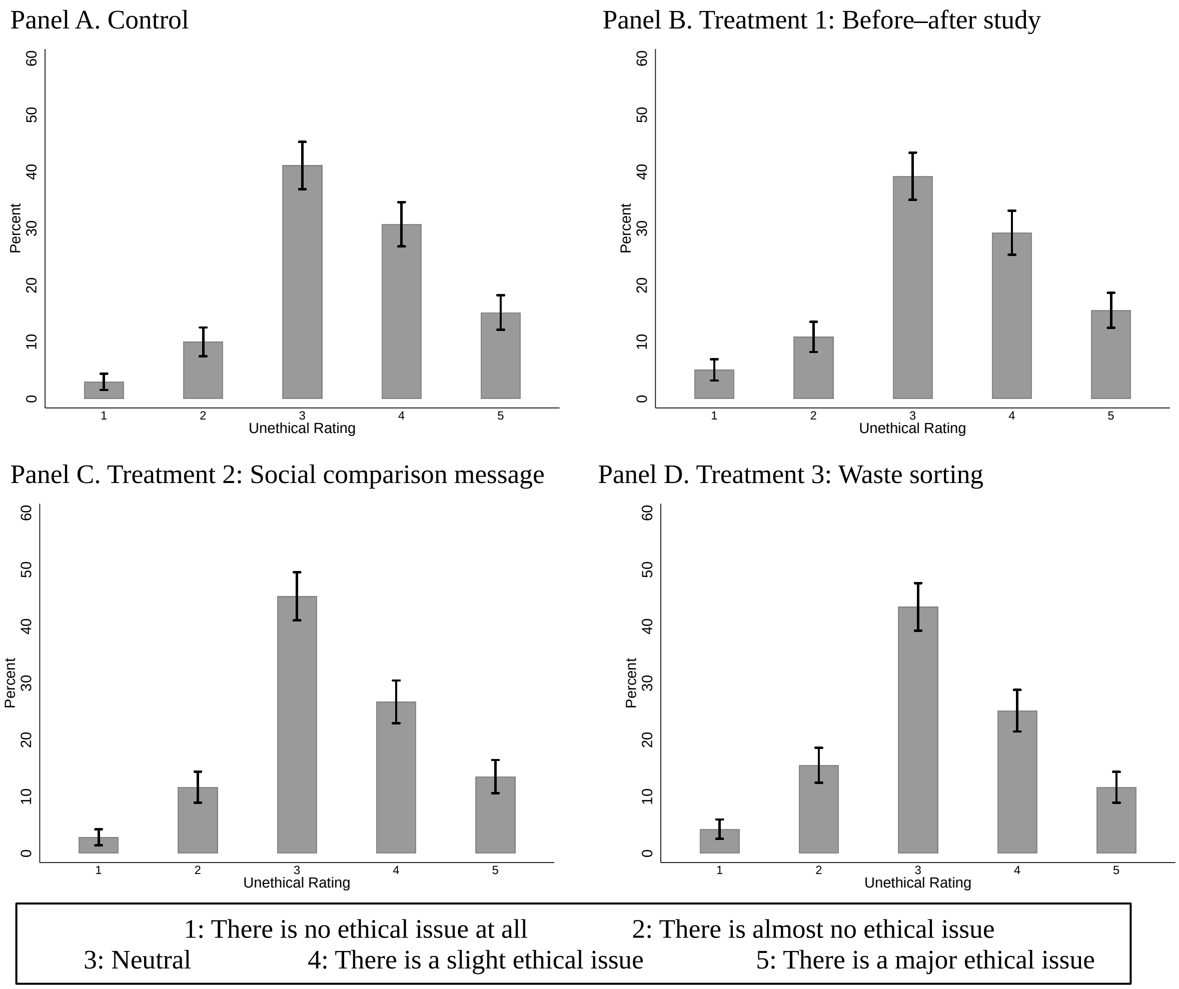
Panel B. Treatment 1: No informed consent

Panel C. Treatment 2: Sampling by a researcher

Panel D. Treatment 3: No holiday party

Figure 2: Response to the randomized survey (Fryer, Levitt, and List, 2015)

Notes: This figure shows the distribution (percentages) of the survey response to the question "Do you recognize any ethical issues in this study?" for the four descriptions based on Fryer, Levitt, and List (2015). The vertical bars and caps are 95 % confidence intervals.



Panel A. Control

Panel B. Treatment 1: Before–after study

Panel C. Treatment 2: Social comparison message

Panel D. Treatment 3: Waste sorting

Figure 3: Response to the randomized survey (Landry et al., 2006)

Notes: This figure shows the distribution (percentages) of the survey response to the question “Do you recognize any ethical issues in this study?” for the four descriptions based on Landry et al. (2006). The vertical bars and caps are 95 % confidence intervals.

Appendix A. Descriptions used in the first survey

A1. Description of Thornton (2008)

Study on HIV testing
The AIDS epidemic is a serious issue in the developing world, such as African countries. One of the reasons for this is that there are people who have sexual intercourse with multiple, unspecified partners while unaware of their HIV-positive status.
To prevent the AIDS epidemic, Professor X conducted a campaign in a developing country. The campaign offered a chance of earning reward equivalent to JPY 20 to those who accepted HIV testing and learned the results. In this country, JPY 20 is the amount one can earn doing agricultural work in one day.
The content of the campaign was as follows:

- *HIV testing was free, and anyone could participate.*
- *Of 3,000 adults in a certain region, only 1,500 who were randomly selected by a computer were involved in the campaign.*
- *The remaining 1,500 adults could not take part.*
- *Among the eligible adults, only those who underwent HIV testing and went to testing centers to learn about the results were awarded an amount of money equivalent to JPY 20.*

After the campaign, Professor X surveyed those who were—as well as those who were not—selected for participation in the campaign about whether they had received HIV testing and had gone to testing centers to be notified of the results. Finally, Professor X conducted a study comparing those who were and were not selected for their decision to learn HIV results. Note that the 3,000 adults gave their consent following an explanation that the survey results on their HIV testing status would be used for a certain research objective. However, they were not told that a comparison was made based on the presence/absence of a chance in reward.

A2. Description of Landry et al. (2006)

Study on charitable giving
Researchers sometimes try to obtain donations from citizens to conduct studies that contribute to society. To obtain more donations, Professor X came up with the idea of "offering donors a chance to win a prize in a raffle."
To examine whether this idea actually increases the amount of donations, Professor X conducted a fundraising project for "Natural Hazard Mitigation Research" in an area with 4,800 households.
Overview of the project:

- *Donations were collected through door-to-door visits for all households.*
- *Only 2,400 households who were randomly selected by a computer were asked to donate with a raffle in which one among all donors could win JPY 100,000 (households with a prize).*
- *The remaining 2,400 households were asked to donate without the chance of winning in the raffle (households without a prize).*

- *One of the "households with a prize" that actually donated money won JPY 100,000 in the raffle.*
- *Collected donations were actually used for "Natural Hazard Mitigation Research."*

After the fundraising activities, Professor X counted all the donations from both the "households with a prize" and "households without a prize" and then compared the two groups. Note that the 4,800 households that were solicited for donations were not informed of their involvement in the study.

A3. Description of Allcott (2011)

Study on electricity conservation
Professor X prepared reports that present effective strategies to reduce the electricity bill of households. These "Home Energy Reports" contain data on average household electricity usage and tips to conserve it. To examine the effectiveness of these reports in terms of reducing electricity consumption, a project was run in an area populated by 40,000 households.
Overview of the project:

- *The "Home Energy Reports" were mailed to only 20,000 households that were randomly selected by a computer.*
- *The reports were mailed once a month for three months (3 times total).*
- *The remaining 20,000 households did not receive the reports at all.*

Four months after the manuals started to be mailed out, an electric utility company compared usage between households who were mailed the manuals three times and those to whom the manuals were not mailed. Note that the 40,000 households were not informed of their involvement in the study.

A4. Description of Hanna et al. (2016)

Study on smoke from cooking stoves
In many developing countries, each household has a stove for cooking. Note that those stoves emit smoke during cooking. This has become a social problem since women and children become sick by inhaling smoke.
To tackle this problem, Professor X carried out a project in a developing country, whereby "improved cooking stoves" were constructed free of charge in an area with 1,600 households.
Overview of the project:

- *Stove construction was carried out over two periods over 5 years.*
- *For the first three years, stoves were built for only 800 households selected by a lottery.*
- *The remaining 800 households awaited their turn.*

At the end of the third year, Professor X observed the health status of both for those whom "improved cooking stoves" were constructed and for those whom these stoves had not been built. Finally, Professor X compared the health status of the two groups. Note that the 1,600 households were informed that they were involved in a study. However, they were not informed that they were compared based on the presence/absence of the improved stove.

A5. Description of Hosono and Aoyagi (2018)

Study on recyclable waste sorting
An increasing amount of waste has become a social problem. To tackle this problem, it has been suggested to decrease the amount of waste by sorting and recycling it.
Professor X carried out a project using stamps and gifts in a poor region populated by 500 households in a developing country to increase household waste sorting.
Overview of the project:

- *Households can get a stamp on their card if recyclable waste is sorted upon disposal.*
- *Households were gifted with laundry detergent if they gathered a certain number of stamps.*
- *Due to budgetary reasons, cards for collecting stamps were distributed to only 250 households that were randomly selected by a computer.*
- *The remaining 250 households could not participate in collecting stamps.*

Two months after the stamp collection began, Professor X surveyed household waste disposed of by those who received the stamp card and who did not. Then, Professor X examined whether they had sorted recyclable waste from other waste. Furthermore, Professor X measured the weight of waste by type (cans, plastic, etc.) and compared the weight between the two groups. Note that the 500 households received an explanation regarding their involvement and were informed that their household waste was weighed. However, they were not informed that they were compared based on whether they had the stamp card.

Appendix B. Descriptions used in the second survey

2.1. Description of Treatment 2 in the survey for Fryer et al. (2015)

Study on a preschool
Recent findings show that the care and education one receives in early childhood affect one's academic achievement and lifetime earnings in adulthood. Following these findings, Professor X established a preschool in a low-income area.
Overview of the preschool:

- *The preschool is free of charge.*
- *This preschool uses a curriculum called "Tools of the Mind" to foster patience and social skills.*
- *Inside the preschool is similar to a small "town," where one can experience various types of jobs.*
- *Children of this preschool are surveyed periodically.*
- *Followup surveys are planned for every few years following graduation.*

Overview of the selection:

- *Parents and their children from 140 families living in the area are defined as the research subjects.*
- *Only 70 children selected based on a lottery were admitted to enroll in the preschool.*
- *The remaining 70 children were not able to enroll in the preschool.*
- *However, the children who were not able to enroll, as well as their parents, were regularly invited to parties held on holidays.*

After the preschool was opened, Professor X invited the children who were enrolled and their parents—as well as the children who were not able to enroll and their parents—to regularly held parties and surveyed them. The surveys were periodically conducted for over 10 years, even after the children entered primary school. Finally, Professor X conducted a study comparing children who attended the preschool with those who were not able to enroll. Note that the parents of the 140 children who became subjects of the study received an explanation regarding them being the subjects of the study, and they gave their consent.

2.2. Description of Treatment 3 in the survey for Fryer et al. (2015)

Study on a preschool
Recent findings show that the care and education one receives in early childhood affect one's academic achievement and lifetime earnings in adulthood. Following these findings, Professor X established a preschool in a low-income area.
Overview of the preschool:

- *The preschool is free of charge.*
- *This preschool uses a curriculum called "Tools of the Mind" to foster patience and social skills.*
- *Inside the preschool is similar to a small "town," where one can experience various types of jobs.*
- *Children of this preschool were surveyed periodically.*
- *Followup surveys were planned for every few years following graduation.*

Professor X called for applicants to this preschool.
Overview of admissions:

- *Parents and their children in 140 families applied for admission.*
- *Only 70 children selected based on a lottery were admitted.*
- *The remaining 70 children were not able to enroll in the preschool.*

After the preschool was opened, Professor X conducted periodical surveys for the children who were enrolled and their parents as well as the children who were not able to enroll and their parents. The surveys were periodically conducted for over 10 years, even after the children entered primary school. Finally, Professor X conducted a study comparing children who attended the preschool with those who were not able to enroll. Note that the parents of the 140 children who became subjects of the study received an explanation regarding them being the subjects of the study, and they gave their consent.

2.3. Description of Treatment 1 in the survey for Landry et al. (2006)

Study on charitable giving
Researchers sometimes try to obtain donations from citizens to conduct studies that contribute to society. To obtain more donations, Professor X came up with the idea of "offering donors a chance to win a prize in a raffle."
To examine whether this idea actually increases the amount of donations, Professor X conducted a fundraising project for "Natural Hazard Mitigation Research" in an area with 4,800 households.
Overview of the project:

- *Donations were collected through door-to-door visits for all households.*

- *One year later, donations were again collected through door-to-door visits.*
- *In the door-to-door visits in the first year, the widely used practice of solicitation was used (no prize phase).*
- *In the visits in the second year, households were asked to donate with a raffle in which one donor could win JPY 100,000 (prize phase).*
- *One of the households that actually donated in the "prize phase" won JPY 100,000 in the raffle.*
- *Collected donations were actually used for "Natural Hazard Mitigation Research."*

After the fundraising activities, Professor X counted all the donations in both the "no prize phase" and "prize phase" and then compared the two phases. Note that the 4,800 households that were solicited for donations were not informed of their involvement in the study.

2.4. Description of Treatment 2 in the survey for Landry et al. (2006)

Study on charitable giving
Researchers sometimes try to obtain donations from citizens to conduct studies that contribute to society. To obtain more donations, Professor X came up with the idea of "telling donors the result of solicitation in neighboring town."
To examine whether this idea actually increases the amount of donations, Professor X conducted a fundraising project for "Natural Hazard Mitigation Research" in an area with 4,800 households.
Overview of the project:

- *Donations were collected through door-to-door visits for all households.*
- *Only 2,400 households who were randomly selected by a computer were asked to donate through a flyer that states the following: "In the neighboring town, 80% of the households donated" (households with a message).*
- *The remaining 2,400 households were asked to donate using a flyer without this message (households without a message).*
- *Collected donations are actually used for "Natural Hazard Mitigation Research."*

After the fundraising activities, Professor X counted all the donations from both the "households with a message" and "households without a message" groups and then compared them. Note that the 4,800 households that were solicited for donations were not informed of their involvement in the study.

2.5. Description of Treatment 3 in the survey for Landry et al. (2006)

Study on recyclable waste sorting
The increasing cost of solid waste management has become a social problem. To tackle this problem, it has been suggested to separate food waste from garbage and recycle it to decrease the amount of waste. To increase sorting food waste at home, Professor X came up with the idea of "offering recyclers a chance to win a prize in a raffle."
To examine whether this idea actually increases the amount of food waste sorted, Professor X conducted a recycling campaign project in an area with 4,800 households.
Overview of the project:

- *In cooperation with the city government, the municipal collection of food waste separately from garbage was begun.*

- *Sorting of food waste was solicited through door-to-door visits for all households.*
- *Only 2,400 households who were randomly selected by a computer were asked to sort with a raffle in which one among all recyclers could win JPY 100,000 (households with a prize).*
- *The remaining 2,400 households were asked to sort without the chance of winning in the raffle (households without a prize).*
- *One of the "households with a prize" that actually sorted food waste won JPY 100,000 in the raffle.*
- *Collected food waste was composted and used by farmers in the area.*

After the campaign, Professor X measured the amount of sorted waste for both the "households with a prize" and "households without a prize" groups and then compared them. Note that the 4,800 households that were involved in the sorting campaign were not informed of their involvement in the study.

Online Appendix 1. Additional Tables

Table A1: Survey response time and characteristics (OLS)

Dependent variable: Time spent on the first survey (seconds)		
	(1)	(2)
Female	-655.751	-1022.648
	(713.732)	(648.026)
Age	21.759	8.542
	(21.392)	(22.982)
Married	-379.050	-547.440
	(425.457)	(419.803)
Living with children	1200.081**	1251.696**
	(585.969)	(568.889)
Full-time employee	660.288	716.329
	(648.012)	(627.491)
Part-time employee	1553.957	2540.860*
	(1048.779)	(1325.376)
Temporary/contract employee	-318.174	232.391
	(524.304)	(561.940)
Self-employed	-154.851	679.092
	(585.592)	(621.967)
Housewife/househusband	1393.794	1736.837
	(1106.674)	(1069.023)
Unemployed/retired	-550.859	13.776
	(414.779)	(579.761)
Household income		0.309
(10 thousand yen)		(1.142)
Constant	-1087.996	-514.549
	(1019.103)	(737.741)
Province dummy variables	Yes	Yes
Observations	2107	1645
Pseudo-R^2/ R^2	0.033	0.051

Notes: This table reports the estimates from linear regression analyses on the association between time spent for the first survey and characteristics of respondents. The median of the dependent variable is 205 seconds (3.4 minutes), and the average is 1,374 seconds (23 minutes). The coefficients are reported. Robust standard errors are in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Table A2: Summary statistics of the randomized survey experiments by group for Table 11

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Allcott (2011)			Hanna et al. (2016)			Hosono and Aoyagi (2018)		
	Prof. X	Company	*P*-value (1) vs (2)	Prof. X	NPO	*P*-value (4) vs (5)	Prof. X	IDA	*P*-value (7) vs (8)
Female	0.482	0.459	0.309	0.480	0.483	0.899	0.482	0.466	0.486
	(0.500)	(0.499)		(0.500)	(0.500)		(0.500)	(0.499)	
Age	46.607	47.411	0.209	46.701	46.367	0.601	46.633	47.120	0.447
	(14.065)	(14.024)		(14.062)	(14.068)		(14.034)	(14.375)	
Married	0.610	0.598	0.596	0.608	0.614	0.817	0.611	0.589	0.337
	(0.488)	(0.491)		(0.488)	(0.487)		(0.488)	(0.492)	
Living with children	0.379	0.381	0.932	0.378	0.398	0.360	0.378	0.388	0.670
	(0.485)	(0.486)		(0.485)	(0.490)		(0.485)	(0.488)	
Household income	534.991	538.564	0.781	534.529	543.525	0.480	534.022	549.268	0.235
(10 thousand yen)	(249.133)	(249.175)		(249.475)	(245.305)		(248.779)	(252.662)	
Full-time employee	0.251	0.229	0.265	0.247	0.269	0.261	0.252	0.215	0.061
	(0.433)	(0.420)		(0.431)	(0.444)		(0.434)	(0.411)	
Part-time employee	0.125	0.107	0.211	0.124	0.125	0.935	0.125	0.108	0.260
	(0.331)	(0.309)		(0.329)	(0.331)		(0.331)	(0.311)	
Temporary/contract employee	0.051	0.059	0.429	0.052	0.049	0.787	0.052	0.053	0.871
	(0.220)	(0.236)		(0.222)	(0.217)		(0.221)	(0.225)	
Self-employed	0.055	0.063	0.444	0.056	0.047	0.391	0.055	0.059	0.722
	(0.228)	(0.243)		(0.230)	(0.213)		(0.228)	(0.236)	
Living in Tokyo	0.126	0.112	0.368	0.125	0.117	0.591	0.125	0.118	0.614
	(0.332)	(0.316)		(0.331)	(0.322)		(0.331)	(0.323)	
Living in Osaka	0.072	0.069	0.774	0.071	0.081	0.363	0.071	0.080	0.447
	(0.258)	(0.253)		(0.256)	(0.274)		(0.257)	(0.271)	

Notes: This table reports the means for each of the groups in the analyses of Table 11. Standard deviations are in parentheses. Columns 3, 6 and 9 report *p*-values for the differences.

Table A3: Comparisons of the six studies for respondents who spent a long time on the survey (coefficients)

	Ordered logit		OLS	
	(1)	(2)	(3)	(4)
Fryer et al. (2015)	-0.453***	-0.443***	-0.248***	-0.241***
	(0.095)	(0.095)	(0.053)	(0.052)
Thornton (2008)	0.055	0.054	0.045	0.045
	(0.092)	(0.092)	(0.051)	(0.051)
Landry et al. (2006)	0.720***	0.729***	0.411***	0.412***
	(0.093)	(0.092)	(0.052)	(0.051)
Allcott (2011)	-0.292***	-0.294***	-0.145**	-0.150***
	(0.102)	(0.101)	(0.057)	(0.056)
Hanna et al. (2016)	0.072	0.079	0.045	0.049
	(0.116)	(0.116)	(0.065)	(0.064)
Order (1–6)	-0.075***	-0.076***	-0.039***	-0.040***
	(0.013)	(0.014)	(0.008)	(0.007)
Female		0.328***		0.184***
		(0.078)		(0.043)
Age		0.013***		0.007***
		(0.003)		(0.001)
Constant			3.224***	2.787***
			(0.051)	(0.092)
Control *implementers*	Yes	Yes	Yes	Yes
Other control variables	No	Yes	No	Yes
Observations	5670	5670	5670	5670
Pseudo-R^2/ R^2	0.015	0.022	0.043	0.063

Notes: This table reports the estimates from the same regression analyses as Table 3 but dropping samples where time spent on the survey is in the bottom 10% (shorter than 49 seconds). Five studies are compared to Hosono and Aoyagi (2018). The coefficients are reported. Standard errors, clustered at the respondent level, are in parentheses. Columns 2 and 4 include other variables in Table 2 as well as *Time spent on the survey*. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Table A4: Results of the randomized survey experiment with interactions over time (Fryer, Levitt, and List, 2015)

	(1)	(2)	(3)	(4)
	Ordered logit		OLS	
T1: Deleting the informed consent statement	0.019	0.017	0.053	0.052
	(0.157)	(0.158)	(0.072)	(0.072)
	[0.886]	[0.899]	[0.466]	[0.481]
T2: Samples are selected rather than self-selection	-0.036	-0.040	0.023	0.019
	(0.154)	(0.155)	(0.073)	(0.073)
	[0.778]	[0.752]	[0.760]	[0.796]
T3: Deleting holiday parties statement	-0.207*	-0.207*	-0.077	-0.079
	(0.152)	(0.152)	(0.067)	(0.067)
	[0.092]	[0.094]	[0.251]	[0.241]
Long time	-0.938***	-0.939***	-0.397***	-0.397***
	(0.162)	(0.162)	(0.084)	(0.084)
T1 × Long time	0.644***	0.647***	0.242**	0.242**
	(0.227)	(0.227)	(0.116)	(0.116)
	[0.006]	[0.006]	[0.036]	[0.037]
T2 × Long time	0.633***	0.637***	0.254**	0.256**
	(0.230)	(0.230)	(0.121)	(0.121)
	[0.008]	[0.006]	[0.036]	[0.034]
T3 × Long time	0.385*	0.386*	0.147	0.148
	(0.226)	(0.226)	(0.116)	(0.116)
	[0.098]	[0.096]	[0.204]	[0.201]
Order (1/2)		-0.195**		-0.077*
		(0.080)		(0.041)
Constant			3.042***	3.157***
			(0.052)	(0.079)
Multiple-Hypothesis Testing	0.031	0.032	0.162	0.153
Observations	2146	2146	2146	2146
Pseudo-R^2/ R^2	0.011	0.012	0.025	0.026

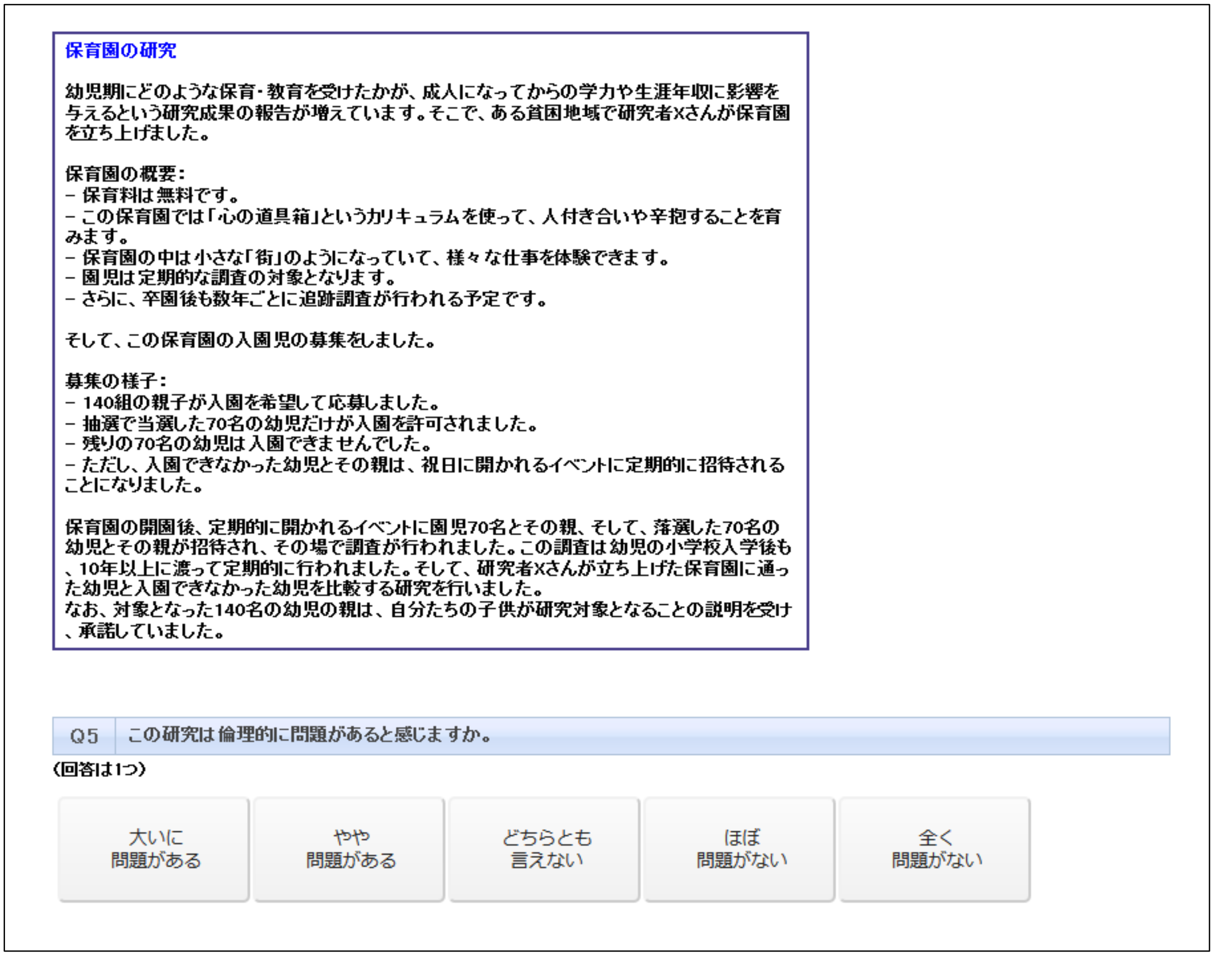
Notes: This table reports the estimates from regression analyses where the dummy variable of time spent on the survey being longer than the median is incorporated into the analysis of Table 7. The coefficients are reported. Standard errors are in parentheses in columns 1 and 2. Robust standard errors are in parentheses in columns 3 and 4. The randomization-*t* *p*-values in brackets. Inference in each column is based on a randomization inference procedure of Young (2019). ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. The row of *Multiple-Hypothesis Testing* reports the randomization-*t* *p*-values for multiple-hypothesis testing computed based on a randomization inference procedure of Young (2019), which applies the procedure of Westfall and Young (1993). It tests the null hypothesis that all treatment effects in each equation (each column) are zero.

Table A5: Results of the randomized survey experiment with interactions with time (Landry et al., 2006)

	(1)	(2)	(3)	(4)
	Ordered logit		OLS	
T1: Before–after study without control	0.015	0.017	0.006	0.007
	(0.153)	(0.153)	(0.072)	(0.072)
	[0.909]	[0.901]	[0.939]	[0.921]
T2: Treatment is a message rather than a raffle	-0.143	-0.145	-0.067	-0.069
	(0.152)	(0.152)	(0.069)	(0.069)
	[0.266]	[0.259]	[0.338]	[0.320]
T3: Promoting waste sorting rather than donations	-0.317**	-0.325**	-0.158**	-0.164**
	(0.154)	(0.154)	(0.072)	(0.071)
	[0.014]	[0.011]	[0.027]	[0.021]
Long time	0.762***	0.751***	0.346***	0.340***
	(0.157)	(0.157)	(0.081)	(0.081)
T1 × Long time	-0.136	-0.128	-0.101	-0.097
	(0.224)	(0.225)	(0.121)	(0.121)
	[0.547]	[0.570]	[0.402]	[0.421]
T2 × Long time	-0.029	-0.015	-0.015	-0.006
	(0.222)	(0.222)	(0.115)	(0.115)
	[0.896]	[0.947]	[0.895]	[0.954]
T3 × Long time	-0.149	-0.138	-0.082	-0.076
	(0.223)	(0.223)	(0.117)	(0.117)
	[0.506]	[0.534]	[0.474]	[0.508]
Order (1/2)		-0.125		-0.085**
		(0.079)		(0.042)
Constant			3.271***	3.403***
			(0.050)	(0.081)
Multiple-Hypothesis Testing	0.072	0.057	0.126	0.100
Observations	2146	2146	2146	2146
Pseudo-R^2/ R^2	0.015	0.015	0.029	0.031

Notes: This table reports the estimates from regression analyses where the dummy variable of time spent on the survey being longer than the median is incorporated into the analysis of Table 8. The coefficients are reported. Standard errors are in parentheses in columns 1 and 2. Robust standard errors are in parentheses in columns 3 and 4. The randomization-*t* *p*-values in brackets. Inference in each column is based on a randomization inference procedure of Young (2019). ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. The row of *Multiple-Hypothesis Testing* reports the randomization-*t* *p*-values for multiple-hypothesis testing computed based on a randomization inference procedure of Young (2019), which applies the procedure of Westfall and Young (1993). It tests the null hypothesis that all treatment effects in each equation (each column) are zero.

Online Appendix 2. Additional Figures



保育園の研究

幼児期にどのような保育･教育を受けたかが、成人になってからの学力や生涯年収に影響を与えるという研究成果の報告が増えています。そこで、ある貧困地域で研究者Xさんが保育園を立ち上げました。

保育園の概要：
- 保育料は無料です。
- この保育園では「心の道具箱」というカリキュラムを使って、人付き合いや辛抱することを育みます。
- 保育園の中は小さな「街」のようになっていて、様々な仕事を体験できます。
- 園児は定期的な調査の対象となります。
- さらに、卒園後も数年ごとに追跡調査が行われる予定です。

そして、この保育園の入園児の募集をしました。

募集の様子：
- 140組の親子が入園を希望して応募しました。
- 抽選で当選した70名の幼児だけが入園を許可されました。
- 残りの70名の幼児は入園できませんでした。
- ただし、入園できなかった幼児とその親は、祝日に開かれるイベントに定期的に招待されることになりました。

保育園の開園後、定期的に開かれるイベントに園児70名とその親、そして、落選した70名の幼児とその親が招待され、その場で調査が行われました。この調査は幼児の小学校入学後も、10年以上に渡って定期的に行われました。そして、研究者Xさんが立ち上げた保育園に通った幼児と入園できなかった幼児を比較する研究を行いました。
なお、対象となった140名の幼児の親は、自分たちの子供が研究対象となることの説明を受け、承諾していました。

Q5 この研究は倫理的に問題があると感じますか。

（回答は1つ）

Figure A1: Screenshot of the survey

Notes: This figure shows a screenshot of the survey. Fryer, Levitt, and List (2015) is shown (in Japanese).

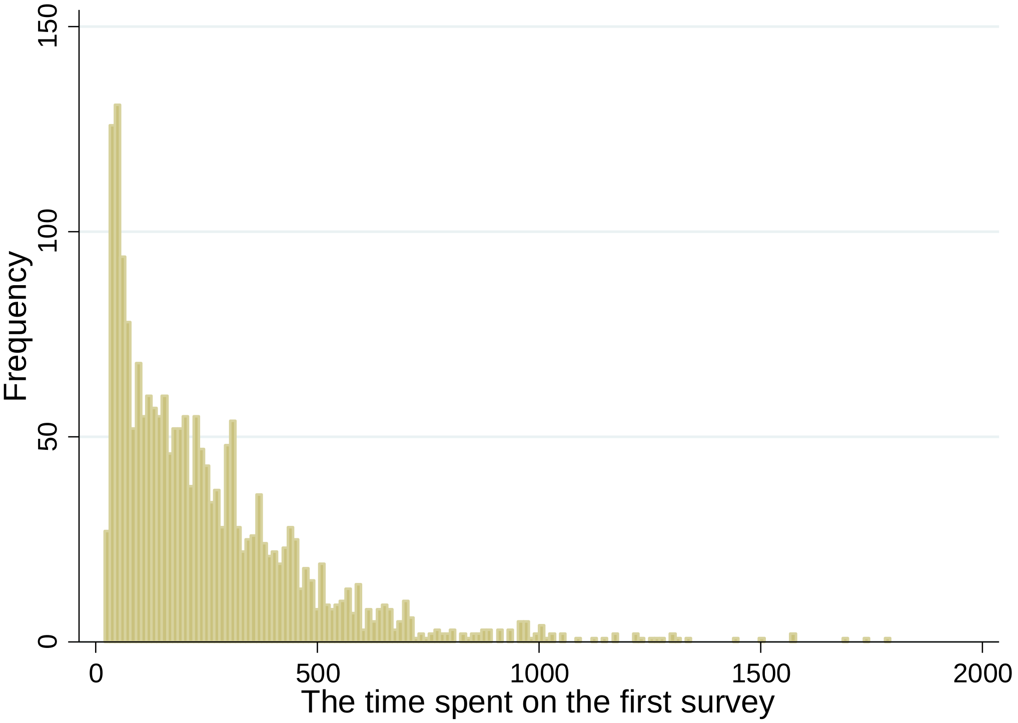


Figure A2: Distribution of time spent on the first survey

Notes: This figure shows the distribution of the time spent on the first survey. The vertical axis shows the density. The horizontal axis shows the time in seconds. This figure only shows the distribution shorter than 1,800 seconds, while the maximum value was 184,468 seconds.