UNIVERSITA' COMMERCIALE "LUIGI BOCCONI" PhD SCHOOL

PhD program in Business Administration and Management

Cycle: 33°

Disciplinary Field (only the code): SECS-P/07

Consumer Aversions Revisited: From Mechanisms to Solutions

Advisor: Joachim VOSGERAU

PhD Thesis by

Burcak BAS

ID number:

1822162

Year 2022

CHAPTER 1

WHY DO CONSUMERS CONDEMN AND APPRECIATE EXPERIMENTS?

Burcak Bas*

*Doctoral student and corresponding author

Department of Marketing - Bocconi University, Milan, Italy

Email. burcak.bas@unibocconi.it

Rachele Ciulli

The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Email. ciulli@wharton.upenn.edu

Joachim Vosgerau

Department of Marketing - Bocconi University, Milan, Italy

Email. joachim.vosgerau@unibocconi.it

WHY DO CONSUMERS CONDEMN AND APPRECIATE EXPERIMENTS?

ABSTRACT

Field experiments have surged in recent years in marketing research and practice. Whether consumers approve of such experiments, however, is a contested topic. One stream of research maintains that consumers object to experiments but approve of the implementation of an experiment's untested treatments (Meyer et al., 2019a). Another stream of research claims that consumers view experiments run by corporations as acceptable as their worst treatment, but never as less acceptable (Mislavsky, Dietvorst, & Simonsohn, 2020). We reconcile these contradictory findings and explain when consumers are averse to experiments and when they appreciate them. When consumers strongly believe in the existence of normative standards of best practice, conducting an experiment is perceived as a violation of the standard and the experiment is evaluated as *worse* than the implementation of *either* of its treatments. When an experiment is instead designed to discover a new standard of best practice, conducting an experiment is evaluated as *better* than the implementation of *either* of its treatments (i.e., experiment appreciation). Eight preregistered studies (N = 2,676) provide empirical support for our account of experiment aversion and appreciation, offering guidelines on how to communicate experiments to consumers.

KEYWORDS

experiments, evidence-based practices, ethics, marketing communications, public opinion, research

INTRODUCTION

Field experiments have surged in recent years in marketing research, marketing practice, and in the social sciences in general (Nelson, Simester, & Sudhir, 2020; Buyalskaya, Gallo, & Camerer, 2021). Field experiments help avoid concerns about endogeneity and selection and allow for causal inferences that are fundamental for the design of interventions in marketing and elsewhere (Gneezy, 2017; Lambrecht & Tucker, 2018; Zoumpoulis, Simester, & Evgeniou, 2015). For example, in the marketing field, experiments have been used to explore what influences the effectiveness of online advertising (Goldfarb & Tucker, 2011), to test whether email personalization can make email campaigns more effective (Munz, Jung, & Alter, 2020), to test whether social advertising is effective across product categories (Huang, Aral, Hu, & Brynjolfsson, 2020), or to test the validity of company-provided field experimental tools such as A/B testing on Facebook (Gordon, Zettelmeyer, Bhargava, & Chapsky 2019).

The development and implementation of evidence-based practices hinges on the extent to which experimentation on consumers is accepted by consumers (Meyer, 2014). In the last two years, two streams of research have come to contradicting conclusions. One stream of research suggests that consumers disapprove of experiments in a wide range of domains (Heck, Chabris, Watts, & Meyer, 2020; Meyer, 2015, Meyer et al., 2019a; 2019b). This so-called "experiment aversion" is puzzling as consumers have been shown to judge an experiment that tests the comparative effectiveness of two treatments as inappropriate while at the same time judging the implementation of either of its treatments as appropriate. Another stream of research maintains that consumers view experiments run by corporations on their customers as acceptable as their worst treatment but never as less acceptable (Mislavsky, Dietvorst, & Simonsohn 2019; 2020).

We propose and test a theory to reconcile these contradictory findings. According to our theory, consumers can condemn and appreciate almost identical experiments depending on their lay beliefs about normative standards of best practice. When a normative standard is believed to exist in a given domain, running an experiment is perceived as violating the standard and consumers disapprove of it (i.e., experiment aversion). When instead an experiment is designed to discover or improve a normative standard of best practice, consumers appreciate it (i.e., experiment appreciation). Our findings provide guidelines for promoting experiments to the public.

ACCEPTABILITY OF EXPERIMENTS

Experiment Aversion

Recent research by Meyer (2015), Meyer et al. (2019a), and Heck, Chabris, Watts, and Meyer (2020) suggests that consumers are generally averse to experiments in a wide range of domains. For example, respondents were asked to evaluate a charity's plan to help people in a low-income country escape poverty. The charity considered two options, either providing adults below a certain income level a sturdy roof for their homes, or offering adults below a certain income level a one-month training in a trade of their choice. Respondents evaluated both options as equally appropriate. However, respondents disapproved of the charity running an experiment in which half of all adults below a certain income level would be randomly assigned to receive a sturdy roof for their home, and the other half to receive one month of training in a trade of their choice (after a year, the charity would provide everyone in the country whichever resource—roof or training—turned out to help more people escape extreme poverty). The simultaneous disapproval of the experiment and approval of the implementation of either of the experiment's untested treatments was coined "A/B illusion" and "experiment aversion," and is defined as "…the widespread tendency to view a field

experiment designed to study the effects of an existing or proposed practice as more morally suspicious than an immediate, universal implementation of an untested practice" (Meyer, 2015: 278). Experiment aversion is a puzzling phenomenon, to say the least, as it seems irrational to oppose random assignment to treatments when at the same time approving of the implementation of either treatment.

Consumers are not Averse to Experiments, They are Averse to Bad Treatments

Contradicting the claim that consumers are generally averse to experiments, Mislavsky, Dietvorst, and Simonsohn (2020) demonstrated that consumers evaluate corporate experiments as acceptable as their worst treatment but never as less acceptable. The authors call this the "critical condition" hypothesis, according to which the acceptability of an experiment is determined by its least acceptable treatment. For instance, respondents were told that Amazon considered changing the way it recommends products. Specifically, the company contemplated whether to recommend its most profitable or most highly rated products to customers. Respondents rated the former as less acceptable than the latter. And respondents who were told that Amazon decided to run an experiment to test the two recommendation systems by randomly assigning customers to one or the other rated the experiment as acceptable as its worst treatment—recommending the most profitable products—but not worse. The authors argue that consumers are averse to bad treatments but not to experiments testing them. Table 1 provides examples of scenarios in which experiment aversion has been observed and in which it has not been observed.

Experiment Aversion or Critical Condition?

Experiment aversion has been attributed to a wide range of possibilities, both at the conceptual and methodological levels. At the conceptual level, Meyer et al. (2019a) have argued that experiment aversion may be driven by lack of consent, an illusion of knowledge in which respondents believe experts already do or should know "what works", and to

experimentation/randomization per se. However, the authors conclude that none of these possibilities are sufficient to explain all cases in which experiment aversion has been observed. In addition, none of these could explain why experiment aversion occurs in the scenarios investigated by Meyer and colleagues but not in corporate experiments (Mislavsky, Dietvorst, & Simonsohn, 2020).

Mislavsky, Dietvorst, and Simonsohn (2019) argued that a statistical artifact is responsible for Meyer et al.'s (2019a) findings. Specifically, the authors argued that what has been interpreted as experiment aversion may simply reflect heterogeneous preferences. If some respondents dislike treatment A and others dislike treatment B, both groups will dislike an experiment that randomly assigns participants to either A or B. "To illustrate, imagine an experiment giving people a dessert with either dairy (A) or peanuts (B). If 30% of people are lactose-intolerant and another 30% have peanut allergies, A and B are each objectionable to 30% of people. However, the experiment is objectionable to 60%" (Mislavsky, Dietvorst, & Simonsohn, 2019:1). To alleviate this statistical artifact, the authors suggest measuring evaluation of treatments in a within-participant design and comparing the evaluation of an experiment to the minimum evaluation of its treatments.

Not all findings of experiment aversion, however, can be explained by heterogeneous preferences (Mislavsky, Dietvorst, & Simonsohn, 2019). Meyer et al. (2019a), for example, found experiment aversion in the medical domain where a doctor tried out two different drugs to treat high blood pressure. The two drugs are described only as drug A and drug B, making heterogeneity in drug evaluations unlikely. Furthermore, heterogeneous preferences would predict that the proportion of respondents objecting to treatment A and those objecting to treatment B equals the proportion of respondents objecting to the experiment. Meyer et al. (2019a), however, found in most of their studies that the proportion objecting to the experiment exceeded the sum of proportions objecting to the treatments. Finally, Heck,

Chabris, Watts and Meyer (2020) re-ran Meyer et al.'s (2019a) studies following the recommendations by Mislavsky, Dietvorst and Simonsohn (2019) and still found experiment aversion in all scenarios but one.

Concluding, experiment aversion is a robust and well-established phenomenon. Why and when it occurs, however, is unknown.

SOLVING THE PUZZLE:

WHY CONSUMERS CONDEMN AND APPRECIATE EXPERIMENTS

Why do consumers object to some experiments but not to others? Researchers and physicians have speculated that experiments in healthcare may be seen as aversive for three reasons. First, an experiment might reflect an ethical dilemma in which individual patients' rights are sacrificed for potential benefits to society (Hellman & Hellman, 1991). Second, experiments may—due to random assignment to experimental conditions—be perceived as depriving patients of the judgment and medical expertise of physicians (Heck, Chabris, Watts, & Meyer, 2020); Lantos, 2020; Lynch, Greiner, & Cohen, 2020; Meyer et al., 2019a). And finally, experiments might signal lack of physician knowledge about the "best" treatment (Lantos, 2020). All three explanations have in common that experiments—at least in healthcare—may be perceived as unethical.

We propose that such perceptions of unethicality are a consequence of lay beliefs about normative standards of best practice, not only in healthcare but in all domains in which experiments are conducted on consumers. Normative standards of best practice consist of abstract ideas according to which an established set of principles or practices exist that decision-makers/experts are morally obliged to adhere to; they are called "normative" because their adherence is perceived as morally required. Consumers may have little to no knowledge at all what such principles and best practices look like, what is important is that

they believe that they exist (Hellman & Hellman, 1991; Lantos, 2020; Lynch, Greiner, & Cohen, 2020).

Let's consider the best drug scenario by Meyer et al. (2019a) in which a doctor administers either drug A or drug B to patients with hypertension, or runs an experiment in which half of the patients receives drug A and the other half receives drug B. For the treatment of a specific form of hypertension, consumers may believe that there exists one drug or treatment that is superior to all other drugs/treatments, and hence doctors are morally required to prescribe that drug/treatment to provide the best possible treatment to their patients. When consumers strongly believe in the existence of such a normative standard of best practice, running an experiment will be perceived as violating the standard, as the doctor who randomly assigns patients to receive drug A or drug B must be violating the standard of best practice for at least one group of patients. The doctor who prescribes only drug A or only drug B, in contrast, appears to be acting in accordance with the normative standard of best practice for the treatment of hypertension as long as patients receive only one treatment. Consequently, consumers disapprove of the experiment, but they would approve of the implementation of either of its treatments. The same logic may apply to the charity scenario. Consumers may believe that there is a best way to help people in low-income countries escape extreme poverty, and hence the charity who provides only sturdy roofs or only training in a trade of choice appears to be acting in accordance with the normative standard of best practice. But the charity running an experiment will be perceived as violating that standard.

Under which conditions will consumers relinquish their objection to experiments? According to our theorizing there are three possibilities: First, when consumers do not hold strong beliefs in the existence of normative standards of best practice. This, we contend, is the case for many experiments that corporations conduct on their customers (Mislavsky,

Dietvorst, & Simonsohn, 2020; see Table 1). Second, when consumers hold strong beliefs in the existence of standards of best practice but don't believe adherence to these standards is morally required. Third, when it is pointed out that extant normative standards of best practice are not perfect and can be further improved (for example, by conducting an experiment).

It follows from our theorizing that it may even be possible to reverse experiment aversion, that is to make consumers appreciate experiments (i.e., judging an experiment as better than the implementation of either of its treatments). This may happen when experiments are conducted to discover a new normative standard of best practice, for example testing new drugs for the treatment of a disease for which no treatment yet exists. Consumers in these instances may recognize that the required information can only be gleaned from conducting an experiment.

OVERVIEW OF STUDIES

We test our hypotheses in eight preregistered studies (N = 2, 676). Following the extant literature, we operationalize experiment aversion as an experiment being rated as less appropriate than its worst-rated treatment, where appropriateness of treatments is measured within-participants (Heck, Chabris, Watts, & Meyer, 2020, Mislavsky, Dietvorst, & Simonsohn, 2019; 2020). Analogously, we operationalize experiment appreciation as an experiment being rated as more appropriate than its best rated treatment.

Study 1 tests whether participants hold stronger beliefs in normative standards of best practice for scenarios/domains in which experiment aversion has previously been demonstrated. In Study 2A, we identify two new scenarios/domains for which participants hold strong beliefs in normative standard of best practice, and in Study 2B we test whether in these scenarios/domains experiment aversion occurs. Study 3 tests whether experiment

aversion is observed for corporate experiments when a moral obligation to follow standards of best practice is introduced, and Study 4 tests the opposite, whether experiment aversion in a medical scenario disappears when moral obligation to follow standards of best practice is removed. In Study 5, we test whether experiment aversion is caused by perceived violations of normative standards of best practice—as we claim—or by perceived unfairness in treatments of study-participants. Study 6 shows that experiment aversion disappears when participants are informed that extant standards of best practice can be improved, and Study 7 demonstrates experiment appreciation for participants who are informed that the experiment will help to establish a new normative standard of best practice. We conclude with a discussion of how businesses leaders and public policy makers can frame experiments in such a way that consumer do not object to, but actually approve of experiments.

TRANSPARENT REPORTING

All eight studies are preregistered[1]. In all studies, participants were recruited either from the online crowd-sourcing platform MTurk (https://www.mturk.com/) or Prolific (https://www.prolific.co/) in exchange for monetary compensation. Participants were required to have an approval rating of at least 95%, more than 50 previously completed HITs, and to reside in the United States. MTurk workers were recruited through Cloud Research's (https://www.cloudresearch.com/) approved participant pool; CloudResearch provides a user-friendly interface for recruiting and conducting online experiments with MTurk workers. MTurk workers, especially those with approval ratings above 95%, have been shown to provide high quality data (Peer, Vosgerau, & Acquisti, 2014).

[1] In several pre-registrations of our studies, we referred to normative standards of best practice as "golden standard" or "gold standard." However, we later abandoned these terms to avoid confusion with experiments generally being referred to as the "gold standard for hypothesis testing."

We have complied with all relevant ethical regulations and obtained informed consent from all participants. As preregistered, participants who failed an attention check were not allowed to start our studies. All experimental materials, preregistrations, data, codes, and analyses are accessible here: https://researchbox.org/188&PEER_REVIEW_passcode=EQISHC. We use two-sided t-tests in our analyses, and compute Cohen's *d* with the mean and standard deviation of the groups' difference in paired sample t-tests and pooled standard deviation in independent sample t-tests. Additional analyses are provided in the Online Appendix.

STUDY 1

Study 1 tested the hypothesis that people hold stronger beliefs in the existence of a normative standard of best practice for contexts/scenarios in which experiment aversion has previously been documented (versus not).

Method

Sample. Two hundred MTurk workers accessed and completed Study 1 (*female* = 98, M_{age} = 39.50, SD_{age} = 12.04).

Design. Study 1 employed a 2 (experiment aversion: present (3 scenarios) vs. absent (5 scenarios); within-participant) design. We selected 3 scenarios for which Meyer et al. (2019a) had found evidence for experiment aversion: Best drug, Charity, and Teacher wellbeing. We selected 5 scenarios for which no experiment aversion had been previously observed, one from Meyer et al. (2019a; Basic income) and four from Mislavsky, Dietvorst, and Simonsohn (2020; Social networking, Ride-sharing, Shipping, and Product recommendation; cf. Table 1). For each scenario, we formulated a statement expressing the

existence of a normative standard of best practice, which is defined by three characteristics: 1. there is only one treatment that qualifies as best practice, 2. all consumers in question should receive that treatment, and 3. decision-makers are morally obliged to deliver that treatment. For instance, for the Charity scenario, the statement read: "In an ideal world, a citizen in a low-income country would receive the same help from a charity, either a sturdy roof for her home or one month of training in a trade of her choice, no matter which low-income country s/he lived in." In the best drug scenario, the statement read: "In an ideal world, a patient suffering from a specific form of high blood pressure would receive the same treatment, no matter which doctor treated her/him." For the other 6 statements please see the Online Appendix.

Participants were told: "Below are 8 statements. We would like to know whether you believe these statements should be true in an ideal world. Indicate your agreement or disagreement on the scale below each statement." (1 = Strongly disagree, 4 = Neither agree nor disagree, 7 = Strongly agree). Statements were provided in random order.

Results and Discussion. As predicted, participants agreed more with the statements for scenarios in which experiment aversion had previously been documented (all pairwise $\text{scenario}_{\text{experiment aversion}}$ vs. $\text{scenario}_{\text{no experiment aversion}}$ comparisons $t(199) > 3.5$, $p < .001$, $d > 0.25$, see Figure 1 and Table 2).

The results of study 1 suggest that lay beliefs about normative standards of best practice may underlie experiment aversion. Participants held stronger beliefs in the existence of normative standards of best practice for contexts/scenarios in which experiment aversion had previously been documented than for contexts/scenarios in which no experiment aversion had previously been documented.

STUDIES 2A AND 2B

If stronger beliefs in the existence of normative standards of best practice are indeed responsible for experiment aversion to occur, we should be able to predict whether experiment aversion will occur in new scenarios based on how strongly consumers believe in the existence of a normative standard for these scenarios. In study 2A, we tested the strength of such beliefs in two new scenarios, a financial scenario involving investment consultancy, and a judiciary scenario involving sentencing for driving under the influence (DUI). To-date, neither scenario/domain has been used as a test for experiment aversion.

Study 2A

Method

Sample. One hundred and ninety-nine Prolific Academic workers accessed and completed the study (*female* = 119, M_{age} = 33.47, SD_{age} = 11.97).

Design. Participants were provided with four statements expressing beliefs in normative standards of best practice, and—as in Study 1—asked to indicate their agreement with these statements (1= Strongly disagree; 7=Strongly agree). Two of the statements were taken from Study 1, namely the statement for product recommendation and for social networking, for which participants had expressed weak beliefs in the existence of a normative standard. The other two statements were formulated for the two new judiciary and financial advice scenarios.

The statement for the judiciary scenario read "In an ideal world, a first-time driving under the influence (DUI) offender would receive the same punishment, either a $500 monetary fine or community-service for three weekends, no matter which judge is judging him/her." The statement for the financial advice scenario read "In an ideal world, someone who wants to invest in the stock market would receive the same financial advice no matter which financial advisor is advising her/him."

Results. Participants agreed more with the statements regarding the financial and the judiciary scenarios than with statements regarding the product recommendation and social networking scenarios (all pairwise $\text{scenario}_{\text{financial, judiciary}}$ vs. $\text{scenario}_{\text{reference standard}}$ comparisons $t(198) > 5.5$, $p < .001$, $d \geq .40$, see Figure 2 and Table 3). Given the relative strong belief in the existence of a normative standard for both scenarios, we predicted that consumers will judge an experiment testing treatments as worse than the implementation of the treatments (i.e., experiment aversion). Study 2B tested this prediction.

Study 2B

Methods

Sample. Two hundred Prolific Academic workers accessed and completed Study 2B (*female* = 106, M_{age} = 32.12, SD_{age} = 10.27).

Design. In a 2 (evaluation: experiment vs. treatments, between-subject) x 2 (scenario: judiciary vs. financial, within-subject) design, participants read about two scenarios and were asked to evaluate the appropriateness of the decision. For instance, for the evaluation of the judiciary scenario/treatments, participants read "Mr. Smith has been arrested for driving under the influence (DUI). This was Mr. Smith's first arrest ever. The judge wants to make sure that people like Mr. Smith will not drink and drive again, so he decides to sentence Mr. Smith to a $500 monetary fine [community-service for three weekends]. How okay is it for the judge to do this?" (1 = "It's really bad"; 4 = "It's okay"; 7 = "It's really good").

For the evaluation of the financial advice scenario/treatments, participants read "Despite the pandemic, several industries experienced tremendous growth over the past year. Especially internet-based service and technology stocks have soared. Company A and company B are two major internet-based service companies, and their stocks look like excellent investment opportunities for 2021 and beyond. Mr. Jones, a financial advisor, wants

to provide solid investment advice to his clients. He decides to advise his clients to invest in Stock A/B. Stock A/B performed well over the last 12 months."

For participants who evaluated the experiments, the judiciary scenario read "The judge wants to make sure that people like Mr. Smith will not drink and drive again, so he decides to run an experiment by randomly assigning first-time DUI offenders to one of two possible sentences. Half of first-time DUI offenders will be ordered a $500 monetary fine, the other half will be ordered community-service for three weekends. After a year, the judge will determine which sentence is more effective in reducing the likelihood of DUI re-arrests, and will use that sentence for all new first-time DUI offenders."

The financial advice scenario read "Mr. Jones, a financial advisor, thinks of two different ways to provide solid investment advice to his clients, so he decides to run an experiment by randomly assigning his clients to one of two investments. He advises half of his clients to invest in Stock A, and the other half to invest in Stock B. Both stocks have performed well over the last 12 months. After a year, he will only advise new clients to invest in whichever stock had the higher return." Participants rated the experiments on the same scale as the treatments were rated.

Results and Discussion. As predicted, participants rated conducting the experiment as less appropriate than the implementation of its worst-rated treatment. This pattern was observed for both scenarios (judiciary scenario: $t(198) = 3.36$, $p < .001$, $d = .48$; financial scenario: $t(198) = 6.89$, $p < .001$, $d = .97$ (cf., Figure 3).

In both the judiciary and financial advice scenario, participants indicated a strong belief in the existence of a normative standard of best practice (Study 2A) and exhibited experiment aversion (Study 2B). The results provide further evidence that strong beliefs in normative standards cause participants to evaluate conducting an experiment as worse than the implementation of its least rated treatment.

STUDY 3

The finding that consumers are averse to a financial advisor running an experiment suggests that experiment aversion may also be observed in the corporate domain, a domain in which so far no evidence for experiment aversion has been found (Mislavsky, Dietvorst, & Simonsohn, 2020; cf., Table 1). According to our theorizing, consumers should object to corporate experiments if they morally require adherence to normative standards of best practice. We tested this possibility by adapting the best drug scenario (Meyer et al. 2019a; Heck, Chabris, Watts, & Meyer, 2020) to a corporate setting. Specifically, we tested whether a pharmaceutical company who is providing help to patients with diabetes during times of war by sending one of two drugs will be evaluated more favorably than the same company conducting an experiment to test which drugs to send.

Method

Sample. Two hundred Prolific Academic workers accessed and completed Study 3 (*female* = 140, M_{age} = 35.16, SD_{age} = 13.62).

Design. In a 2 (evaluation: experiment vs. treatments) between-participant design, participants were randomly assigned to one of two experimental conditions. Participants were presented with a scenario in which a pharmaceutical company was testing which drug to deliver to patients with diabetes in Ukraine, which at the time of writing had been invaded by Russia.

Participants in all conditions read "The unprecedented war in Ukraine has made it a priority to ensure support for Ukrainian patients with chronic diseases such as diabetes. Imagine a pharmaceutical company is considering which drugs to deliver to Ukrainian diabetes patients. It could either deliver Drug A or Drug B. Both drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating

diabetes."

Participants in the treatments condition then rated the acceptability of the pharmaceutical company administering drug A or administering drug B to its patients ("How okay is it for the pharmaceutical company to deliver Drug A[B]?") on 7-point scales with end points 1 = "It's really bad"; 4 = "It's okay"; 7 = "It's really good."

Participants in the experiment condition read "The pharmaceutical company decides to run an experiment to test which drug to deliver. One randomly selected group of patients is delivered Drug A. Another randomly selected group is delivered Drug B. After two weeks, the pharmaceutical company will only deliver whichever drug has had the best outcomes for the patients." Participants then rated the acceptability of running the experiment ("How okay is it for the pharmaceutical company to do this?") on the same scale as above.

Results and Discussion. In support of our prediction, the experiment (M = 4.75, SD = 1.75) was rated as less appropriate than its least acceptable treatment (M = 6.38, SD = 1.04), $t(198) = 8.0$, $p < .001$, $d = 1.14$.

The results of Study 3 suggest that consumers can object to corporate experiments, namely when the experiment's treatments require a moral obligation to adhere to standards of best practice. By the same logic, removing moral obligation to adhere to standards of best practice should remove experiment aversion for scenarios where it is typically observed. Study 4 was designed to test this possibility.

STUDY 4

Study 4 is again based on the best drug scenario in which Meyer et al. (2019a) and Heck, Chabris, Watts, and Meyer (2020) had documented experiment aversion. Experiment aversion should be attenuated when the moral requirement to adhere to standards of best practice is removed. In the adapted scenario, Doctor Jones treated patients with heart

arrhythmia and administered either drug A or drug B, or ran an experiment in which half of the patients received drug A and the other received drug B. So far, this corresponds to the original best drug scenario by Meyer et al. (2019a). We introduced an additional scenario in which Doctor Jones administered the drugs not to humans but to fruit flies (so-called drosophilia are indeed used to study chemical components' effects on human atrial fibrillation, cf., Zhang et al., 2014). We predicted that, because the moral status of simple organisms does not equal that of humans (Ringach, 2011), Doctor Jones would not be expected to adhere to normative standards of best practice when treating fruit flies, and hence experiment aversion would occur only in the human but not the fruit fly scenarios.

Method

Sample. Four hundred one MTurk workers accessed and completed Study 4 (*female* = 211, M_{age} = 38.81, SD_{age} = 12.75).

Design. Study 4 employed a 2 (moral obligation to adhere to best practice: present vs. absent; between-participant) x 2 (evaluation: experiment vs. treatments; between-participant) design.

Participants in all conditions read "Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating heart arrhythmia (a condition in which the heartbeat is irregular, too fast, or too slow). Doctor Jones thinks of two different ways to provide good treatment to his patients suffering from arrhythmia. He considers Drug A or Drug B. Both drugs are affordable and patients can tolerate their side effects." In the "moral obligation present" conditions, participants then rated either the acceptability of Doctor Jones administering drug A or drug B to his patients ("How okay is it for Doctor Jones to prescribe patients Drug A[B]?") or of Doctor Jones running an experiment in which patients were randomly administered drug A or drug B ("How okay is it for Doctor Jones to do this?"). Both dependent variables were measured on 7-point scales with end points 1 =

"It's really bad"; 4 = "It's okay"; 7 = "It's really good." In the "moral obligation absent" conditions, participants were informed that before prescribing the drugs to his patience, Doctor Jones decided to give it to fruit flies [run an experiment with fruit flies]. "Fruit fly heart cells resemble human ones to such a degree that they are well-suited to study drugs to treat heart arrhythmia." The scenarios are provided in the Online Appendix.

Results. In support of our prediction, a significant interaction of appropriateness ratings was observed, $F(1, 397) = 44.47$, $p < .001$. When Doctor Jones treated humans, the experiment was rated as less appropriate than its least acceptable treatment ($t(198) = 7.04$, $p < .001$, $d = 1.00$). When Doctor Jones instead treated fruit flies, the experiment was rated as more appropriate than its least acceptable treatment, $t(199) = 2.40$, $p = .02$, $d = 0.34$ (cf., Figure 4).

Discussion. Using the best drug scenario, Study 4 showed that experiment aversion disappears when the moral obligation to adhere to standards of best practice is removed.

An alternative explanation for the results of Studies 3 and 4 (and also of Study 2B) may be that consumers do not object to experiments that violate normative standards of best practice, but rather object to experiments in which not all participants receive the same treatment, as such experiments may be perceived as unfair. Study 5 tests this possibility.

STUDY 5

According to our theorizing, when consumers strongly believe in the existence of a normative standard of best practice, conducting an experiment—but not the implementation of either of its treatments—is perceived as violating the standard. According to the alternative explanation, when consumers strongly believe in the existence of a normative standard of best practice, conducting an experiment—but not the implementation of either of its treatments—is perceived as unfair because not all study-participants receive the same

treatment in the experiment.

The two explanations can be tested against each other as they make opposing predictions for which kind of experiments will be perceived as objectionable. Assuming that violations of standards are judged as more objectionable than adherence to standards are judged as beneficial (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001; Weber, 1994), our account predicts that experiments in which each participant is assigned to both treatments (a within-participant design) should be perceived as violating the standard for each participant. Experiments in which each participant is randomly assigned to only one treatment (a between-participant design), in contrast, should be perceived as violating the standard for only half of the participants.

The alternative account of consumers objecting to unfair treatments makes the opposite prediction. Experiments employing within-participant manipulations should be perceived as less objectionable than experiments employing between-participant manipulations, because with the former experimental design all participants receive the same treatments.

Method

Sample. Three hundred MTurk workers accessed and completed Study 5 (*female* = 140, M_{age} = 39.15, SD_{age} = 11.74).

Design. Study 5 employed a 2 (experiment aversion: present vs. absent; between-participant) x 3 (scenario: scenario 1 vs. scenario 2 vs. scenario 3; within-participant) design. We selected the three scenarios from Meyer et al. (2019a) in which experiment aversion had been documented (Best drug, Charity, and Teacher wellbeing), and three scenarios from Mislavsky, Dietvorst and Simonsohn (2020) for which no experiment aversion had been found (Social networking, Ride-sharing and Product recommendation). For each scenario,

participants were asked a comprehension check question which they had to answer correctly to continue with the study.

Participants were provided either with the first or the second set of three scenarios, and were told that a researcher, who wanted to run an experiment, was trying to decide which experimental design to use, a within-participant or a between-participant design. For example, in the best drug scenario, participants were told that Doctor Jones wanted to test which of two drugs is most effective to treat high blood pressure. Doctor Jones was considering two ways to test the drugs:

> "Experiment A: 500 patients (randomly chosen) will be given drug A. Another 500 patients (also randomly chosen) will be given drug B. After a year, Doctor Jones will compare the results for both groups."
>
> "Experiment B: 1000 patients (randomly chosen) will be given both drugs, for the first six months the patients will be given drug A, and for the remaining six months they will be given drug B. After a year, Doctor Jones will compare the results."

Participants were asked to choose which experiment they thought was more acceptable ("experiment A", "experiment B", or "either experiment would be ok"). For the other scenarios please see the Online Appendix. Experiment A and Experiment B—within- and between-participant designs of the same experiment—were counterbalanced. To make within- and between-participant experiments comparable in terms of the potential number of norm violations, we specified the same sample size for both types of experimental designs.[2]

Results. As predicted, for the three scenarios in which experiment aversion had been previously documented, the majority of participants chose the between- over the within-participant design as more acceptable ($\chi^2(2)_{\text{best drug}} = 60.41$, $p < .001$; $\chi^2(2)_{\text{charity}} = 19.04$, $p < .001$; $\chi^2(2)_{\text{teacher well-being}} = 13.04$, $p = .001$). In contrast, for the three scenarios in which no experiment aversion had been previously observed, neither experimental design was

[2] In the preregistration of Study 5 we called this "distributive fairness."

preferred ($\chi^2(2)_{\text{product recommendation}} = 0.42$, $p = .81$; $\chi^2(2)_{\text{ride-sharing}} = 3.74$, $p = .15$; $\chi^2(2)_{\text{social networking}} = 2.69$, $p = .26$; cf., Figure 5). We analyzed choice shares also with a multinomial logistic regression on the IV 'normative standard of best practice' (Meyer (present) vs. Mislavsky (absent) scenarios) with errors clustered at the participant level, see Table 4.

Discussion. Study 5 tested two competing explanations for why consumers object to experiments. According to our violations of normative standards account, consumers find experiments using within-participants manipulations to be more objectionable than experiments using between-participant manipulations, because the latter implies that the normative standard is violated only for half of the study-participants. According to the unfair treatment account, consumers find experiments using between-participants manipulations as more objectionable than those using within-participants manipulations because in the latter design all study-participants receive the same treatments. The results of Study 5 provided support for the violation of normative standards account, and thus rule out perceptions of unfair treatments as an explanation.

STUDY 6

So far, we have alleviated experiment aversion by removing the moral obligation to adhere to standards of best practice (Study 4). In the final two studies, we alleviate (Study 6) and reverse (i.e., demonstrate experiment appreciation, Study 7) by directly changing consumers' beliefs in normative standards of best practice. Specifically, in Study 6 we tell participants that extant normative standards of best practice are not perfect and can be further improved (for example, by conducting an experiment). And in Study 7, we explain to participants that an experiment is conducted to discover a new normative standard of best practice. Consumers in these instances may recognize that the required information can only be gleaned from conducting an experiment.

Method

Sample. Seven hundred seventy-six Prolific Academic workers accessed and completed Study 6 (*female* = 391, M_{age} = 37.87, SD_{age} = 12.42).

Design. Study 6 employed a 2 (explanation: control vs. extant standards are not perfect, between-participant) x 2 (evaluation: experiment vs. treatments, between-participant) x 3 (scenario: best drug, charity, teacher well-being, within-participant) design. Statistically, we predicted a two-way interaction between evaluation (experiment vs. treatments) and explanation (absent vs. present).

To this end, we again used the three scenarios in study 1 for which experiment aversion had been previously documented. In the control condition, participants were given the description of the three scenarios and asked to evaluate either the experiment or its treatments (the control condition is hence a direct replication). In the 'extant normative standards are not perfect' condition, participants were additionally informed that experts were uncertain which treatment is superior, and which treatment had lesser side effects. For instance, in the best drug scenario for treating hypertension, participants were told that there was no consensus among physicians on which drug is superior, and that Doctor Jones had observed side-effects such as headache, vomiting, and general tiredness for both drugs. In order to find out which drug is better, Doctor Jones then either decided to conduct an experiment or prescribed one or the other drug to his patients. For the other scenarios please see the Online Appendix.

For each of the three scenarios, participants were asked either to rate the acceptability of running an experiment (e.g., "How okay is it for Doctor Jones to do this?") or to rate the acceptability of implementing the treatments (e.g., "How okay is it for Doctor Jones to

prescribe Drug A/B?"). Both dependent variables were measured on 7-point scales with end points 1 = "it's really bad"; 4 = "It's okay"; 7 = "It's really good."

Results and Discussion. A 2 (explanation: control vs. extant standards are not perfect) x 2 (evaluation: experiment vs. treatment) ANOVA with scenario as a repeated factor yielded the predicted interaction of explanation and evaluation, $F(1, 772) = 15.65$, $p < .001$. In the control condition, we replicated experiment aversion—experiments were rated as less appropriate than their least acceptable treatments—for all three scenarios, $t(388)_{\text{best drug}} = 7.26$, $p < .001$, $d = 74$; $t(388)_{\text{charity}} = 2.82$, $p = .005$, $d = .29$; $t(388)_{\text{teacher well-being}} = 5.12$, $p < .001$, $d = .52$. In the 'extant standards are not perfect' condition, no experiment aversion was found in any of the three scenarios, $t(384)_{\text{best drug}} = .93$, $p = .35$, $d = .09$; $t(384)_{\text{charity}} = .1.57$, $p = .118$, $d = .16$; $t(385)_{\text{teacher well-being}} = .60$, $p = .55$ (cf., Figure 6). In addition to the repeated measures ANOVA, we analyzed ratings also in 3 separate 2 (explanation: control vs. extant standards are not perfect) x 2 (evaluation: experiment vs. treatment) ANOVAs, one for each scenario. The predicted interaction of explanation and evaluation was significant for the best drug ($F(1, 772) = 20.06$, $p < .001$) and the teacher well-being scenario ($F(1, 772) = 10.18$, $p = .001$), but not for the charity scenario ($F(1, 772) = 0.73$, $p = .39$).

Study 6 shows that providing an explanation that extant standards of best practice are not perfect can help mitigate experiment aversion, since participants seem to realize that the necessary information to improve extant standards is best obtained by running an experiment.

STUDY 7

Study 7 tested our hypothesis whether consumers' objection to experiments could be reversed and turned into experiment appreciation by informing them that an experiment can

help discover or establish a new normative standard of best practice.[3] To this end, we adapted the best drug scenario and applied it to the Covid-19 pandemic. At the time the study was run (November 17th, 2020), no drug or vaccine had received emergency approval from the American Food and Drug Administration, so no normative treatment for Covid-19 existed yet. We modified the drug scenario such that Doctor Jones was treating Covid-19 rather than hypertension.

Method

Sample. Four hundred MTurk workers accessed and completed Study 7 (*female* = 207, M_{age} = 38.93, SD_{age} = 11.29).

Design. The study employed a 2(evaluation: experiment vs. treatments; between-participant) x 2 (normative standard: present vs. absent; between participant) design.

In the 'normative standard' conditions, participants read: "Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating Covid-19. Doctor Jones thinks of two different ways to provide good treatment to his Covid-19 patients." Participants evaluating the treatments further read: "He considers Drug A or Drug B. Both drugs are affordable and patients can tolerate their side effects." Participants evaluating the experiment read instead: "So he decides to run an experiment by randomly

[3] In the preregistration of Study 7, we had operationalized experiment appreciation as the mean rating of an experiment being significantly higher than the mean rating of its treatments. We later decided to operationalize experiment appreciation as the mean rating of an experiment being significantly higher than the maximum rating of its treatments, because experiment appreciation then signals that an experiment adds value above and beyond the implementation of its best treatment. Statistically, however, it makes little difference whether experiment ratings are compared to the mean or the maximum rating of its treatments because most participants in our studies rated treatments quite similarly.

assigning his Covid-19 patients to one of two test conditions. Half of the patients will be administered Drug A, and the other half will be administered Drug B. Both drugs are affordable and patients can tolerate their side effects."

In the 'no normative standard established yet' conditions, participants read: "Research labs across the world are racing to find a safe and effective drug to cure patients of Covid-19. Doctor Jones has developed two new drugs, drug A and drug B. However, Doctor Jones does not know yet whether the drugs are effective in triggering an immune system response, and whether they have (potentially severe) side-effects." Participants evaluating the treatments further read: "To test the drugs, Doctor Jones considers administering one of the two new drugs to a group of volunteers." Participants evaluating the experiment read instead: "To test the drugs, Doctor Jones decides to run an experiment with volunteers. Doctor Jones randomly assigns participants to one of two test conditions. Half of the participants will be administered Drug A, and the other half will be administered Drug B." Participants then rated either the acceptability of running the experiment "How okay is it for Doctor Jones to do this?" or the acceptability of the decision to administer drug A or drug B "How okay is it for Doctor Jones to administer Drug A/B?" Both dependent variables were measured on 7-point scales with end points 1 = "It's really bad"; 4 = "It's okay"; 7 = "It's really good."

Results and Discussion. As predicted, we replicated experiment aversion in the 'normative standard of best practice' condition. Participants rated the experiment as less acceptable than its least acceptable treatment $t(198) = 5.53$, $p < .001$, $d = 0.78$. In the 'no normative standard of best practice' condition, evaluations reversed: participants rated the experiment as more appropriate than its most acceptable treatment, $t(198) = 4.02$, $p < .001$, $d = 0.57$, (cf., Figure 7).

In line with our theorizing, experiment aversion can be reversed by explaining to participants that no standard of best practice exists yet, and an experiment will help finding or

establishing a new standard of best practice. Participants in this case judge running an experiment as more appropriate than implementing either of its treatments, as they realize that the necessary information can only be gleaned from conducting the experiment. This finding is obviously important for all practitioners who aim at developing evidence-based practices in marketing and elsewhere.

GENERAL DISCUSSION

What do consumers think about experiments conducted on consumers? One stream of research claims that consumers object to such experiments but approve of the implementation of an experiments' untested treatments or policies. If that was true, Amazon, Google, governmental bodies, charities, researchers, in short anyone wanting to test treatments or develop evidence-based best practices would have to conceal their intentions and run experiments without the explicit knowledge and consent of participants. Another stream of research claims that consumers view experiments run by corporations as acceptable as their worst treatment, but never as less acceptable. According to this view, consumers may be fine with Amazon, Netflix, and Google experimenting on them, but it may be difficult to get consumers' approval for experiments in public policy like the Oregon Health Insurance Experiment on the effect of expanding public health insurance in the United States (Finkelstein et al., 2012) or experiments studying ways to alleviate global poverty that were awarded the 2019 Nobel Prize (The Sveriges Riksbank Prize, 2019).

We propose a theory of lay beliefs that reconciles these contradictory claims and explains when and why consumers object to or approve of experiments. In eight preregistered studies, we provide empirical support for our account of experiment aversion and appreciation: When people strongly believe in a normative standard of best practice, conducting an experiment—especially an experiment using a within-participant design—

violates this standard and is objected to. Consumers relinquish their objection to experiments—and come even to appreciate them—when it is explained to them how the experiment will help improve extant standards or discover new normative standards of best practice.

Note that our theory does not explain why consumers in some contexts have a stronger belief in the existence of normative standards of best practice than in others. Our theory only states that when they do, they perceive experiments as violations of such standards and object to them. We show this to be the case for three scenarios/contexts in which extant research had documented experiment aversion to occur, and for 5 scenarios/contexts in which extant research had documented no experiment aversion to occur (Study 1). We also identified two new scenarios/contexts for which consumers hold strong beliefs in the existence of normative standards of best practice—a judiciary and a financial consulting scenario—and show that participants accordingly object to experiments in both scenarios/context (Studies 2A and 2B).

A few alternative explanations require attention. First, consumers may be averse to random assignment of study-participants to experimental treatments because it may be perceived as unfair or interpreted as showcasing a decision-makers' lack of knowledge/expertise. Both accounts, however, would predict that consumers judge experiments employing within-participant designs (no random assignment as all participants are assigned to all treatments) as more acceptable than experiments employing between-participant designs (random assignment of participants to different treatments). Study 5 tested this and found the opposite to be the case.

Second, one might argue that consumers are averse to experiments when they are run by individuals, but would approve of experiments conducted by corporations/organizations. Contrary to this hypothesis, however, Study 3 shows that participants disapprove of a

pharmaceutical company running an experiment, and Study 7 shows that participants appreciate Doctor Jones conducting an experiment to test the efficacy of two newly developed drugs for treating Covid-19.

Third, one might argue that high stakes, potential harm for study-participants, or lack of consent are the causes of experiment aversion. If true, consumers should object to all experiments with high stakes and potential harm for study-participants. The findings of Studies 6 and 7, however, demonstrate that consumers are not averse and even appreciate high-stakes experiments in healthcare that bear potential harm for study-participants.

Our theory and empirical findings highlight the importance of lay beliefs about normative standards of best practice in determining how consumers perceive experiments. Consequently, consumers can object to corporate experiments (Study 2B and Study 3) and appreciate experiments in healthcare and public policy (Study 6 and 7).

Our findings have important implications for practitioners, researchers, and decision-makers who want to run experiments to test and develop evidence-based interventions and policies, especially when policies would have a significant impact on consumer-wellbeing and societal welfare (Manchanda et al., 2005). Our findings show that consumers can be convinced of the usefulness of experimenting on consumers by expounding the ways in which experimental results will help improve or discover new standards of best practice. This can be achieved by explaining to consumers the purpose of conducting an experiment and pointing out the potential negative effects that the implementation of untested treatments may have, or by elucidating consumers that in a given scenario or context there is no agreed-upon standard of best practice, and the best experts can do is to test in an experiment which treatment will yield the most desirable outcome.

REFERENCES

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. *Rev. of Gen. Psychol.* 5: 323-370.

Buyalskaya A, Gallo M, Camerer CF (2021) The golden age of social science. *Proc. Natl. Acad. Sci. U.S.A.* 118(5): 1-11.

Finkelstein, A, et al., (2012) The Oregon health insurance experiment: evidence from the first year. *Q. J. Econ.* 127(3): 1057-1106.

Gneezy, A (2017) Field experimentation in marketing research. *J. Mark. Res.* 54(1): 140-143.

Goldfarb, A, Tucker, C (2011) Online display advertising: Targeting and obtrusiveness. *Mark. Sci.*, 30(3): 389-404.

Gordon, BR, Zettelmeyer, F, Bhargava, N, Chapsky, D (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Mark. Sci.* 38(2): 193-225.

Heck PR, Chabris CF, Watts DJ, Meyer MN (2020) Objecting to experiments even while approving of the policies or treatments they compare. *Proc. Natl. Acad. Sci. U.S.A.* 117(32): 18948-18950.

Hellman S, Hellman DS (1991) Of mice but not men. Problems of the randomized clinical trial. *N. Engl. Med.* 324: 1585–1589.

Huang, S, Aral, S, Hu, YJ, Brynjolfsson, E (2020) Social advertising effectiveness across products: A large-scale field experiment. *Mark. Sci.* 39(6): 1142-1165.

Lambrecht, A, Tucker, CE (2018) Field experiments. In *Handbook of Marketing Analytics*. Edward Elgar Publishing: 32-51

Lantos JD (2020) Randomized trials are deeply offensive. *Am. J. Bioeth.* 20: 3–5.

Lynch HF, Greiner DJ, Cohen IG (2020) Overcoming obstacles to experiments in legal practice. *Science* 367(6482): 1078–1080.

Manchanda P et al. (2005) Understanding firm, physician and consumer choice behavior in the pharmaceutical industry. *Mark. Lett.* 16(3): 293-308.

Meyer MN (2014) Misjudgements will drive social trials underground. *Nature* 511: 265.

Meyer MN (2015) Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colo. Tech. L. J.* 13: 273–331.

Meyer MN et al. (2019a) Objecting to experiments that compare two unobjectionable policies or treatments. *Proc. Natl. Acad. Sci. U.S.A.* 116(22): 10723–10728.

Meyer MN et al. (2019b) Reply to Mislavsky et al.: Sometimes people really are averse to experiments. *Proc. Natl. Acad. Sci. U.S.A.* 116(48): 23885–23886.

Mislavsky R, Dietvorst BJ, Simonsohn U (2019) The minimum mean paradox: A mechanical explanation for apparent experiment aversion. *Proc. Natl. Acad. Sci. U.S.A*, 116(48): 23883-23884.

Mislavsky R, Dietvorst B, Simonsohn U (2020) Critical condition: People don't dislike a corporate experiment more than they dislike its worst condition. *Mark. Sci.* 39(6): 1092-1104.

Munz, KP, Jung, MH, Alter, AL (2020) Name similarity encourages generosity: A field experiment in email personalization. *Mark. Sci.* 39(6): 1071-1091.

Nelson, L, Simester, D, Sudhir, K (2020) Introduction to the special issue on marketing science and field experiments. *Mark. Sci.* 39(6): 1033-1038.

Peer E, Vosgerau J, Acquisti A (2014) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Methods* 46(4): 1023-1031.

Ringach DL (2011) The use of nonhuman animals in biomedical research. *Am. J. Med. Sci.* 342(4): 305-313.

Rozin P, Royzman EB (2001) Negativity bias, negativity dominance, and contagion. *Pers. Soc. Psychol. Rev*. 5: 296-320.

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019. (n.d.). https://www.nobelprize.org/prizes/economicsciences/2019/summary/ [Accessed 29 March 2021].

Weber E (1994) From subjective probabilities to decision weights: The effect of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychol. Bull.* 115: 228-242.

Zhang D et al. (2014) Activation of histone deacetylase-6 induces contractile dysfunction through derailment of α-tubulin proteostasis in experimental and human atrial fibrillation. *Circulation*. 129(3): 346-358.

Zoumpoulis, S, Simester, D, Evgeniou, T (2015) Run field experiments to make sense of your big data. *Harvard Bus. Review*.

Table 1. The scenarios used in previous research: Scenarios in which experiment aversion was previously observed (left column) and not observed (right column).

	Experiment aversion observed	No experiment aversion observed
Scenarios from Meyer et al. 2019a	**Best drugs** Dr. Jones prescribes Drug A vs. Drug B for high blood pressure. **Charity** To help citizens in a low-income country, a charity provides a sturdy roof for homes vs. one month training in a trade. **Teacher well-being** Elementary school teachers receive a yearly bonus vs. additional vacation days during summer and winter breaks.	**Basic income** Citizens out of work for at least 12 months receive the equivalent of $1000 per month for 6 months, or until they find a job, whichever comes first.
Scenarios from Mislavsky, Dietvorst, & Simonsohn, 2020		**Social networking** Users of a social networking website see only positive vs. only negative content. **Shipping** Shipping company offers senders of packages standard shipping vs. priority shipping. **Ride-sharing** Ride-sharing company offers a 15% discount vs. $1 credit for each $10 spent **Product recommendation** Amazon recommends most profitable products vs. highly rated products to its customers.

Table 2. Pairwise comparisons for Study 1 (*N* = 200): Pairwise comparisons of lay beliefs about normative standards of best practice for scenarios for which experiment aversion had been documented (underlined) versus not.

Statement Pair	Paired Differences		Paired t-tests (df = 199)
	Mean	SD	
Shipping vs. Best drug	-0.59	2.00	$t = 4.17$, $p < .001$
Shipping vs. Teacher well-being	-0.65	1.84	$t = 4.95$, $p < .001$
Shipping vs. Charity	-0.66	1.66	$t = 5.63$, $p < .001$
Product recommendation vs. Best drug	-1.47	2.18	$t = 9.55$, $p < .001$
Product recomm. vs. Teacher well-being	-1.53	2.03	$t = 10.65$, $p < .001$
Product recommendation vs. Charity	-1.54	2.06	$t = 10.55$, $p < .001$
Social networking vs. Best drug	-1.44	2.32	$t = 8.79$, $p < .001$
Social networking vs. Teacher well-being	-1.50	2.32	$t = 9.10$, $p < .001$
Social networking vs. Charity	-1.51	2.21	$t = 9.68$, $p < .001$
Ride-sharing vs. Best drug	-0.75	2.09	$t = 5.04$, $p < .001$
Ride-sharing vs. Teacher well-being	-0.80	1.80	$t = 6.28$, $p < .001$
Ride-sharing vs. Charity	-0.82	1.78	$t = 6.48$, $p < .001$
Basic income vs. Best drug	-0.48	1.87	$t = 3.59$, $p < .001$
Basic income vs. Teacher well-being	-0.53	1.63	$t = 4.61$, $p < .001$
Basic income vs. Charity	-0.55	1.45	$t = 5.32$, $p < .001$

Table 3. Pairwise comparisons for Study 2A (*N* = 199): Pairwise comparisons of lay beliefs about normative standards of best practice for current scenarios on the financial and the judiciary domains (underlined) versus scenarios used as reference standards from Study 1.

Statement Pair	Paired Differences		Paired t-tests (df = 198)
	Mean	SD	
Product recommendation vs. Financial	-.83	2.09	$t = 5.63$, $p < .001$
Product recommendation vs. Judiciary	-.92	2.22	$t = 5.82$, $p < .001$
Social networking vs. Financial	-.89	2.06	$t = 6.09$, $p < .001$
Social networking vs. Judiciary	-.97	2.22	$t = 6.17$, $p < .001$

Table 4. Regression results for Study 5 (*N* = 300): Multinomial logistic regression results with IV 'normative standard of best practice' (Meyer (present) vs. Mislavsky (absent) scenarios) on answer choices, with errors clustered at the participant level. The reference category is "between-participant experiment".

Choice	IV	Beta	SE	*z*	*p*	95% Conf. Interval		1/Odds Ratio
Within-participant experiment	Intercept	-0.09	0.15	-0.59	0.552	-0.378	0.202	2.15
	Meyer = 1 Mislavsky = 0	-0.77	0.21	-3.68	<.001	-1.174	-0.359	
Either is fine	Intercept	-0.28	0.16	-1.77	0.077	-0.581	0.030	1.65
	Meyer = 1 Mislavsky = 0	-0.50	0.22	-2.29	0.022	-0.935	-0.073	
LR $\chi^2(2)$: 24.5								
Pseudo R^2: 0.0128								
Number of observations: 900								
Number of clusters: 300								

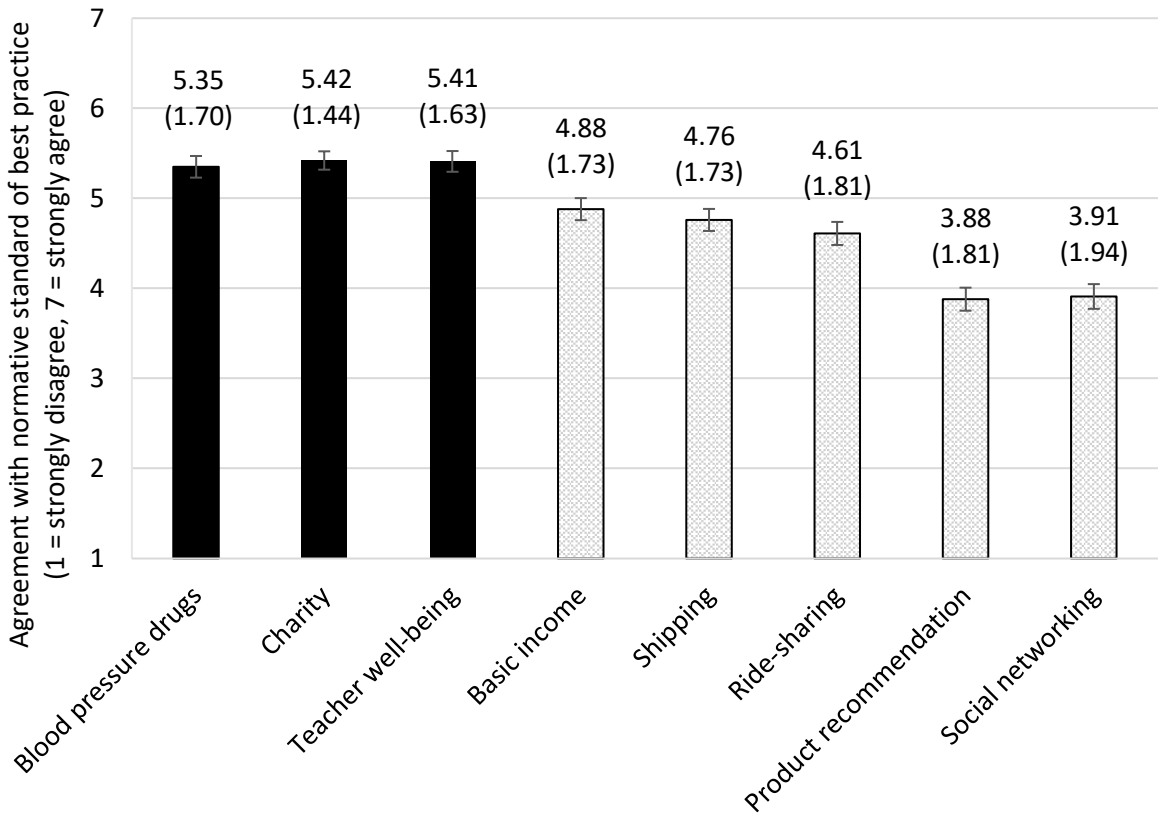

Figure 1. Agreement with normative standard of best practice in Study 1 (*N* = 200): Average agreement (std. dev.) with statements indicating a normative standard of best practice (error bars indicate +/- 1 standard error; dark (versus light) bars indicate previously documented experiment aversion (versus no experiment aversion))

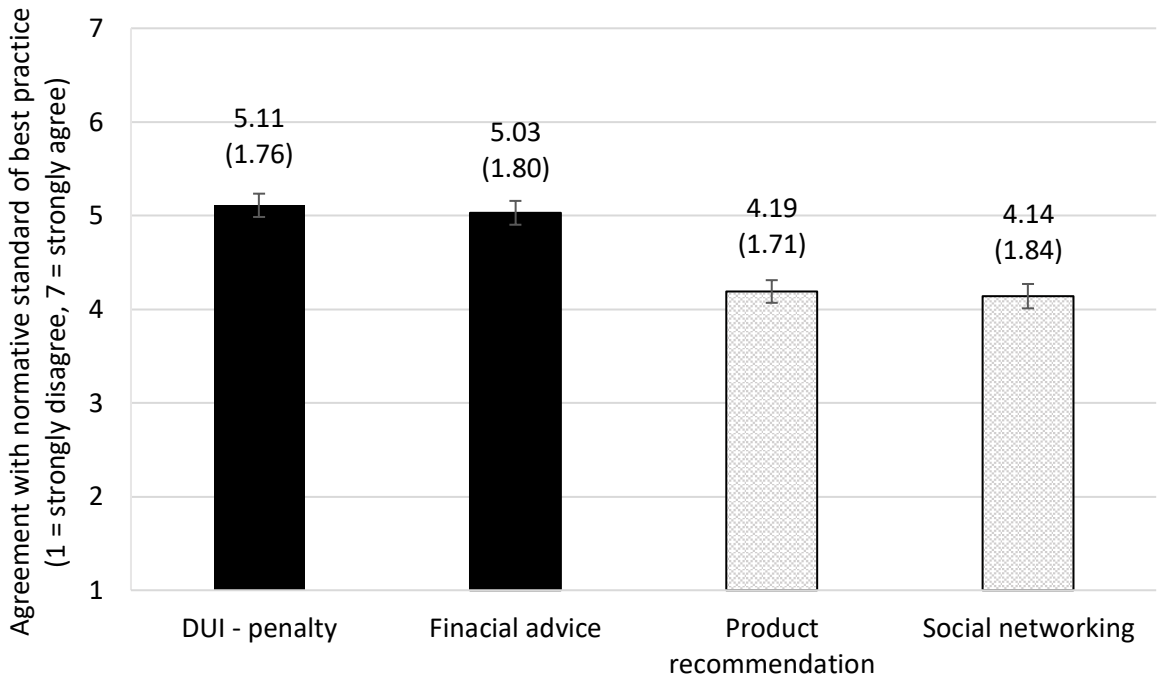

Figure 2. Agreement with normative standard of best practice in Study 2A (*N* = 199): Average agreement (std. dev.) with statements indicating a normative standard of best practice (error bars indicate +/- 1 standard error; dark (versus light) bars indicate previously documented experiment aversion (versus no experiment aversion))

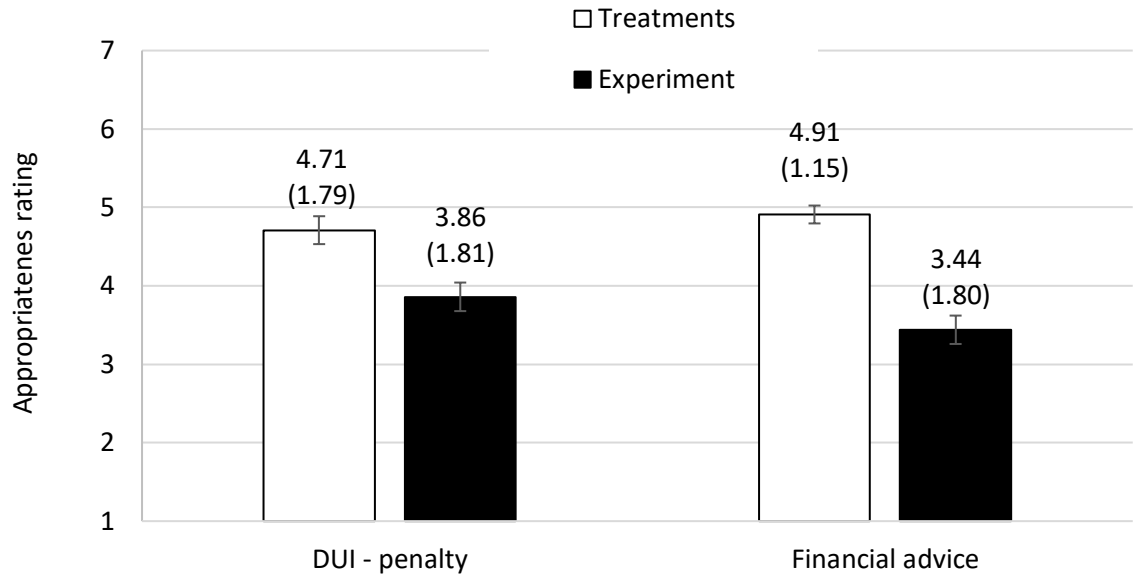

Figure 3. **Appropriateness ratings in Study 2B (*N* = 200):** Appropriateness ratings of experiments (mean, (std. dev.)) and their treatments (mean of minimum, (std. dev.)) for the judiciary scenario and the financial scenario. Error bars indicate +/- 1 standard error.

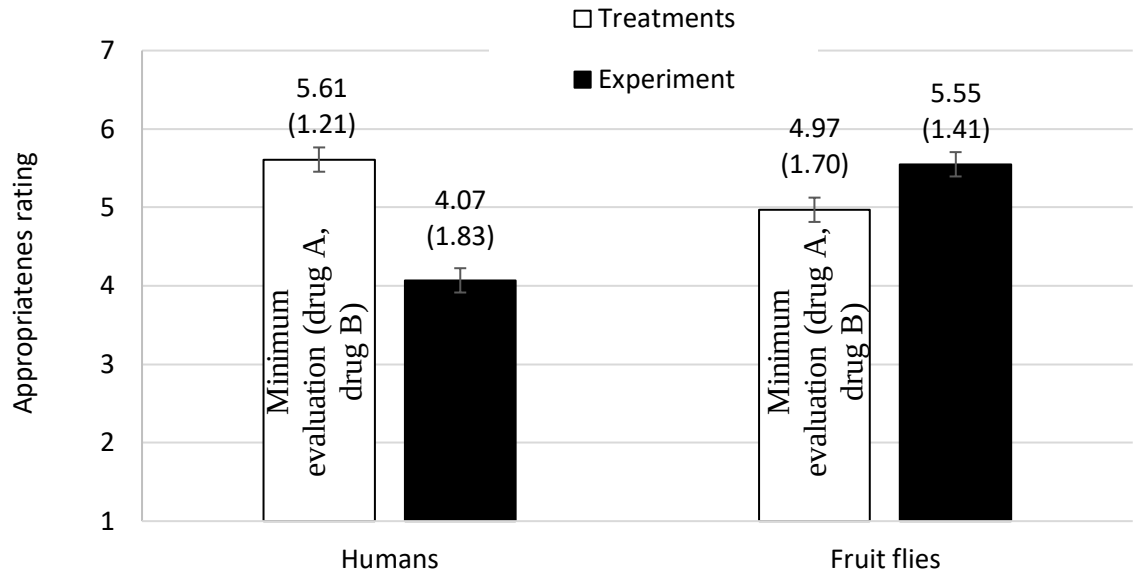

Figure 4. **Appropriateness ratings in Study 4 (*N* = 401):** Appropriateness ratings of experiments (mean, (std. dev.)) and their treatments (mean of minimum, (std. dev.)) when there is a moral obligation to follow normative standards of best practice (i.e., humans) versus not (i.e., fruit flies). Error bars indicate +/- 1 standard error.

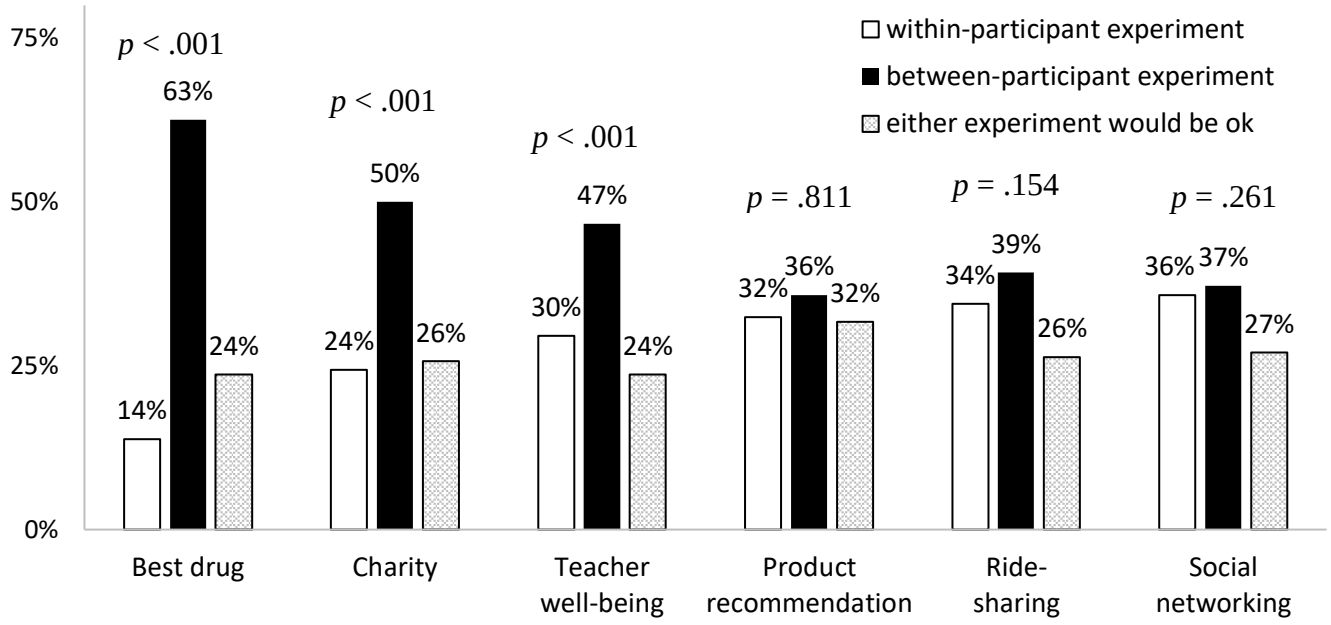

Figure 5. Choice frequencies in Study 5 (*N* = 300): Choice frequencies for most acceptable experimental design (between- versus within-participant) for the scenarios for which people believe in a normative standard (left three columns) and scenarios for which people do not believe in a normative standard (right three columns). P-values are estimated with $\chi^2(2)$ tests.

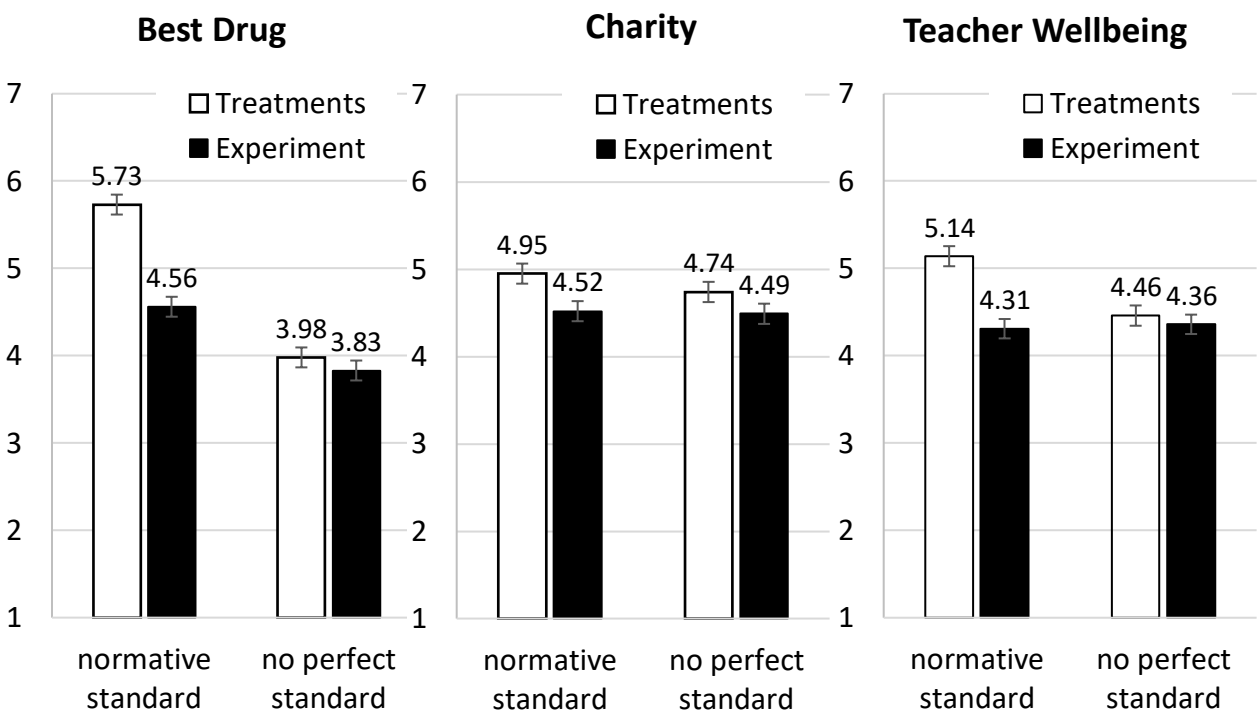

Figure 6. **Appropriateness ratings in Study 6 (*N* = 776):** Appropriateness ratings of experiments (mean) and treatments (min) for the best drug, charity, teacher well-being scenarios, when extant standards are not perfect (vs. control). Error bars indicate +/- 1 standard error.

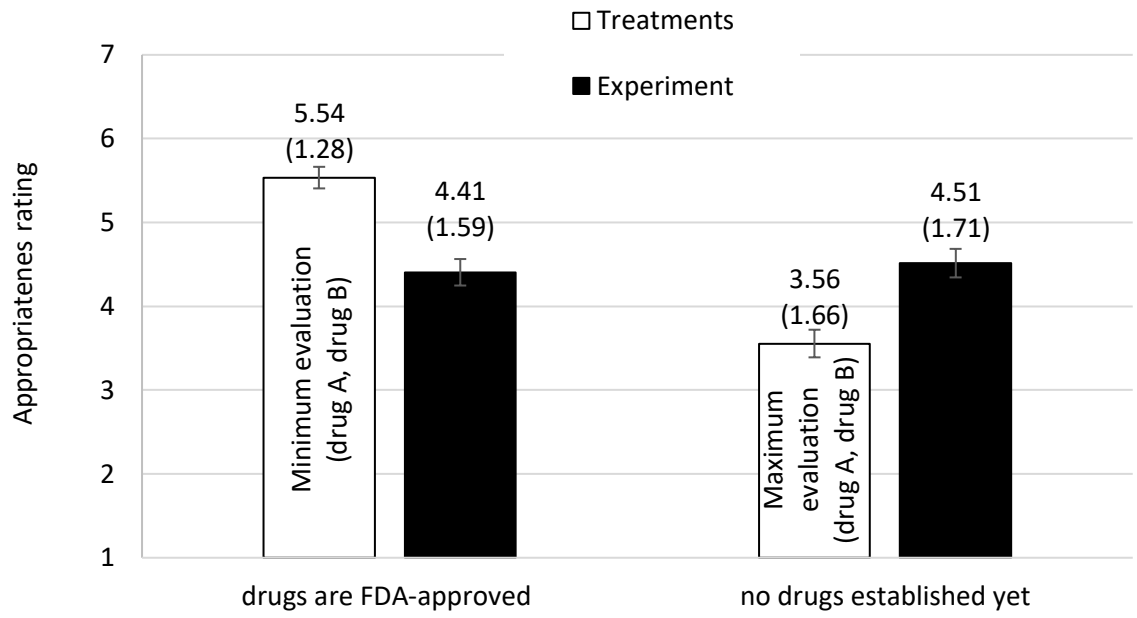

Figure 7. Appropriateness ratings in Study 7 (*N* = 400): Appropriateness ratings of experiments (mean, (std. dev.)) and their treatments when there is a normative standard of best practice (i.e., drugs are FDA approved; mean of minimum, (std. dev.)) versus the normative standard is yet to be established (i.e., no drugs established yet; mean of maximum, (std. dev.)). Error bars indicate +/- 1 standard error.

Why Do Consumers Condemn and Appreciate Experiments?

Online Appendix

The stimuli used in Study 1 - Study 7, and other analyses conducted, if any, are provided below. Where relevant, we refer to the findings on experiment aversion by Meyer et al. (1) or the absence of experiment aversion by Mislavsky et al. (2).

STUDY 1

The statements in Study 1 read as follows:

Blood pressure drugs: “In an ideal world, a patient suffering from a specific form of high blood pressure would receive the same treatment, no matter which doctor treated her/him.”

Charity: “In an ideal world, a citizen in a low-income country would receive the same help from a charity, either a sturdy roof for her home or one month of training in a trade of her choice, no matter which low-income country s/he lived in.”

Teacher well-being: “In an ideal world, a teacher at an elementary school would receive the same benefit, either a yearly bonus or additional vacation days during summer and winter breaks, no matter which school s/he worked for.”

Basic income: “In an ideal world, a citizen unable to find work would receive the same basic income, either a fixed amount $XX per month for 12 months (the basic income would stop as soon as s/he finds a job) or half of the $XX per month for 6 months (even if s/he finds a job during that time), no matter what country s/he lived in.”

Social networking: “In an ideal world, a user of a social networking website would see the same content, either all positive or all negative, no matter which social networking website s/he used.”

Shipping: “In an ideal world, the sender of a package would be offered the same shipping option, either standard shipping or priority shipping, no matter which shipping company s/he chose.”

Ride-sharing: “In an ideal world, a user of a ride-sharing company would be offered the same discount, either a 15% discount or a $1 credit for every $10 spent, no matter which ride-sharing company s/he chose.”

Product recommendation: “In an ideal world, an online shopper would be recommended the same products, either the products that are most profitable to the online retailer or the most highly rated products, no matter which website s/he used.”

In Study 1, we additionally tested whether a participant's level of education is related to the extent to which s/he believes in the existence of a normative standard of best practice (mean belief). The extent to which participants overall believed in the existence of a normative standard of best practice did not correlate with participants' education level, $r(200) = .043$, $p = .548$. For exploratory purposes, we also asked for participants' race, income, STEM degree, religion, and political ideology. We also regressed age, gender, race, income, STEM degree, religion and political ideology on the overall belief in the existence of a normative standard of best practice. None of these variables were significant.

STUDY 4

The stimuli used in Study 4 were as follows:

Moral obligation – Treatments
"Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating heart arrhythmia (a condition in which the heartbeat is irregular, too fast, or too slow).
Doctor Jones thinks of two different ways to provide good treatment to his patients suffering from arrhythmia. He considers Drug A or Drug B.
Both drugs are affordable and patients can tolerate their side effects."

No moral obligation – Treatments
"Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating heart arrhythmia (a condition in which the heartbeat is irregular, too fast, or too slow).
Doctor Jones thinks of two different ways to provide good treatment to his patients suffering from arrhythmia. He considers Drug A or Drug B. Both drugs are affordable and patients can tolerate their side effects.
Before he prescribes one of the drugs to his patience, he decides to give it to fruit flies. Fruit fly heart cells resemble human ones to such a degree that they are well-suited to study drugs to treat heart arrhythmia."

Moral obligation - Experiment
"Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating heart arrhythmia (a condition in which the heartbeat is irregular, too fast, or too slow).
Doctor Jones thinks of two different ways to provide good treatment to his patients suffering from arrhythmia, so he decides to run an experiment by randomly assigning his patients to one of two test conditions. Half of the patients will be prescribed Drug A, and the other half will be prescribed Drug B.
Both drugs are affordable and patients can tolerate their side effects."

No moral obligation - Experiment:
"Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating heart arrhythmia (a condition in which the heartbeat is irregular, too fast, or too slow).
Doctor Jones thinks of two different ways to provide good treatment to his patients suffering from arrhythmia, so he decides to run an experiment with fruit flies. Fruit fly heart cells resemble human ones to such a degree that they are well-suited to study drugs to treat heart arrhythmia.
Doctor Jones randomly assigns fruit flies with heart arrhythmia to one of two test conditions. Half of the fruit flies will be given Drug A, and the other half will be given Drug B.
Both drugs are affordable and fruit flies can tolerate their side effects."

STUDY 5

The stimuli used in Study 5 including a between- and within-participant experiment were as follows:

Best drug

"Doctor Jones wants to test which of two drugs is most effective to treat high blood pressure, drug A or drug B.
Both drug A and drug B have been approved by the U.S. Food and Drug Administration as safe and effective for treating high blood pressure.
However, Doctor Jones is not sure of how to test the drugs.

Experiment A: 500 patients (randomly chosen) will be given drug A. Another 500 patients (also randomly chosen) will be given drug B. After a year, Doctor Jones will compare the results for both groups.

Experiment B: 1000 patients (randomly chosen) will be given both drugs, for the first six months the patients will be given drug A, and for the remaining six months they will be given drug B. After the year, Doctor Jones will compare the results."

Charity

"Last year, a charity received a large number of donations.
The director of this charity wants to test which strategy to adopt to help people in low-income countries escape extreme poverty.
sturdy roof: Adults below a certain income level receive a sturdy roof for their home
trade training: Adults below a certain income level receive one month of training in a trade of their choice
However, the director is not sure of how to test these strategies.

Experiment A: 500 adults (randomly chosen) will receive a sturdy roof. Another 500 adults (also randomly chosen) will receive trade training. After a year, the director will compare the results for both groups.

Experiment B: 1000 adults (randomly chosen) will receive both strategies, for the first six months the adults will receive a sturdy roof, and for the remaining six months they will receive trade training. After the year, the director will compare the results."

Teacher wellbeing

"Research has shown that increasing teacher happiness and well-being can result in better student outcomes.

The superintendent of a school district wants to test which benefit to provide to teachers to improve student outcomes.
Benefit 1: Give a financial bonus
Benefit 2: Give additional vacation days during summer and winter breaks
However, the superintendent is not sure of how to test these strategies.

Experiment A: 500 teachers (randomly chosen) will be given a financial bonus (benefit 1). Another 500 teachers (also randomly chosen) will begiven additional vacation days (benefit 2). After a year, the superintendent will compare the results for both groups.

Experiment B: 1000 teachers (randomly chosen) will be given both benefits, for six months the teachers will be given a financial bonus (benefit 1), and for the remaining six months they will be given additional vacation days (benefit 2). After the year, the superintendent will compare the results."

Social networking
"A social networking website wants to test which ad displays are most effective.
The website thinks of two different ads:
Display 1: Show only sad ads
Display 2: Show only happy ads
However, the social networking website is not sure of how to test these strategies.

Experiment A: 500 users (randomly chosen) will only be shown sad ads. Another 500 users (also randomly chosen) will be shown only happy ads. After a month, the company will compare the results for both groups.

Experiment B: 1000 users (randomly chosen) will be provided with both ads, for the first two weeks the users will be shown only sad ads, and for the remaining two weeks they will be shown only happy ads. After the month, the company will compare the results."

Ride-sharing
"A ride-sharing company (e.g., Uber, Lyft) wants to test which promotional discounts are most effective.
The company thinks of two different discounts:
Discount 1: Customers get a $1 credit towards future rides for every $10 they spend
Discount 2: Customers get 15% off every ride
For both discounts the company will pay for the price difference (i.e., the drivers will not be paid less).
However, the company is not sure of how to test these strategies.

Experiment A: 500 customers (randomly chosen) will receive discount 1. Another 500 customers (also randomly chosen) will receive discount 2. After a year, the company will compare the results for both groups.

Experiment B: 1000 customers (randomly chosen) will receive both discounts, for the first six months the customers will receive discount 1,and for the remaining six months they will receive discount 2. After the year, the company will compare the results."

Product recommendation
"Imagine Amazon wants to test which recommendation strategies are most effective. Amazon thinks of two different ways to recommend items:
Option 1: Recommend items that are not selling well
Option 2: Recommend the items that are most profitable to Amazon
However, Amazon is not sure of how to test these strategies.

Experiment A: 500 customers (randomly chosen) will be recommended items that are not selling well (option 1). Another 500 customers (also randomly chosen) will be recommended items that are most profitable for Amazon (option 2). After a year, Amazon will compare the results for both groups.

Experiment B: 1000 customers (randomly chosen) will be provided with both options, for the first six months the customers will be recommended items that are not selling well, and for the remaining six months they will be recommended items that are most profitable for Amazon. After the year, Amazon will compare the results."

Table S1. **Choice frequencies for Study 5 (*N* = 300):** Choice frequency (%) of "within participant experiment", "between-participant experiment" or "either is fine" options as a function of a normative standard of best practice being present (Meyer scenarios) or absent (Mislavsky scenarios).

Choice	Meyer (%)	Mislavsky (%)	Total (%)
Within-participant experiment	22.59%	34.23%	28.33%
Between-participant experiment	53.07%	37.39%	45.33%
Either is fine	24.34%	28.38%	26.33%
Total	100%	100%	100%

STUDY 6

The stimuli used in Study 6 were as follows:

1. Best drug Scenario

Control - Treatments
"Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating high blood pressure.
Doctor Jones thinks of two different ways to provide good treatment to his patients with high blood pressure. He considers Drug A or Drug B.
Both drugs are affordable, and patients can tolerate their side effects."

Control - Experiment
"Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating high blood pressure.
Doctor Jones thinks of two different ways to provide good treatment to his patients suffering from high blood pressure, so he decides to run an experiment by randomly assigning his patients to one of two test conditions. Half of the patients will be prescribed Drug A, and the other half will be prescribed Drug B.
Both drugs are affordable, and patients can tolerate their side effects.
After a year, he will only prescribe to new patients, whichever drug has had the best outcomes for his patients"

Extant standards are not perfect - Treatments
"Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating high blood pressure.
Doctor Jones thinks of two different ways to provide good treatment to his patients suffering from high blood pressure. Drug A and Drug B are the two valid options.
However, there is no consensus among physicians on which drug is superior. In addition, Doctor Jones has observed that some patients being prescribed Drug A and Drug B have complained about negative side-effects such as headache, vomiting, and general tiredness."

Extant standards are not perfect - Experiment
"Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating high blood pressure.
Doctor Jones thinks of two different ways to provide good treatment to his patients suffering from high blood pressure. Drug A and Drug B are the two valid options.
However, there is no consensus among physicians on which drug is superior. In addition, Doctor Jones has observed that some patients being prescribed Drug A and Drug B have complained about negative side-effects such as headache, vomiting, and general tiredness.
To test this, Doctor Jones decides to run an experiment with his patients.
He randomly assigns patients to one of two test conditions. Half of the patients will be prescribed Drug A, and the other half will be prescribed Drug B.
After a year, he will only prescribe to new patients, whichever drug has had the best outcomes for his patients."

2. Teacher well-being Scenario

Control - Treatments

“Research has shown that increasing teacher happiness and well-being can result in better student outcomes.
The superintendent of a school district thinks of two different ways to improve how well his elementary school students do.
He considers assigning teachers to either receive a yearly bonus or to receive additional vacation days during summer and winter breaks.”

Control - Experiment
“Research has shown that increasing teacher happiness and well-being can result in better student outcomes.
The superintendent of a school district thinks of two different ways to improve how well his elementary school students do, so he decides to run an experiment by randomly assigning teachers to one of two test conditions. Half of the teachers will receive a yearly bonus. The other half will receive additional vacation days during summer and winter breaks.
After a year, the superintendent will give all teachers whichever benefit turns out to result in better student outcomes.”

Extant standards are not perfect - Treatments
“Research has shown that increasing teacher happiness and well-being can result in better student outcomes.
The superintendent of a school district thinks of two different ways to improve how well his elementary school students do.
He considers assigning teachers to either receive a yearly bonus or to receive additional vacation days during summer and winter breaks.
However, there is no consensus among superintendents who had previously adopted these strategies about which is superior. In addition, it is not clear whether they have any negative side-effects (for example, teachers who get an additional bonus may become motivated by money rather than their intrinsic desire to teach students, thereby in the long run leading to a reduction rather than an increase in teaching quality, and teachers who get additional vacation days may invest less time in preparing their courses).”

Extant standards are not perfect - Experiment
“Research has shown that increasing teacher happiness and well-being can result in better student outcomes.
The superintendent of a school district thinks of two previously untested ways to improve how well his elementary school students do.
He considers assigning teachers to either receive a yearly bonus or to receive additional vacation days during summer and winter breaks.
However, there is no consensus among superintendents who had previously adopted these strategies about which is superior. In addition, it is not clear whether they have any negative side-effects (for example, teachers who get an additional bonus may become motivated by money rather than their intrinsic desire to teach students, thereby in the long run leading to a reduction rather than an increase in teaching quality, and teachers who get additional vacation days may invest less time in preparing their courses).
To test these strategies, the superintendent decides to run an experiment by randomly assigning teachers to one of two test conditions. Half of the teachers will receive a yearly bonus. The other half will receive additional vacation days during summer and winter breaks.
After a year, the superintendent will give all teachers whichever benefit turns out to result in better student outcomes.”

3. Charity Scenario

Control - Treatments
“Last year, a charity received a large number of donations.
The director of this charity thinks of two different ways to help people in a low-income country escape extreme poverty.
He considers assigning adults below a certain income level to either receive a sturdy roof for their home or to receive one month of training in a trade of their choice.”

Control - Experiment
“Last year, a charity received a large number of donations.
The director of this charity thinks of two different ways to help people in a low-income country escape extreme poverty, so he decides to run an experiment by randomly assigning people to one of two test conditions. Half of all adults below a certain income level will receive a sturdy roof for their home. The other half will receive one month of training in a trade of their choice.
After a year, the director will begin providing everyone in the country whichever resource (roof or training) turns out to help more people escape extreme poverty.”

Extant standards are not perfect – Treatments
“Last year, a charity received a large number of donations.
The director of this charity thinks of two different ways to help people in a low-income country escape extreme poverty.
He considers assigning adults below a certain income level to either receive a sturdy roof for their home or to receive one month of training in a trade of their choice.
However, there is no consensus among charities that had previously adopted these strategies on which strategy is superior. In addition, it is not clear whether these options indeed help people escape extreme poverty, and/or whether they have potentially negative side-effects, (for example, getting a sturdy roof for free may reduce the motivation to make other improvements to houses, and those receiving one month of training in a chosen trade may expect similar training support in the future, thus effectively lowering their motivation to self-improve in the long run).”

Extant standards are not perfect – Experiment
“Last year, a charity received a large number of donations.
The director of this charity thinks of two previously untested ways to help people in a low-income country escape extreme poverty.
He considers assigning people to either receive a sturdy roof for their home or to receive one month of training in a trade of their choice.
However, there is no consensus among charities that had previously adopted these strategies on which strategy is superior. In addition, it is not clear whether these options indeed help people escape extreme poverty, and/or whether they have potentially negative side-effect (for example, getting a sturdy roof for free may reduce the motivation to make other improvements to houses, and those receiving one month of training in a chosen trade may expect similar training support in the future, thus effectively lowering their motivation to self-improve in the long run).
To test these strategies, the superintendent decides to run an experiment by randomly assigning people to one of two test conditions. Half of all adults below a certain income level will receive a sturdy roof for their home. The other half will receive one month of training in a trade of their choice.

After a year, the director will begin providing everyone in the country whichever resource (roof or training) turns out to help more people escape extreme poverty."

ONLINE APPENDIX REFERENCES

Meyer MN et al. (2019) Objecting to experiments that compare two unobjectionable policies or treatments. *Proc. Natl. Acad. Sci. U.S.A.* 116(22): 10723–10728.

Mislavsky R, Dietvorst B, Simonsohn U (2020) Critical condition: People don't dislike a corporate experiment more than they dislike its worst condition. *Mark. Sci.* 39(6): 1092-1104.

CHAPTER 2

UNDERSTANDING OVERHEAD AVERSION

Burcak Bas*

*Doctoral student and corresponding author

Department of Marketing - Bocconi University, Milan, Italy

Email. burcak.bas@unibocconi.it

Zachary Estes

Bayes Business School - City, University of London, London, UK

Email. zachary.estes@city.ac.uk

Aradhna Krishna

Ross School of Business - University of Michigan (Ann Arbor), Michigan, USA

Email. aradhna@umich.edu

UNDERSTANDING OVERHEAD AVERSION

ABSTRACT

Decisions about how to manage financial resources include choices on charitable giving. However, referred to as overhead aversion, people often tend to avoid charities, which allocate a high percentage to fundraising and administration expenses. The prevailing explanation for overhead aversion is perceived personal impact, which suggests that donors care about the impact they make on a cause. An alternative explanation is based on perceived charity efficiency, according to which people base their donation decisions on how competent and likely the charity is in delivering on its promise. In three pre-registered studies (N = 1,279), we disentangle these two competing mechanisms. In reversal of prior theory, we find evidence for perceived efficiency (vs. personal impact). Our findings have implications for charities to boost donation decisions while also contributing to consumers' perceived financial well-being.

Keywords: charitable giving, charity efficiency, overhead aversion, personal impact

People regularly decide how to manage their financial resources (Greenberg & Hershfield, 2018), including charitable giving (Browning & Finke, 2016). Indeed, many people engage in charitable giving, and donations in the United States approximated $472 billion in 2020 (Pruitt & Bergdoll, 2021). Charity evaluators, such as Charity Navigator and CharityWatch, facilitate donation decisions by providing ratings of charities on various metrics, including overhead – namely, the portion of donations allocated to administrative and fundraising expenses instead of contributing to the cause (Coupet & Berrett, 2018; Gneezy, Keenan, & Gneezy, 2014; Tinkelman & Mankaney, 2007). However, overhead increases the price of giving (Bowman, 2006) and donors tend to be overhead averse, giving lower donations to charities that have higher overhead costs (Caviola, et al., 2014; Charles, Sloan, & Schubert, 2020; Gneezy et al., 2014; Portillo & Stinn, 2018). This is especially concerning, given that overhead aversion feeds the non-profit starvation cycle, such that a charity underinvesting in organizational infrastructure ultimately becomes less effective (Lecy & Searing, 2015).

Because overhead aversion suppresses donations, it is important to reduce or eliminate it, to increase donations, while also contributing to consumers' perceived financial well-being. Researchers have worked on reducing overhead aversion. One suggestion is to have a third-party donor cover a charity's overhead, so that 100% of a donors' money goes for charitable purposes. Gneezy et al. (2014) robustly show that third-party coverage is a good way to decrease overhead aversion. In their studies, they compare donations when a third-party covers a charity's overhead versus when it does not, and they show that the former increases donations. They also show that third-party coverage results in similar donations as when a charity has no overhead. Thus, they conclude that third party coverage eliminates overhead aversion. Other suggestions include providing information

on the purpose of the overhead and omitting the word "overhead" altogether (Qu & Daniel, 2021).

Another way to tackle overhead aversion and increase donations is to understand its psychological drivers, in a way that charities can also better signal to donors that their money is well-spent, and thereby contributing to consumers' perceived financial wellbeing. Previous literature has identified perceived personal impact (also called the "efficacy hypothesis") and perceived charity efficiency ("efficiency hypothesis") as potential explanations of overhead aversion (Gneezy et al., 2014). The personal impact account suggests that donors care about the impact they make on a cause (Bekkers & Wiepking, 2011). Accordingly, donors prefer charities with low overhead because they allow for a greater contribution to a cause. This view, which has garnered support in the previous literature (Gneezy et al., 2014), is in line with the literatures on impact philanthropy (Duncan, 2004) and "warm glow" (Andreoni, 1990). However, an alternative explanation is perceived charity efficiency, according to which people base their donation decisions on how competent and likely the charity is to deliver on its promise. This view implies that donors prefer charities with low overhead because they appear efficient.

It is important to note that personal impact and charity efficiency are naturally correlated. Low overhead signals both that the charity is highly efficient, and that the donor has a larger impact on the cause. Given their significant and different theoretical and practical implications, it is important to isolate personal impact and charity efficiency, and to evaluate their relative prevalence and strength in donation decisions. If personal impact is the driver of overhead aversion, then charities should focus on donor-based interventions that enhance donors' perceptions of contributing to a cause. Alternatively, if charity efficiency underlies overhead aversion, then charities should undertake charity-based interventions that enhance perceptions of the charity's efficiency. This would also help

contribute to donors' perceived financial well-being, signaling that their money is well-spent.

We conducted three preregistered experiments (*N = 1,279*) to test charity efficiency as a main driver of overhead aversion. Study 1 attempted to cleanly separate personal impact from charity efficiency through money vs. time donations. Study 2 tested our assumption that lower overhead signals higher charity efficiency. Finally, Study 3 tested if perceived efficiency drives overhead aversion and influences donations, while controlling for personal impact.

Results

Study 1

Changing overhead amount can change both efficiency and impact (Gneezy et al., 2014). Study 1 (*N* = 778) uses a time vs. money donation manipulation to cleanly separate the two. Time donations (i.e., volunteering) are becoming more commonplace and have been in recent research focus (Liu & Aaker, 2008; Macdonnell & White, 2015). We argue that the signal of efficiency should not differ between money and time donations; higher overhead should reduce perceived efficiency for both money and time donations. However, the donor's personal impact should differ between donation units (money vs. time). High overhead costs should reduce the donor's personal impact when donating money, but not when donating time. This is because when donating money, the donation and overhead are in the same units (i.e., money), so the overhead costs reduce the donor's personal impact. When donating time, however, the donation (time) and overhead (money) are in different units, so the overhead does not reduce personal impact. Therefore, if personal impact underlies overhead aversion, then overhead aversion should only occur in money donations, not in time donations. Alternatively, if overhead aversion is due to charity efficiency, then overhead aversion should occur in both money and time donations. The predictions by each

account are provided in Figure 1 (Panels A and B). This study was preregistered: https://aspredicted.org/blind.php?x=n2q9jb.

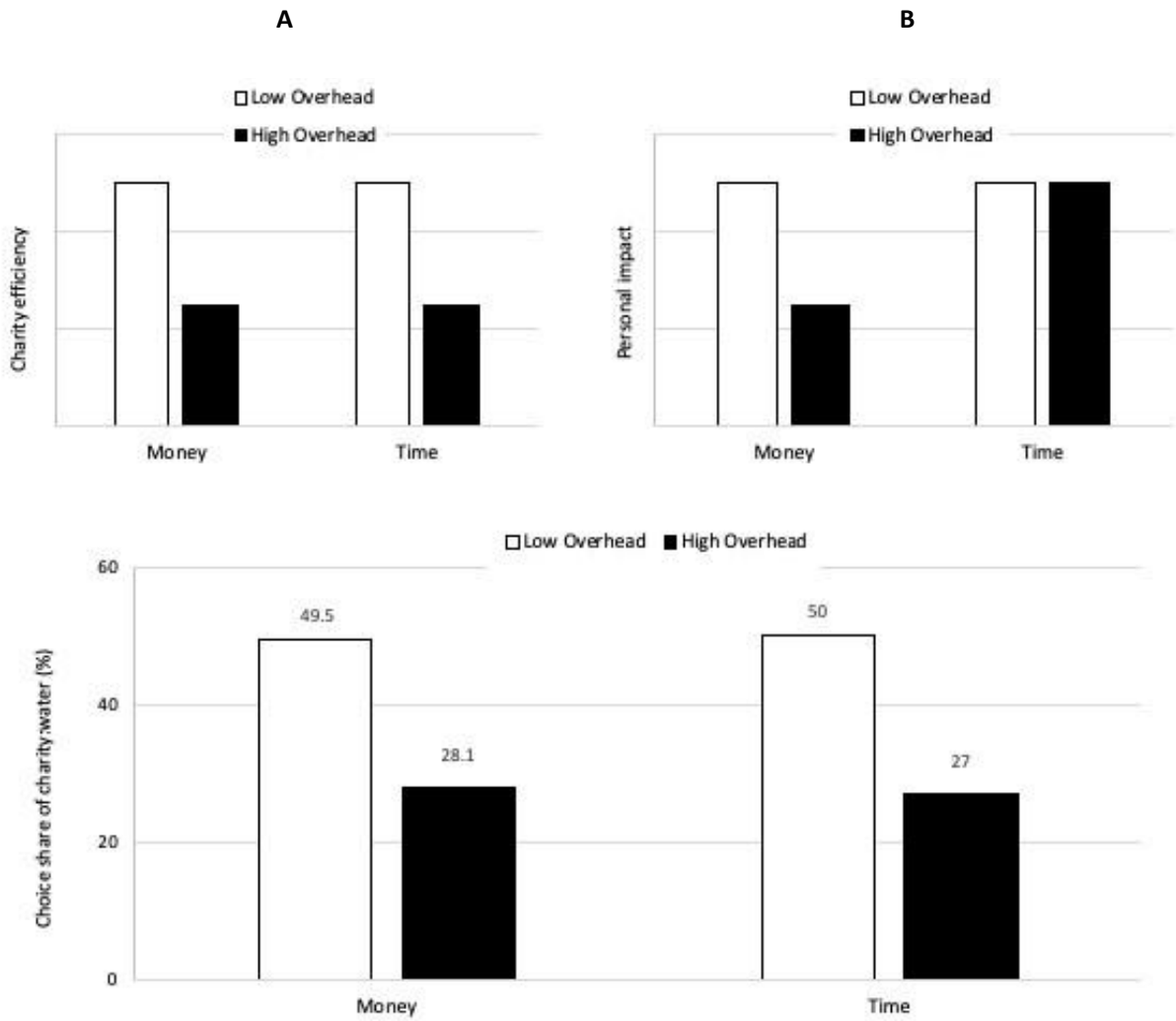

Figure 1. Panel A visualizes the prediction of the efficiency hypothesis (i.e., charity efficiency). Panel B visualizes the prediction of the efficacy hypothesis (i.e., personal impact). Panel C provides the results on the effect of overhead on the willingness to donate money and time.

This pattern of results is more consistent with a charity efficiency account, rather than a personal impact account, as argued earlier. However, one might still contend that there can be many reasons for choosing the low overhead charity in the time condition. This would imply that; charity efficiency is not necessarily ruled in as a mechanism.

Study 2

Study 2 (N = 300) tested more directly and independent of the distinction between money and time donations whether people perceive a charity's overhead as a signal of its efficiency, i.e., whether a higher overhead denotes lower efficiency. This is an important assumption to test since prior literature has (implicitly) assumed that it does not. For instance, Gneezy et al. (2014) assume that third party coverage of overhead increases personal impact without affecting perceived charity efficiency. In fact, this is how the authors suggest that personal impact, and not perceived charity efficiency, drives overhead aversion. But does third party coverage also affect perceived efficiency?

We randomly assigned participants to one of three conditions (no overhead = 0% overhead; 50% overhead; 50% overhead covered by a third-party donor). Participants evaluated the efficiency of a hypothetical charity (3 items, $\alpha = .95$), Good Lives For All, which was described as improving the living conditions of people who have lost their jobs and are affected by poverty. This study was preregistered: https://aspredicted.org/blind.php?x=eg3cy7.

As predicted, the overhead manipulation significantly affected perceived efficiency, $F(2, 297) = 45.32$, $P < .001$. The charity was perceived as being more efficient when there was 0% overhead (M = 5.84; SD = 1.20) versus when there was 50% overhead -- whether not covered (M = 4.02; SD = 1.43; $t(198) = 9.78$, $P < .001$, d = 1.38) or covered (M = 4.63; SD = 1.50; $t(198) = 6.35$, $P < .001$, d = .89); and when the 50% overhead was covered by a third-party or not ($t(198) = 2.94$, $P = .004$, d = .42). These results support our hypotheses that (a) a charity's overhead affects its perceived efficiency, and (b) when overhead is covered (vs. not), participants perceive the charity to be more efficient. Thus, in the results of Gneezy et al. (2014), perceived efficiency may also explain the mitigation of overhead aversion by third-party coverage of the overhead.

Study 3

Finally, Study 3 (N = 201) tested whether perceived efficiency drives overhead aversion, while controlling for personal impact. We again manipulated whether overhead was covered by a third-party donor or not. We randomly assigned participants to one of two conditions (overhead covered vs. overhead), and informed them about two charities, Kids Korps USA and charity:water. This study was preregistered: https://aspredicted.org/blind.php?x=8ia9ip.

All participants were told that there is no overhead associated with donations made to Kids Korps, whereas charity:water has 50% overhead. However, for charity:water we manipulated whether its 50% overhead was covered by a third-party or not. Participants chose which of the two charities should receive a $100 donation, and then rated charity efficiency (3 items, α = .88) and personal impact (3 items, α = .89) of charity:water. Corroborating previous research, more participants chose charity:water when overhead was covered (59.8%) vs. not (43.4%; $\chi^2(1, 201) = 5.39$, $P = .020$). charity:water was also perceived as being more efficient when overhead was covered ($M = 5.11$, $SD = 1.26$) than when it was not ($M = 4.40$, $SD = 1.27$), $t(199) = 3.95$, $P < .001$, $d = .56$. While perceived personal impact on charity:water was directionally higher when overhead was covered ($M = 5.28$, $SD = 1.17$) versus not ($M = 5.04$, $SD = 1.19$), consistent with Gneezy et al's (2014) assumption, in this study the difference was not significant ($t(199) = 1.39$, $P = .167$). A mediation analysis (Hayes, 2017) with charity efficiency and personal impact as parallel mediators revealed that the effect of a third-party covering overhead on choice of charity:water was mediated by charity efficiency, $B = .39$, $SE = .18$, $CI = (.1160, .7983)$, but not by personal impact, $B = .22$, $SE = .18$, $CI = (-.0922, .6125)$. Thus, contrasting prior theory, this study indicates that overhead aversion is due to charity efficiency, but not personal impact.

Discussion

In sum, although we replicate prior findings on overhead aversion, contrary to prior theory, we demonstrate that perceived efficiency underlies overhead aversion. Study 1 manipulates both overhead (high vs. low) and donation unit (money vs. time) and provides a critical test on the drivers of overhead aversion. For instance, if overhead aversion were due to personal impact, then it should not occur with time donations (i.e., volunteering). However, we found overhead aversion in both time and money donations. Study 2 shows that overhead ratio affects the charity's perceived efficiency, and that third-party coverage of overhead increases perceived efficiency. This finding shows generally that efficiency may underlie overhead aversion, and specifically that efficiency could explain Gneezy et al.'s finding that third-party coverage of overhead increases donations. Indeed, Study 3 demonstrates that charity efficiency mediates the effect of third-party coverage on donation decisions. Thus, our primary theoretical contribution is to provide a novel explanation of overhead aversion, thereby also explaining prior findings (e.g., Gneezy et al. 2014) in a new way. Our findings provide guidelines for charities to increase donations, without necessarily lowering overhead, based on charity efficiency perceptions as a major driver of overhead aversion. Importantly, our mechanism can contribute to consumers' perceived financial wellbeing by signaling them that their money is well-spent, thereby helping charities to increase donations.

Materials and Methods

Study 1

This study was conducted as part of a series of independent studies on MTurk. 801 participants completed the study, in exchange for a small monetary compensation. 778 valid participants (*female* = 339, M_{age} = 39.31) who passed the question on attention check were included in the analyses. In a 2 (donation unit: money vs. time) x 2 (overhead: high

vs. low) between-participant design, participants were randomly assigned to one of four conditions. In the beginning of the study, participants responded to an attention check - whether they were willing and able to dedicate attention to this research. All participants were presented with information about two charities, Kids Korps USA and charity:water. In the money (time) conditions, they were asked to decide which of the two charities they would donate to (volunteer for). Participants were told that there is no overhead associated with donations made to Kids Korps. For charity:water, we manipulated whether the overhead is high (i.e., 50%) vs. low (i.e., 5%). The order of the charities was counterbalanced. The scenarios were as follows:

1. Money – High overhead:

"In today's study we will ask you to give $100 to one of two non-profits. Your choice is whether to give the $100 to "Kids Corps." or to "charity:water".

Kids Corps.: Kids Korps is a non-profit organization that engages young people in volunteerism and teaches them about leadership and civic responsibility. There is no overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar donated, the entire $1 will go to "Kids Korps."

charity:water: charity:water is a non-profit organization that brings clean and safe drinking water to people in developing nations. There is 50% overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar donated, 50 cents will go to "charity:water" and 50 cents will be used to cover our costs."

2. Money – Low overhead:

"In today's study we will ask you to give $100 to one of two non-profits. Your choice is whether to give the $100 to "Kids Corps." or to "charity:water".

Kids Corps.: Kids Korps is a non-profit organization that engages young people in volunteerism and teaches them about leadership and civic responsibility. There is no overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar donated, the entire $1 will go to "Kids Korps."

charity:water: charity:water is a non-profit organization that brings clean and safe drinking water to people in developing nations. There is 5% overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar donated, 95 cents will go to "charity:water" and 5 cents will be used to cover our costs."

3. Time – High overhead:

"In today's study we will ask you to volunteer 5 hours of your time for one of two nonprofits. Your choice is whether to volunteer for "Kids Corps." or to "charity:water".

Kids Corps.: Kids Korps is a non-profit organization that engages young people in volunteerism and teaches them about leadership and civic responsibility. There is no overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar donated, the entire $1 will go to "Kids Korps."

charity:water: charity:water is a non-profit organization that brings clean and safe drinking water to people in developing nations. There is 50% overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar donated, 50 cents will go to "charity:water" and 50 cents will be used to cover our costs."

4. Time – Low overhead:

"In today's study we will ask you to volunteer 5 hours of your time for one of two nonprofits. Your choice is whether to volunteer for "Kids Corps." or to "charity:water".

Kids Corps.: Kids Korps is a non-profit organization that engages young people in volunteerism and teaches them about leadership and civic responsibility. There is no overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar donated, the entire $1 will go to "Kids Korps."

charity:water: charity:water is a non-profit organization that brings clean and safe drinking water to people in developing nations. There is 5% overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar donated, 95 cents will go to "charity:water" and 5 cents will be used to cover our costs."

Participants were then asked to respond to three multiple choice pre-evaluation comprehension check questions. The answers were presented in random order. Participants were allowed to proceed to the questions on choice only after they correctly answered these three questions. Otherwise, the participants were shown that the validation had failed, which allowed them to change their answer. Participants in the money (vs. time) conditions viewed: "Please tell us which organization you would like to give $100 to (volunteer 5 hours of your time for):" The answer choices were (Kids Corps; charity:water).

Study 2

We recruited 300 participants (*female* = 164; M_{age} = 32.59) from Prolific in exchange for a small monetary compensation. We randomly assigned participants to one of three

between-participant treatment conditions (no overhead; 50% overhead; 50% overhead covered). All participants first responded to an attention check - whether they were willing and able to dedicate attention to this research. Participants who failed this question were not allowed to start the study. We then provided participants with a scenario describing a hypothetical charity. The conditions were as follows:

1. No overhead:

"Good Lives For All is a charity that helps improve the living conditions of people who have lost their jobs and are affected by poverty.

0% of all money raised by "Good Lives For All" is spent on overhead (i.e., administrative and fundraising costs); all the money raised goes to the people they help.
So, for every dollar you would donate, the entire dollar would go to the people they help.

Right now, Good Lives For All is in need of money-donors."

2. 50% overhead:

"Good Lives For All is a charity that helps improve the living conditions of people who have lost their jobs and are affected by poverty.

50% of all money raised by "Good Lives For All" is spent on overhead (i.e., administrative and fundraising costs); the rest goes to the people they help. So, for every dollar you would donate, 50 cents would go to the people they help and 50 cents would be used to cover the costs.

Right now, Good Lives For All is in need of money-donors."

3. 50% overhead covered:

"Good Lives For All is a charity that helps improve the living conditions of people who have lost their jobs and are affected by poverty.

50% of all money raised by "Good Lives For All" is spent on overhead (i.e., administrative and fundraising costs); the rest goes to the people they help. However, someone else has already promised to cover this overhead cost. So, for every dollar you would donate, the entire dollar would go to the people they help.

Right now, Good Lives For All is in need of money-donors."

Participants were then asked to respond to two multiple choice pre-evaluation comprehension check questions. The answers were presented in random order. Participants

were allowed to proceed to the questions on perceived efficiency only after correctly answering both questions. Otherwise, participants were shown that the validation has failed, which allowed them to change their answer.

All participants then evaluated the perceived efficiency of the charity ("Please rate the extent to which you agree or disagree with the following statements"; (3 items: "Good Lives For All is efficient."; "Good Lives For All is effective."; "Good Lives For All is competent.")). All three items were on the same page and presented in random order. Each item was rated from 1 ("strongly disagree") to 7 ("strongly agree").

Study 3

We recruited 201 participants (*female* = 94; M_{age} = 31.44) from Prolific in exchange for a small monetary compensation. We randomly assigned participants to one of two treatment conditions (overhead vs. overhead covered). All participants first responded to an attention check - whether they were willing and able to dedicate attention to this research. Participants who failed this question were not allowed to start the study.

They were then presented with information about two charities, Kids Korps USA and charity:water, and asked to decide which of the two charities should receive a $100 donation. Participants were told that there is no overhead associated with donations made to Kids Korps. For charity:water, we manipulated whether the 50% overhead is covered by a third-party. The two conditions were otherwise identical. The scenarios were as follows:

1. Overhead:

"In today's study we will ask you to give $100 to one of two non-profits. Your choice is whether to give the $100 to "Kids Corps." or to "charity:water".

Kids Corps.: Kids Korps is a non-profit organization that engages young people in volunteerism and teaches them about leadership and civic responsibility. There is no overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar you'll donate the entire $1 will go to "Kids Korps."

charity:water: charity:water is a non-profit organization that brings clean and safe drinking water to people in developing nations. There is 50% overhead (i.e., spending on

administrative and fundraising costs) associated with this donation, so for every dollar you'll donate 50 cents will go to "charity:water" and 50 cents will be used to cover our costs."

2. Overhead covered:

"In today's study we will ask you to give $100 to one of two non-profits. Your choice is whether to give the $100 to "Kids Corps." or to "charity:water".

Kids Corps.: Kids Korps is a non-profit organization that engages young people in volunteerism and teaches them about leadership and civic responsibility. There is no overhead (i.e., spending on administrative and fundraising costs) associated with this donation, so for every dollar you'll donate the entire $1 will go to "Kids Korps."

charity:water: charity:water is a non-profit organization that brings clean and safe drinking water to people in developing nations. There is 50% overhead (i.e., spending on administrative and fundraising costs) associated with this donation, but someone else already covered this cost for your contribution, so for every dollar you'll donate the entire $1 will go to "charity:water.""

Participants were then asked to respond to two multiple choice pre-evaluation comprehension check questions, and they were allowed to proceed to the next page only after they correctly answered both questions. Otherwise, the participants were shown that the validation failed, which would allow them to change their answer. Once the correct answers were validated, participants indicated their choice of charity: "Please tell us which organization you would like to give $100 to: Kids Korps or charity:water." The order was counterbalanced.

Once participants provided their choice, they were asked about two mediators, namely perceived efficiency (3 items: "charity:water is competent."; "charity:water is effective."; "charity:water is efficient.") and perceived efficacy (3 items: "charity:water allows me to have a personal impact."; "charity:water helps me make an impact."; "I have an impact via charity:water."). Each item of a given mediator was randomly presented and rated from 1 ("strongly disagree") to 7 ("strongly agree"). The mediators were presented in different pages and in random order.

REFERENCES

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464-477.

Bekkers, R., & Wiepking, P. (2011). A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly*, *40*(5), 924-973.

Bowman, W. (2006). Should donors care about overhead costs? Do they care?. *Nonprofit and Voluntary Sector Quarterly*, *35*(2), 288-310.

Browning, C., & Finke, M. S. (2016). Conducting research in financial planning. In *Handbook of Consumer Finance Research* (pp. 103-114). Springer, Cham.

Caviola, L., Faulmüller, N., Everett, J. A., Savulescu, J., & Kahane, G. (2014). The evaluability bias in charitable giving: Saving administration costs or saving lives?. *Judgment and Decision Making*, *9*(4), 303.

Charles, C., Sloan, M. F., & Schubert, P. (2020). If someone else pays for overhead, do donors still care?. *The American Review of Public Administration*, *50*(4-5), 415-427.

Coupet, J., & Berrett, J. L. (2019). Toward a valid approach to nonprofit efficiency measurement. *Nonprofit Management and Leadership*, *29*(3), 299-320.

Duncan, B. (2004). A theory of impact philanthropy. *Journal of Public Economics*, *88*(9-10), 2159-2180.

Gneezy, U., Keenan, E. A., & Gneezy, A. (2014). Avoiding overhead aversion in charity. *Science*, *346*(6209), 632-635.

Greenberg, A. E., & Hershfield, H. E. (2019). Financial decision making. *Consumer Psychology Review*, *2*(1), 17-29.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process*

analysis: A regression-based approach. Guilford publications.

Lecy, J. D., & Searing, E. A. (2015). Anatomy of the nonprofit starvation cycle: An analysis of falling overhead ratios in the nonprofit sector. *Nonprofit and Voluntary Sector Quarterly*, *44*(3), 539-563.

Liu, W., & Aaker, J. (2008). The happiness of giving: The time-ask effect. *Journal of Consumer Research*, *35*(3), 543-557.

Macdonnell, R., & White, K. (2015). How construals of money versus time impact consumer charitable giving. *Journal of Consumer Research*, *42*(4), 551-563.

Portillo, J. E., & Stinn, J. (2018). Overhead aversion: Do some types of overhead matter more than others?. *Journal of Behavioral and Experimental Economics*, *72*, 40-50.

Pruitt, A., & Bergdoll, J. (2021, June 17). *Americans gave a record $471 billion to charity in 2020*. Fast Company. Retrieved October 15, 2021, from https://www.fastcompany.com/90647835/americans-gave-a-record-471-billiontocharity-in-2020.

Qu, H., & Daniel, J. L. (2021). Is "Overhead" A Tainted Word? A Survey Experiment Exploring Framing Effects of Nonprofit Overhead on Donor Decision. *Nonprofit and Voluntary Sector Quarterly*, *50*(2), 397-419.

Tinkelman, D., & Mankaney, K. (2007). When is administrative efficiency associated with charitable donations?. *Nonprofit and Voluntary Sector Quarterly*, *36*(1), 41-64.

CHAPTER 3

USING BENCHMARKS TO MITIGATE OVERHEAD AVERSION AND INCREASE DONATIONS

Burcak Bas*

*Doctoral student and corresponding author

Department of Marketing - Bocconi University, Milan, Italy

Email. burcak.bas@unibocconi.it

Zachary Estes

Bayes Business School - City, University of London, London, UK

Email. zachary.estes@city.ac.uk

Aradhna Krishna

Ross School of Business - University of Michigan (Ann Arbor), Michigan, USA

Email. aradhna@umich.edu

USING BENCHMARKS TO MITIGATE OVERHEAD AVERSION AND INCREASE DONATIONS

ABSTRACT

Donors are averse to overhead, tending to donate less to charities that allocate a high percentage of donations to administrative and fundraising expenses (i.e., overhead aversion). This especially works against charities that operate in high overhead industries (e.g., art museums), even if these charities are otherwise well run and curtail overhead expenses versus their peers. How can one help such charities increase donations? In this research, we reveal one strategy to mitigate such overhead aversion. We suggest that people judge charity overhead relative to that of peer charities, so that using benchmarks (i.e., providing the average overhead of similar charities) can impact donations. In four preregistered studies, we document that benchmarking mitigates overhead aversion, inducing higher donations to a charity that has lower overhead versus its peers, even if it has a high overhead per se. In understanding why this happens, we uncover the novel role of perceived charity efficiency - when relative overhead is lower, benchmarking increases perceived charity efficiency, which then mitigates the negative effect of absolute high overhead. Our findings add to prior work on the drivers of overhead aversion and provide guidelines for charities with high overheads to reduce overhead aversion and increase donations.

Keywords: benchmark, charitable giving, charity efficiency, overhead aversion, relative thinking

During the Covid-19 pandemic, charities such as United Way, Feeding America, and the Salvation Army played a big role in helping address the ensuing joblessness and poverty (Barrett 2021). According to Forbes' American Top 100 Charities List, these three charities are among the top registered charities in the U.S. operating in the domain of domestic needs (Barrett 2021). A charity is "an organization whose primary objectives are philanthropy and social well-being (e.g., educational, religious, or other activities serving the public interest or common good)" (Reiling 1958). Thus, charities are instrumental in lifting societal well-being. To function, charities depend on donations from the public and from corporations. In this paper, we investigate how to increase donations to charities.

Previous research has documented several ways to increase donations, such as telling donors that seed money has already been raised (Gneezy, Keenan, and Gneezy 2014); matching donations where another party gives the same amount as that given by a donor (Gneezy Keenan, and Gneezy 2014; Meier 2007); contingent match incentives where matching donations are made contingent on a specific percentage of other donors boosting their donations in a certain way, e.g., by making them recurrent donations (Anik, Norton, and Ariely 2014); and low donation defaults, whereby the default for a potential donor is set to a low amount, which decreases average donation amounts but increases the number of donors enough to increase total donations (Goswami and Urminsky 2016). However, a major hindrance to charitable donations is *overhead aversion*. In this paper, we suggest a way to increase donations, which focuses on reducing overhead aversion.

What is overhead aversion? An overhead ratio is the portion of donations allocated to administrative and fundraising expenses, instead of contributing to the cause (Coupet and Berrett 2018; Gneezy, Keenan, and Gneezy 2014; Tinkelman and Mankaney 2007). Although such overhead costs are often unavoidable, donors think of them as "a diversion of funds from program expenses" (Tinkelman and Mankaney 2007, p. 42). Donors want to see how

their money will be spent - specifically how much of their donations go to charity overhead versus the program - and many charities, such as Oxfam and Unicef, post information on their expenditures. For maximum social impact, donors expect the large chunk of financial resources to be devoted to program services rather than for overhead expenses towards facilities, equipment, and staff support (Bowman 2006; Buchheit and Parsons 2006; Keating and Frumkin 2003; Parsons 2007). Thus, *overhead aversion* is the tendency to donate less to charities that have high overhead costs (Caviola, et al. 2014; Charles, Sloan, and Schubert 2020; Gneezy, Keenan, and Gneezy 2014; Portillo and Stinn 2018).

Overhead aversion is concerning for charities and for societal welfare for various reasons. Primary among these reasons is that overhead ratio varies not only within industries, but also across industries, such that what is considered high or low overhead might not be informative and relevant for a given industry. Many charity aggregators and watchdogs present charities' overhead ratio (or "program ratio," which is the proportion spent on the cause, calculated as "1 - overhead ratio"). For instance, Charity Navigator reveals that charities in the realm of food and humanitarian relief have low overhead, presumably because they mostly have non-cash operations, and therefore they have a lower overhead ratio than most other charities rated on the website. In contrast, many charities in the arts and culture, such as museums, have expensive collections and property to maintain, and hence their overhead ratio is above-average (see Table 1). Thus, even if a museum had excellent control of its overhead costs and had lower overhead costs than most other museums, donors' overhead aversion could still substantially limit their donations to that museum.

Unfortunately, donors' overhead aversion leads charities in high overhead industries to underinvest in their organizational infrastructure (Gregory and Howard 2009), rendering them less effective in the long run (Lecy and Searing 2015). How can one help such charities reduce overhead aversion and increase donations? That is the question we focus on here.

Table 1. Examples of overhead ratios in two charitable industries.

Category		Charity Name	Overhead
Food Banks (N = 43; $M_{overhead}$ = 6.16%)	Low Overhead	SeaGate Food Bank	1.30%
		Food Bank of the Albemarle	3.00%
		United Food Bank and Services of Plant City	3.00%
		Toledo Northwestern Ohio Food Bank, Inc.	3.30%
		Food Bank of Northeast Louisiana	3.50%
		Shared Harvest Foodbank	3.70%
	Moderate Overhead	Community Action Services and Food Bank	4.70%
		Second Harvest Food Bank of North Central Ohio	4.70%
		FoodLink for Tulare County	4.80%
		Anne Arundel County Food Bank	5.20%
		North Country Food Bank, Inc.	5.50%
		Food Bank of Siouxland	5.80%
	High Overhead	Wichita Falls Area Food Bank	8.00%
		Fredericksburg Regional Food Bank	8.10%
		Community Table	10.80%
		Ballard Food Bank	12.10%
		Food Bank of North Central Arkansas	14.50%
		Fairbanks Community Food Bank Service	24.70%
Children's Museums (N = 32; $M_{overhead}$ = 24.31%)	Low Overhead	Golisano Children's Museum of Naples	10.50%
		The Children's Museum of Cleveland	12.40%
		The Children's Museum of Atlanta	12.80%
		Children's Museum of Houston	15.20%
		Children's Museum Tucson-Oro Valley	15.80%
		Imagine Children's Museum	16.60%
	Moderate Overhead	DuPage Children's Museum	22.00%
		Mississippi Children's Museum	22.10%
		Minnesota Children's Museum	22.60%
		Kohl Children's Museum of Greater Chicago	22.80%
		Long Island Children's Museum	23.50%
		Children's Museum of Pittsburgh	23.70%
	High Overhead	Dayton Society of Natural History	32.40%
		Children's Museum of Tacoma	35.90%
		Lincoln Children's Museum	36.30%
		Madison Children's Museum	36.50%
		Louisiana Children's Museum	37.90%
		Children's Museum & Theatre of Maine	45.30%

Notes: Data retrieved from Charity Navigator. Filtering criteria for size (i.e., annual expenses) in the Food Bank industry was up to $13.5 million. This corresponds to the size of charities in the Children's Museums industry, where no filtering criteria have been applied.

Recent literature has indicated a few methods to overcome overhead aversion. Gneezy, Keenan, and Gneezy (2014) robustly show that having a third-party donor cover a charity's overhead, so that 100% of donors' money goes for charitable purposes, abates overhead aversion and increases donations. Qu and Daniel (2021) propose another solution centering around communicating how the overhead is spent. They show that if people are provided the reason for the higher overhead (e.g., building long-term organizational capacity), it can reverse overhead aversion. Additionally, omitting the word "overhead" and just using "building long-term organizational capability" further increases donations.

We suggest a different strategy to decrease overhead aversion. We suggest that people judge a target charity's overhead relative to that of its peer charities, so that their donations are highly impacted by the provision of information relating to the overhead of peer charities. For instance, imagine a charity that allocated 30% of donations to overhead. How willing would one be to donate to this charity? Now imagine that, for the same charity with 30% overhead, one is also informed that similar charities allocate 50% of donations to overhead. Would willingness to donate to this charity change?

We argue that what matters to donors may not be the charity's overhead ratio, per se, but rather how high or low the overhead is relative to peer organizations. In four preregistered studies, we address how charities that have lower than average overhead can mitigate overhead aversion and encourage potential donors to donate. As such, we reveal a novel, simple, and effective strategy – benchmarking (i.e., providing the average overhead of similar charities) - to increase donations. In suggesting that charities communicate overhead benchmarks, we develop an intervention strategy (i.e., a strategy that can change behavior) to encourage donations without lowering the charity's overhead costs, which after all may be important for contributing effectively to the charitable cause. We also distinguish between two potential drivers of overhead aversion – namely, perceptions of personal impact and

charity efficiency. The former is already known to drive donations (Gneezy, Keenan, and Gneezy 2014), but the latter is novel. We show that when a target charity's overhead is lower versus its peers, then benchmarking increases that charity's perceived organizational efficiency, which then mitigates the negative effect of overhead aversion. Thus, we also add to prior work on the drivers of overhead aversion.

We believe that benchmarking could benefit many charities. Grey Matter Research (2018) indicates that nearly 60% of American donors believe that charities spend too much on overhead. Specifically, donors believe on average that 19% is a sensible overhead ratio for charities (Grey Matter Research 2018), whereas 30% is viewed as unacceptably high (Williams 2007). Thus, there are many charities and many industries with overheads higher than donors' expected overhead ratio. There may be charities that really manage their costs well and thus have lower overheads versus other charities in the industry. However, they may still be hurt by overhead aversion because their overhead is higher than donors' expectations. We suggest that they could benefit from providing the "industry average" as a benchmark. Take the case of the Kohl Children's Museum of Greater Chicago, which has an overhead ratio of 22.80% (Charity Navigator 2022; see Table 1). This number may seem high compared to an internal reference overhead ratio that a donor may have in mind, or if a donor scrolls through charities across various industries looking at their overhead ratios. However, the number is low compared to other children's museums, as the average for all Children Museums on Charity Navigator is 24.68% (see Table 1). If the donor is given this average as a benchmark, then donation to the Kohl Children's Museum may increase.

The rest of the paper is organized is follows. Next, we develop our conceptualization and hypotheses, after which we discuss our studies to test the hypotheses.

CONCEPTUAL DEVELOPMENT AND HYPOTHESES

We begin with additional discussion on overhead aversion followed by a discussion on relative thinking.

Overhead Aversion

Charities depend on donations to fulfil their missions of contributing to a cause while sustaining themselves. Referred to as overhead costs, charities must allocate some portion of these donations to cover administrative and fundraising expenses (Coupet and Berrett 2018; Gneezy, Keenan, and Gneezy 2014; Tinkelman and Mankaney 2007). Overhead costs are essential for charities to build organizational infrastructure and to increase organizational effectiveness, in their aim of achieving their missions.

However, the presence of overhead costs results in a clash between donors' desire to directly contribute to the cause and the long-term ability of charities to fulfill their missions (Tinkelman and Mankaney 2007). This makes overhead-related information highly sought by donors (Hope Consulting 2010), and overhead ratios have a large impact on donation decisions (Coupet and Berrett, 2018; Gneezy, Keenan, and Gneezy 2014; Tinkelman and Mankaney 2007). Donors strongly focus on overhead costs and tend to display overhead aversion (Caviola, et al. 2014; Charles, Sloan, and Schubert 2020; Gneezy, Keenan, and Gneezy, 2014; Portillo and Stinn 2018), regardless of how effective these charities might be in contributing to the cause (Metzger and Gunther 2019).

Higher overhead has been shown to reduce donation frequency (Bowman 2006), donation amount (Caviola et al. 2014; Tian, Hung, and Frumkin 2020), and the choice share of a high (versus low) overhead charity (Gneezy, Keenan, Gneezy 2014; Newman et al. 2019; Portillo and Stinn 2018; Qu and Daniel 2021). For instance, using archival data of a workplace giving campaign, Bowman (2006) found a negative association between overhead

ratio and donations. Using a combination of field and lab experiments, Gneezy, Keenan, and Gneezy (2014) demonstrated a causal link between overhead and donations to a charity.

Some other research on overhead aversion has used "program ratio", which is the proportion not spent on overhead (i.e., 1 - overhead ratio), and obtained similar results. Higher program ratios are shown to increase giving intention (Buccheit and Parsons 2006) and donation amount (McDowell, Li, and Smith, 2013; Stout 2001; Van der Heijden 2013). For instance, Stout (2001) found that donation amount is higher for charities with high program ratios. Similarly, Buccheit and Parsons (2006) showed that among donors choosing to compare charities based on financial information, giving intention is higher for the charity with the highest program ratio. These studies of program ratio are broadly consistent with the studies on overhead ratio, as one would expect.

Reasons for overhead aversion: charity efficiency versus personal impact. What does overhead signal to potential donors? Previous literature has identified perceived personal impact (also called the "efficacy hypothesis") and perceived charity efficiency ("efficiency hypothesis") as potential explanations of overhead aversion (Gneezy, Keenan, and Gneezy 2014). The personal impact account suggests that donors care about the impact they make on a cause (Bekkers and Wiepking 2011). Accordingly, donors prefer charities with low overhead because they allow for a greater contribution to a cause and to make a greater difference. This view has garnered support in the previous literature (Gneezy, Keenan, and Gneezy 2014). However, an alternative explanation is perceived charity efficiency, according to which people base their donation decisions on how competent and likely the charity is to deliver on its promise. This view implies that donors prefer charities with low overhead because they appear efficient.

Some previous research has shown that potential donors interpret high overhead ratios as a signal of inefficiency (Lecy and Searing 2015). To our knowledge, however, no prior

research has established a relation between perceived efficiency and charitable donations. In fact, on the assumption that having a third-party donor cover the overhead costs does not affect the charity's efficiency, Gneezy, Keenan, and Gneezy (2014) concluded that perceived efficiency does not underly overhead aversion. Notably, however, they did not actually measure perceived efficiency. We suggest that donors may in fact view a charity's ability to attract a third-party donor to cover overhead expenses as a sign that the charity is efficient. We therefore explicitly measure perceived efficiency and test whether it mediates the presumed effect of benchmarks on donations.

Strategies to decrease overhead aversion. Several intervention strategies have been proposed to mitigate overhead aversion. Gneezy, Keenan, and Gneezy (2014) demonstrated that having a high overhead (e.g., 50%) reduces donations to a charity relative to having a low overhead (e.g., 5%) or no overhead at all (i.e., 0%). They also showed that this overhead aversion is dispelled if it is covered by a third-party donor, whereby the focal donor does not have to pay for the overhead and 100% of her donations are instead devoted to the cause. Qu and Daniel (2021) presented participants with the context of a natural disaster and asked them to choose between two nonprofit organizations. Participants were shown four different scenarios. Choices of the participants across these scenarios revealed that typically donors prefer the organization with lower overhead. However, overhead aversion was diminished when information about the purpose of higher overhead, namely building long-term organizational capacity, was provided; in this case, more people donated to the higher overhead ratio charity (versus when the information was not given). In addition, they found that omitting the word "overhead" and just stating that funds are spent on building long-term organizational capacity further attenuated the negative effect of overhead on donations.

A charity with high overhead may, however, be more effective than a charity with lower overhead (Crutchfield and Grant 2008), especially if they operate in different industries

that have different overhead needs (Charity Navigator 2022). For instance, as stated earlier, charities in the realm of food and humanitarian relief supplies have low overhead expenses, but charities in the arts and culture have high overheads. So, how can otherwise effective charities mitigate overhead aversion and encourage potential donors to donate, without necessarily lowering overhead? We propose a strategic use of relative thinking in the judgment of overhead.

Relative Thinking

People use relative thinking in judgments (Azar 2007; Thaler 1980; Tversky and Kahneman 1981), such that they care about relative and not just absolute differences (Miller 1962). The theoretical foundations of relative thinking can be traced to Weber, whereby, for instance, people respond to changes in physical stimuli such as temperature by comparing it to some other value (Weber 1834). Anecdotal and systematic observations of this phenomenon have been widely documented in many marketing domains, such as that price promotions are judged as percentage discounts, even though rationally only the absolute discount matters (Fetherstonhaugh et al. 1997; Finucane, Peters, and Slovic 2003; Gabor and Granger 1964). Thaler (1980) thus showed that people are generally more sensitive to a saving of $5 on a $25 radio compared to a saving of $5 on a $500 television. Relative thinking is also a fundamental part of prospect theory, whereby individuals use reference points to evaluate their choices (Kahneman and Tversky 1979). Tversky and Kahneman (1981) demonstrated that people had a higher willingness to drive 20 minutes to save $5 on a calculator when it was priced at $15 than at $125.

Reference points and overhead ratio. The notion that consumers tend to think in terms of reference points has been adopted by pricing scholars, as consumers evaluate prices as high or low by comparing them with a reference price (Hardie, Johnson, and Fader 1993; Monroe 2003; Winer 1986), which they retrieve from their memory or their knowledge about

prices of other products (Kalwani et al. 1990). For overhead aversion, this implies that donors might be averse to overhead because it is higher than their internal reference point for overhead. As stated earlier, American donors have an internal reference point for overhead ratios, and on average they consider 19% a sensible amount (Grey Matter Research 2018).

How can relative thinking be strategically used to alter a donor's internal reference point for overhead? We argue that by providing benchmark overheads, charities can influence the reference overhead and thus impact donations. Some prior research has also tried to nudge donations by providing additional information on overhead ratios or program ratios. Much of this research has been done by accounting researchers and focuses on providing information on meeting minimum program ratios (i.e., in other words, being below a maximum overhead ratio). The results have been mixed. Buchheit and Parsons (2006) either provided donors with information on service efforts and accomplishment, including whether the charity met the generally accepted minimum standard for program ratios, or did not provide that information to donors. They found that the provision of this information increased the perceived quality of the charity and the proportion of donors who claimed they would donate to it in the future; however, it did not increase *actual* giving (i.e., foregoing an engraved pencil in order to give $2 to a charity). Importantly, they also showed that when given the chance to obtain comparative financial information on other similar charities, the majority of actual donors were unwilling to do so. This highlights the importance of explicitly providing a benchmark rather than expecting donors to go looking for competitive information by themselves.

Van der Heijden (2013) looks at an experimental situation where people compare several charities together and either do or do not have program ratios for all of them. He finds that the provision of program ratios for all charities increases (hypothetical) donations to charities with the highest program ratios, at the expense of those with the lowest program

ratios, but does not affect those in the middle, thus showing an extreme form of overhead aversion. If participants are given additional information that the lowest program ratio charity does not pass the threshold for acceptable program ratio, then this charity suffers even more, as is to be expected.

Tian, Hung, and Frumkin (2020) look at charities with high overheads and provide a norm for overhead. However, they compare donations when there is no information given to donors -- on the charity's high overhead ratio or a norm -- to donations when there is information on both the overhead ratio and norm. They find that donations decrease when information is given about a high overhead and a norm, consistent with overhead aversion. Notably, they do not examine a situation when donors see a high overhead charity but are given an even higher overhead norm.

As such, Buchheit and Parsons (2006) and Van der Heijden (2013) look at the effect of information about whether a charity crosses a program ratio threshold, and they obtain mixed results, and Tian, Hung, and Frumkin (2020) look at the combined result of providing high overhead information along with a norm. We look at the effect of explicitly providing competitive benchmark ratios (versus not).

One common reference point used by decision-makers to report performance information is benchmarking against competitors (Hammer and Champy 1993). In the case of donation to charities, benchmarking overhead ratio versus competitors can shift the donor's reference point from an internal ratio she may be holding to the externally provided ratio for the competitor. If the target charity is in a high overhead industry, but has a lower overhead than competitors, then the reference overhead will be lowered when the consumer shifts from the (lower) internal ratio to the (higher) external competitor ratio. This change in reference overhead ratio will then increase donations. Let's say a particular donor has an internal overhead reference ratio of 12%. If she is considering donating to Imagine Children's

Museum (see Table 1), its overhead ratio of 16.60% may seem high. But if the average ratio of 24.31% for all Children's Museum charities is provided as a benchmark, this externally provided ratio may become the benchmark, and then the overhead for Imagine Children's Museum may not seem high.

Note that in each charity industry (one can browse Charity Navigator to see various charity categories, e.g., Animal rights, welfare and services; International peace, security and affairs; Environmental Protection and Conservation) some charities have lower overheads versus others. These charities could benefit by benchmarking against other charities with higher overhead or the average for the industry. And if the target charity is in a high overhead industry, it benefits more by providing a benchmark (i.e., a charity in a low overhead industry may get donations anyway).

Relative overhead ratio, charity efficiency, and personal impact.

What would drive the presumed increase in donations when a (higher) benchmark is provided? We suggest that it's a change in perceived charity efficiency. When competitor charities have higher overhead ratios, the target charity appears more efficient. While perceived personal impact may also be affected by benchmarks, the relationship is more tenuous; personal impact should be driven more by absolute overhead ratio than by relative overhead ratio.

Note that even with benchmarking, donors may still be overhead averse – but they would be looking at relative overhead ratios rather than absolute overhead ratios, mitigating overhead aversion if the benchmark provided is higher. That is, providing a (higher) benchmark can increase perceived efficiency and counteract overhead aversion. Additionally, since relative overhead is also indicative of efficiency, the aversion to high relative overhead can also be seen as a desire for higher efficiency. Hence, we predict that an

overhead benchmark increases charitable donations by increasing perceptions of charity efficiency. More formally:

H1: Providing an overhead benchmark that is higher than a given charity's overhead increases donations to that charity.

H2: Perceived charity efficiency mediates the effect of an overhead benchmark on charitable donations.

Providing a benchmark overhead ratio as an intervention strategy may even reverse overhead aversion. This can occur when people are choosing to donate between a lower-than-average overhead charity in a high overhead industry (say, Imagine Children's Museum; see Table 1) versus a higher-than-average overhead charity in a low overhead industry (say, Ballard Food Bank). Without the industry average benchmarks, Imagine Children's Museum's overhead of 16.60% is higher than Ballard Food Bank's overhead of 12.10% and is likely to get lower donations, due to overhead aversion. But with the benchmarks of 24.31% for Children's Museums and 6.16% for Food Banks provided, Imagine Children's Museum might get higher donations than Ballard Food Bank. If charity websites want to shift donations to high overhead industries, this may be one such strategy. That is, charity aggregators and watchdogs could provide average overhead ratio for the industry as one more statistic in the records for each charity. We thus hypothesize that:

H3: Benchmarking can reverse overhead aversion, such that donations are higher for a high-overhead charity with an even higher average-for-the-industry overhead benchmark compared to a low-overhead charity with an even lower average-for-the-industry overhead benchmark.

We now present our studies.

OVERVIEW OF STUDIES

In four preregistered studies, we test the effect of benchmarking on donation decisions, and in mitigating overhead aversion, while also considering donor perceptions of charity efficiency as a driver. Study 1 illustrates that providing a high (vs. low vs. equal) benchmark increases donation intentions (**H1**). Next, Study 2 replicates the observed effect of providing a benchmark on willingness to donate (**H1**), and also demonstrates via mediation that charity efficiency underlies this effect (**H2**). Study 3 tests benchmarking with consequential choice behavior. Study 3 also demonstrates that, in the presence of a benchmark, high overhead (with higher benchmark) elicits higher willingness to donate than low overhead (with lower benchmark), indicating a reversal of overhead aversion (**H3**). Finally, building on Study 3, Study 4 tests the effects of both overhead (low, high) and benchmarking (absent, present) on willingness to donate. We replicate overhead aversion in the absence of a benchmark, and demonstrate that the reversal of overhead aversion in the presence of a benchmark (**H3**) is due to perceived charity efficiency (**H2**). All experiment scripts (Qualtrics), all preregistrations, all data, and all statistical codes are publicly available in a ResearchBox at the following address: https://researchbox.org/553&PEER_REVIEW_passcode=RHQFUM. We now discuss each study in detail.

STUDY 1: Overhead Benchmarks Impact Donations

Study 1 tests our basic hypotheses that the provision of a competitive overhead benchmark can impact donations. More specifically, we tested **H1**, that providing an overhead benchmark that is higher than a given charity's overhead increases donations to that charity. Participants read a brief scenario about The American Brain Tumor Association that has 30% overhead. Importantly, we varied the benchmark (i.e., the average overhead ratio of

similar charities) to be 10% (low benchmark), 30% (equal benchmark) and 50% (high benchmark). We predicted that when the overhead ratio of a target charity is lower (vs. higher) than the benchmark provided, the willingness to donate for that charity will be higher. In addition, we included a control condition whereby the overhead ratio of the target charity is the same as the benchmark provided. This tested whether a high benchmark increases or whether a low benchmark reduces willingness to donate, relative to an equivalent benchmark. This is an important distinction to make, in order to show the positive effect of providing a (higher overhead) benchmark.

Method

Participants and design. Three hundred ninety-eight students (*female* = 215; M_{age} = 20.44; SD_{age} = .84) at a U.S. university participated for course credit as part of a series of studies. In a between-participant design, participants were presented with a charity that has 30% overhead and were assigned randomly to one of three benchmark conditions: low (10%), equal (30%), or high (50%).

Procedure. Participants first reported their gender and age. They then read a brief scenario about The American Brain Tumor Association, which was seeking money-donors. The description read: "Brain disease affects 1 in 6 people. The American Brain Tumor Association aims to advance the understanding and treatment of brain tumors with the goals of improving, extending, and, ultimately, saving the lives of those impacted by a brain tumor diagnosis." Participants in the 10% (vs. 30% vs. 50%) benchmark condition were additionally told that The American Brain Tumor Association allocates 30% of all donations to overhead (i.e., administrative and fundraising costs) and that the average overhead cost for similar charities is 10% (vs. 30% vs. 50%).

To ensure that participants understood the scenario, they were required to answer two comprehension check questions correctly before advancing to the dependent variable: They

were required to provide the correct overhead costs for the target charity (options: 10%, 30%, 50%) and for similar charities (options: 10%, 30%, 50%). Participants were allowed to change any incorrect responses and could advance to the dependent variable only when both responses were correct. Finally, on the next page, they were asked their willingness to donate via 3 items ("I would be [likely to donate money / inclined to donate / willing to donate money] to the charity The American Brain Tumor Association") rated on a scale from "strongly disagree" (1) to "strongly agree" (7).

Results and Discussion

Studies 1, 2, and 4 all used the same dependent variable, willingness to donate. Across those studies, participants' gender and age tended to have significant effects on willingness to donate. Consistent with prior research (Mesch et al. 2011), women generally were more willing to donate than men (Study 1: Cohen's $d = .17$, $p = .09$; Study 2: $d = .30$, $p < .05$; Study 4: $d = .41$, $p < .001$). Participants' age correlated negatively with willingness to donate, though the effect was small (Study 1: $r = -.14$, $p < .01$; Study 2: $r = -.17$, $p < .05$; Study 4: $r = -.09$, $p < .05$): Younger respondents were more willing to donate. Most importantly for the present purposes, however, neither gender nor age significantly moderated any effects in any of our studies. That is, neither gender nor age interacted significantly with the benchmarking factor in any of our studies, and the benchmarking effects reported in each of our studies remained significant when gender and age were included in the analyses. Thus, we do not discuss gender or age further.

A one-way ANOVA indicated a significant effect of benchmarking on willingness to donate (Cronbach $\alpha = .89$), $F(2,395) = 5.57$, $p = .004$. As predicted, providing a high (50%) overhead benchmark ($M = 5.16$; $SD = 1.21$) had a positive and significant effect on willingness to donate, compared to both an equal (30%) benchmark ($M = 4.80$; $SD = 1.25$), $t(265) = 2.36$, $p = .019$, $d = .29$, and a low (10%) benchmark ($M = 4.65$; $SD = 1.36$), $t(263) =$

3.22, $p = .001$, $d = .40$. The equal and low benchmark conditions did not differ significantly, $t(262) = .95$, $p = .341$, $d = .12$. Note that the target charity's overhead ratio was constant (30%) in all three conditions. These results support our basic hypothesis **H1** that a (higher) benchmark will increase donations. Next, we test **H2**.

STUDY 2: The Role of Perceived Charity Efficiency (and Personal Impact)

Study 2 tested whether perceived charity efficiency mediates the effect of an overhead benchmark on charitable donations (**H2**). Participants read a brief scenario about a hypothetical charity, Good Lives For All, that has a 35% overhead. We varied the benchmark to be low (20%) or high (50%). Participants first indicated their willingness to donate to this charity and then rated perceived personal impact and perceived charity efficiency. We predicted that when the overhead ratio of a target charity is lower (vs. higher) than the benchmark provided, the willingness to donate for that charity will be higher. We further hypothesized that this effect will be driven by increased perceptions of charity efficiency. That is, we suggest that relative overhead ratio signals charity efficiency, which then mitigates overhead aversion and increases donations. For comparison, we also measured perceived personal impact, which is known to contribute to overhead aversion (Gneezy, Keenan, and Gneezy 2014).

Method

Participants and design. Two hundred one US residents (*female* = 111; $M_{age} = 31.59$; $SD_{age} = 11.68$), recruited from Prolific, participated for £0.21. We randomly assigned participants to one of two conditions (benchmark: 20% vs. 50%).

Procedure. Participants first reported their gender and age and then completed an attention check in which they were asked, "How much attention are you willing and able to dedicate to this study? If you select VERY LITTLE, then we will have to reject your work

(and will not pay you!).” Response options were “very little” and “a whole lot,” and only participants who answered “a whole lot” were allowed to proceed with the study (those who selected “very little” were directed to a survey exit page).

After completing the attention check, all participants viewed a scenario describing a hypothetical charity, Good Lives For All, that helps improve the living conditions of people who have lost their jobs and are affected by poverty. Participants were told that Good Lives For All allocates 35% of all donations to overhead (i.e., administrative and fundraising costs), when the average overhead cost for similar charities is 20% (vs. 50%).

All participants were first asked their *willingness to donate* (3 items), as in Study 1. They then rated, on separate and randomly ordered pages, *perceived personal impact* (3 items: "Good Lives For All allows me to have a high personal impact."; "I contribute directly to the cause."; "I have a considerable impact on the cause.") and *perceived charity efficiency* (3 items: "Good Lives For All is efficient."; "Good Lives For All is effective.", "Good Lives For All is competent."). The personal impact items were based on the three key properties identified by Carroll and Kachersky (2019). The charity efficiency items were based on Aaker, Vohs, and Mogilner’s (2010) organizational competence measure, because charity efficiency is about “delivering on promise” (i.e., competence; Gneezy, Keenan, and Gneezy 2014). All nine items (three each for willingness to donate, personal impact, charity efficiency) were rated on a scale from “strongly disagree” (1) to “strongly agree” (7).

Results and Discussion

Results are illustrated in Figures 1 and 2. Relative to the low overhead benchmark (20%), the high benchmark (50%) significantly increased willingness to donate, perceived personal impact, and perceived charity efficiency (Figure 1). Both personal impact and charity efficiency mediated the effect of overhead benchmarks on donations (Figure 2).

Effects. As predicted, willingness to donate (α = .95) was higher when the benchmark overhead was 50% (i.e., high benchmark; *M* = 4.82; *SD* = 1.29) than when it was 20% (i.e., low benchmark; *M* = 3.25; *SD* = 1.48), *t*(199) = 8.00, *p* < .001, *d* = 1.13. Perceived personal impact (α = .88) was also higher with the high overhead benchmark (*M* = 4.47; *SD* = 1.30) than with the low benchmark (*M* = 3.44; *SD* = 1.30), *t*(199) = 5.64, *p* < .001, *d* = .80. Finally, perceived charity efficiency (α = .93) was also higher with the high benchmark (*M* = 5.05; *SD* = 1.04) than with the low benchmark (*M* = 3.57; *SD* = 1.33), *t*(199) = 8.78, *p* < .001, *d* = 1.24.

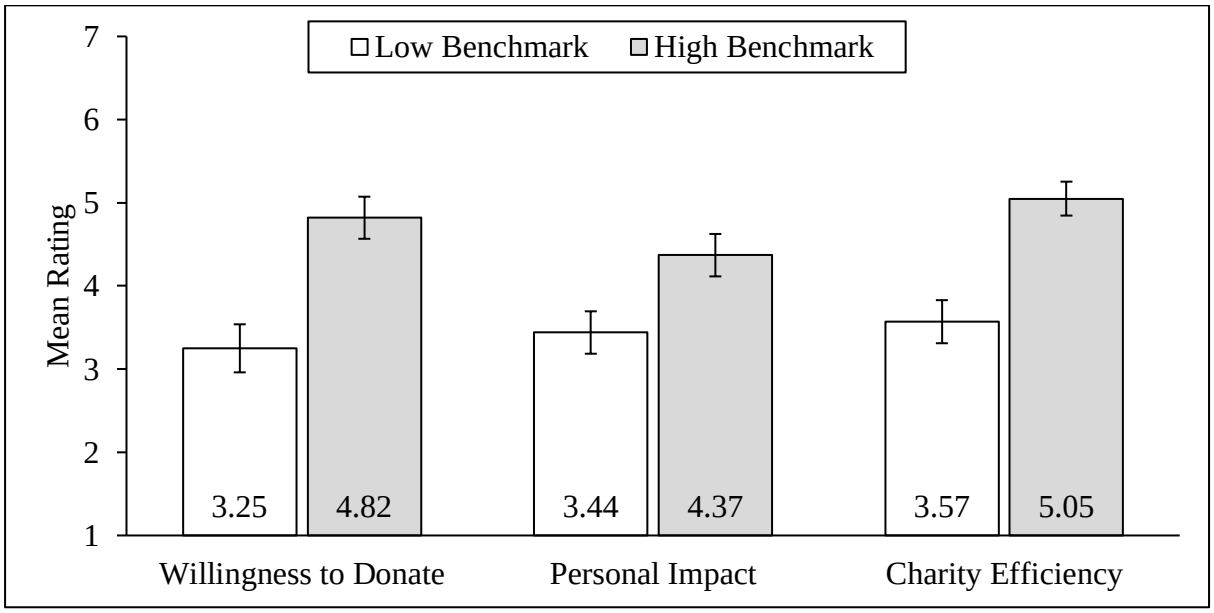

Figure 1. Mean ratings (Study 2).
Notes: Error bars represent 95% confidence intervals.

Process. We first tested for discriminant validity between the presumed mediators (i.e., personal impact and charity efficiency) via a principal components analysis (PCA) with Varimax rotation and two factors. The three items within each measure loaded strongly onto their intended factor (all loadings > .80) and not on the other factor (all cross-loadings < .40), and each factor accounted for a substantial portion of variance in ratings (impact = 40%, efficiency = 44%). Thus, the two presumed mediators exhibited discriminant validity.

Both perceived personal impact and charity efficiency can increase when the benchmark is higher (making relative overhead ratio of the target charity lower), and personal

impact and charity efficiency were significantly correlated, $r = .61$, $p < .001$. A statistical concern about moderately correlated factors such as these is that, when entered simultaneously as predictors of some outcome variable, multicollinearity may distort the effect sizes (and significance) of the predictors. We therefore tested for multicollinearity in a linear regression by entering both personal impact and charity efficiency as predictors of willingness to donate. The variance inflation factor (VIF) was 1.59, which is widely considered acceptable, indicating that multicollinearity was negligible (Johnston, Jones, and Manley 2018).

We next examined the correlations of the two presumed mediators with the outcome variable, willingness to donate. Both personal impact ($r = .69$, $p < .001$) and charity efficiency ($r = .65$, $p < .001$) correlated significantly with willingness to donate. In terms of mediation analysis, the correlation with charity efficiency was barely within the "sweetspot" for meaningful mediation (see Pieters 2017; assuming a lenient criterion with 90% reliability), whereas the correlation with personal impact was outside the sweetspot. Thus, the following mediation analyses should be interpreted cautiously.

Finally, we conducted bootstrap mediation analysis (Hayes 2017, model 4, 10K samples) with benchmark (low = 0, high = 1) as independent variable, charity efficiency and personal impact as parallel mediators, and willingness to donate as dependent variable. The effect of benchmarking on willingness to donate was significantly mediated by both perceived personal impact, $B = .53$, $SE = .13$, $CI_{95} = (.30, .80)$, and perceived charity efficiency, $B = .48$, $SE = .17$, $CI_{95} = (.17, .85)$.

Discussion. These results support **H1** and **H2**. That is, despite the charity's overhead ratio remaining constant across conditions (i.e., overhead = 35%), participants were more willing to donate to that charity when the overhead ratio of other similar charities was higher (i.e., benchmark = 50%) than when it was lower (i.e., benchmark = 20%). This effect of

benchmarking was large (d = 1.13), and it was mediated not only by perceived personal impact, but also by perceived charity efficiency. While personal impact has been shown to mediate donations (Gneezy, Keenan, and Gneezy 2014), showing perceived charity efficiency as a mediator is novel. Thus, evidently, a higher overhead benchmark increases donations to a target charity (**H1**) at least partially by making that charity appear relatively efficient (**H2**).

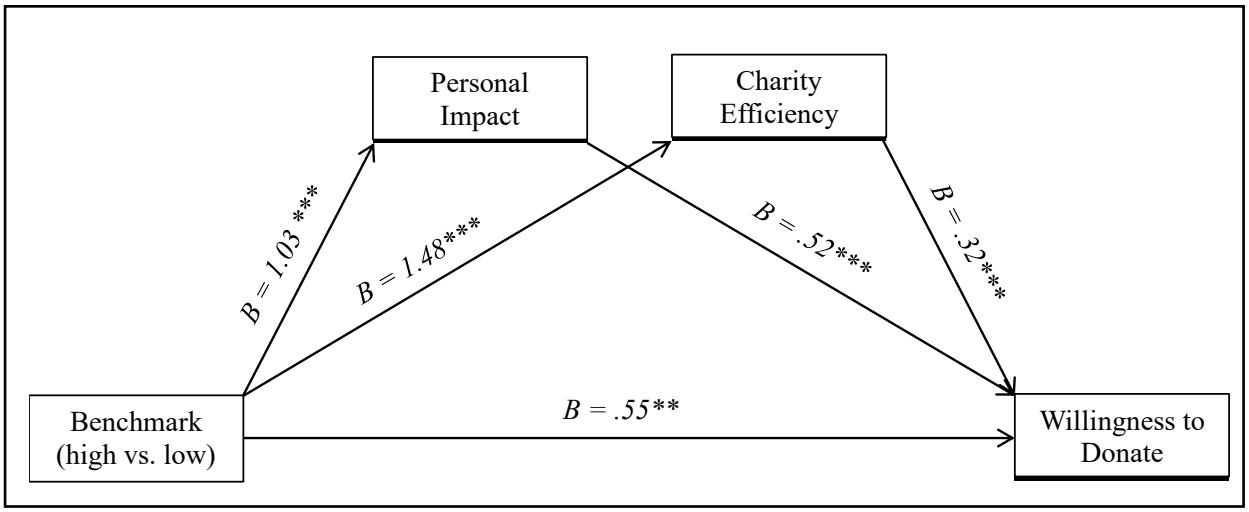

Figure 2. Results of mediation analysis (Study 2).
** $p < .01$.
*** $p < .001$.

STUDY 3 – Using Benchmarks to Reverse Overhead Aversion

Study 3 again tested whether benchmarking can mitigate overhead aversion in donation decisions, but also tested whether it could reverse it (**H3**). In this study, participants' choices were consequential – they were asked to give $100 to one of two charities, and were told that we would randomly select one participant and donate the $100 to that participant's chosen charity on their behalf. Participants read brief descriptions of two real charities (Kids Korps and charity:water) that differed in overhead ratios: one had low overhead (15%) and the other had high overhead (25%). Due to overhead aversion, people should reliably choose to donate the $100 to the charity with lower overhead, as Gneezy, Keenan, and Gneezy

(2014) demonstrated. Critically, however, we presented different overhead benchmarks that contrasted from the target charities' differing overhead ratios. The low-overhead charity was presented with an even lower benchmark (overhead = 15%, benchmark = 5%), whereas the high overhead charity was presented with an even higher benchmark (overhead = 25%, benchmark = 35%).

For example, one of the counterbalanced scenarios read: "About 25% of the money raised by Kids Korps goes towards overhead cost (administrative and fundraising costs). The average overhead cost for similar charities is 35%. About 15% of the money raised by charity:water goes towards overhead cost (administrative and fundraising costs). The average overhead cost for similar charities is 5%." We predicted that these overhead benchmarks would reverse overhead aversion. That is, we predicted that participants would reliably choose to donate the $100 to the charity with the higher overhead ratio (in this scenario, Kids Korps).

Method

Participants and design. Two hundred U.S. residents (*female* = 121; M_{age} = 32.77; SD_{age} = 11.83), recruited from Prolific, participated for £0.12. The study was a simple choice task. All participants chose to donate to one of two charities that differed in overhead ratios and benchmarks, as described next.

Procedure. The study followed the procedure of Gneezy, Keenan, and Gneezy (2014), except that we additionally provided overhead benchmarks. After completing the attention check (see Study 2), participants were informed that they would be asked to give $100 to one of two charities, and that we would randomly select one participant and donate the $100 to that participant's chosen charity on their behalf. Upon advancing to the next page, participants read brief descriptions of two charities, Kids Korps and charity:water. Included in each charity's description was its own overhead ratio, and an overhead benchmark. One

charity was described as having 15% overhead whereas other similar charities had a 5% overhead (i.e., low overhead, lower benchmark), and the other charity was described as having a 25% overhead whereas other similar charities had a 35% overhead (i.e., high overhead, higher benchmark).

Across participants, the two charities (Kids Korps and charity:water) were randomly assigned across "low overhead, lower benchmark" and "high overhead, higher benchmark" conditions; they also appeared first or second in random order. To ensure that participants understood the scenario, they were required to answer four comprehension check questions correctly before advancing to the choice task: They were required to confirm the overhead costs for both charities (options: 15%, 25%) and for their respective benchmarks (options: 5%, 35%). Participants were allowed to change any incorrect responses and could advance to the choice task only after getting all four responses correct. We then asked them to "Please tell us which organization you would like to give $100 to" (options: Kids Korps, charity:water). Finally, participants reported their gender and age. At the conclusion of the study, we randomly selected one participant, donated $100 to his chosen charity (charity:water), and informed him of the donation on his behalf. Per **H3**, we predicted that participants would choose to donate the $100 to the charity with the higher absolute overhead ratio, but lower relative overhead ratio.

Results and Discussion

As predicted, significantly more participants (64%) chose to donate the $100 to the high overhead (=25%), higher benchmark (=35%) charity than to the low overhead (=15%), lower benchmark (=5%) charity, $\chi^2(1, 200) = 15.68$, $p < .001$, $d = .59$. Remarkably, people chose to donate to the charity with higher overhead costs (overhead = 25%) than to an alternative charity with lower overhead costs (overhead = 15%). The benchmark thus

reversed overhead aversion, revealing that benchmarking is an effective strategy to mitigate overhead aversion. Thus, study 3 supports **H3**.

Note that this study pits the two mediators against one another. If donation decisions are determined largely by personal impact, then participants should choose to donate to the charity with lower overhead, because more of their donation would directly impact the cause. In the introduction to this study, we described a scenario in which Kids Korps had overhead costs of 25% but with a 35% benchmark, whereas charity:water had an overhead of 15% with a benchmark of 5%. In this scenario, if donation decisions are based on personal impact, people should prefer to donate to charity:water (overhead = 15%) rather than Kids Korps (overhead = 25%). If people instead prefer to donate to Kids Korps, that would provide strong evidence that charity efficiency affects donation decisions, because Kids Korps appears relatively more efficient (overhead = 25%, benchmark = 35%) than charity:water (overhead = 15%, benchmark=5%). In this way, we can observe whether charity efficiency affects donation decisions even when personal impact works against it. We find that more participants choose the high overhead charity, indicating that they made their choices based on relative efficiency rather than on personal impact, supporting **H2**. However, we did not measure perceived personal impact or perceived charity efficiency. We do this in the next study, while also providing a direct test of overhead aversion without benchmarks.

STUDY 4: Role of Perceived Charity Efficiency in Reversal of Overhead Aversion

Study 3 showed a reversal of overhead aversion in a single choice between two charities, and efficiency was inferred as the mediator. Study 4 tested in a more comprehensive manner whether benchmarking mitigates (and even reverses) overhead aversion (**H3**) and tests more directly whether perceived charity efficiency mediates the effect of overhead benchmarks on charitable donations (**H2**). All participants read a brief scenario

about the Make-A-Wish Foundation. There were two control and two benchmarking conditions. Participants in the control condition were told that the charity has low (15%) or high (35%) overhead. Participants in the benchmark condition additionally read that the average overhead cost for similar charities is lower at 5% (when target charity's overhead is 15%) or higher at 45% (when target charity's overhead is 35%). We predicted that, in the absence of a benchmark, low overhead will elicit higher willingness to donate than high overhead (i.e., overhead aversion). However, in the presence of a benchmark, high overhead (with higher benchmark) will elicit higher willingness to donate than low overhead (with lower benchmark) (i.e., the reversal of overhead aversion, as shown initially in Study 3). In addition, we predicted mediation via efficiency.

Method

Participants and design. Six hundred US residents (*female* = 486; M_{age} = 27.48; SD_{age} = 9.05), recruited from Prolific, participated for £0.31. We randomly assigned participants to conditions of a 2 (high vs. low overhead) × 2 (benchmarking vs. control) between-participant design.

Procedure. Participants first reported their gender and age and completed the attention check (see Study 2). Then, they read a brief scenario about the Make-A-Wish Foundation, which fulfills life-changing wishes for children with critical illnesses. The scenario varied in terms of the charity's overhead costs (low = 15%, high = 35%). Participants in the control conditions were told that "Make-A-Wish Foundation allocates 15% (35%) of all expenses to overhead (i.e., administrative and fundraising costs). That is, for every dollar donated, 15 (35) cents are used to cover the costs." Participants in the benchmark conditions were additionally told that "The average overhead cost for similar charities is 5% (45%)". Thus, the low overhead condition was paired with a lower benchmark and the high overhead condition was paired with a higher benchmark. See the four conditions in Table 2. We did not include

comprehension checks in this study. As in Study 3, participants indicated their willingness to donate (3 items; α = .91) and then, on separate and randomly ordered pages, evaluated their perceived personal impact (3 items; α = .84) and charity efficiency (3 items; α = .89).

Table 2: Experimental conditions used in Study 4.

	Control	Benchmark
Low Overhead	15% Overhead	15% Overhead 5% Benchmark
High Overhead	35% Overhead	35% Overhead 45% Benchmark

Results and Discussion

Effects. As predicted, a 2 (overhead) × 2 (benchmarking) ANOVA on willingness to donate revealed that overhead and benchmarking had a significant interaction effect on willingness to donate, $F(1, 596) = 20.89$, $p < .001$. In the control condition, when there was no benchmark, willingness to donate was higher with low overhead ($M = 5.25$; $SD = 1.22$) than with high overhead ($M = 4.89$; $SD = 1.37$), $t(298) = 2.42$, $p = .016$, $d = .28$, demonstrating overhead aversion. In contrast, when the benchmarks were provided, willingness to donate was higher in the high overhead-higher benchmark condition ($M = 5.07$; $SD = 1.28$) than in the low overhead-lower benchmark condition ($M = 4.42$; $SD = 1.55$), $t(298) = 3.97$, $p < .001$, $d = .46$. This indicates a reversal of overhead aversion, as initially shown in Study 3.

The overhead × benchmarking interaction was also significant in driving perceived personal impact, $F(1, 596) = 8.65$, $p = .003$. In the control condition with no benchmarks, personal impact was higher with low overhead ($M = 4.46$; $SD = 1.35$) than with high overhead ($M = 4.09$; $SD = 1.41$), $t(298) = 2.33$, $p = .020$, $d = .27$, as expected. In the

benchmark conditions, however, perceived personal impact was marginally lower with low overhead-lower benchmark (M = 3.90; SD = 1.42) than with high overhead-higher benchmark (M = 4.20; SD = 1.40), $t(298) = 1.83$, $p = .068$, $d = .21$.

Finally, the interaction was also significant in driving perceived charity efficiency, $F(1, 596) = 28.25$, $p < .001$. In the control conditions, charity efficiency was higher with low overhead (M = 5.64; SD = 1.00) than with high overhead (M = 5.36; SD = 1.09), $t(298) = 2.38$, $p = .018$, $d = .28$, indicating that overhead ratio is perceived as a cue of charity efficiency. In the benchmark conditions, however, charity efficiency was significantly lower with low overhead and lower benchmark (M = 5.00; SD = 1.23) than with high overhead and higher benchmark (M = 5.66; SD = 1.04), $t(298) = 5.03$, $p < .001$, $d = .58$. Thus, the benchmarks strongly affected perceived charity efficiency. The mean ratings are provided in Figure 3.

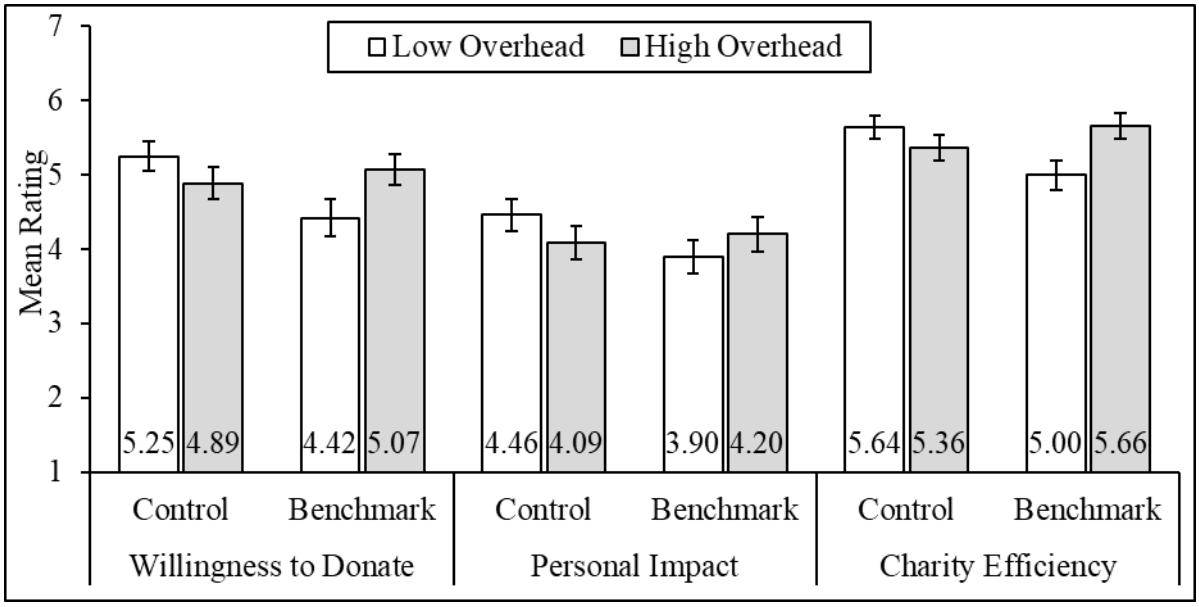

Figure 3. Mean ratings (Study 4).
Notes: Overhead was either low (15%) or high (35%). In the control conditions, only the charity's overhead was presented. In the benchmark conditions, the low overhead (15%) was additionally presented with a lower benchmark (5%) whereas the high overhead (35%) was presented with a higher benchmark (45%). Error bars represent 95% confidence intervals.

Process. Before testing for mediation, we first tested for discriminant validity between the presumed mediators (i.e., personal impact and charity efficiency), as in Study 2. PCA confirmed that the three items within each measure loaded strongly onto their intended factor (all loadings > .80) and not on the other factor (all cross-loadings < .40), and each factor accounted for a substantial portion of variance in ratings (impact = 38%, efficiency = 42%). Thus, the two presumed mediators exhibited discriminant validity. Personal impact and charity efficiency were significantly correlated, $r = .47$, $p < .001$. However, a linear regression with personal impact and charity efficiency as predictors of willingness to donate revealed a VIF of 1.28, indicating that multicollinearity was negligible. We next examined the correlations of the two presumed mediators with the outcome variable, willingness to donate. Both personal impact ($r = .52$, $p < .001$) and charity efficiency ($r = .59$, $p < .001$) correlated significantly with willingness to donate, and both of those correlations were within the "sweetspot" for meaningful mediation, even with Pieters' (2017) strict criterion.

We therefore conducted moderated mediation analysis (Hayes 2017, model 7, 10K samples) with overhead (low = 0, high = 1) as independent variable, benchmark (low = 0, high = 1) as moderator on the A-path, personal impact and charity efficiency as parallel mediators, and willingness to donate as dependent variable. Note that we are theoretically agnostic regarding moderated mediation: We specifically predicted mediation via efficiency in the benchmark condition (**H2**). However, because we expected that efficiency may also mediate the effect within the control condition, it was theoretically unclear whether moderated mediation should occur. Moreover, we expected that impact may also mediate the effects in both control and benchmark conditions, but that was not our primary theoretical interest. As such, we preregistered that **H2** would be supported if there is either unmoderated mediation via efficiency ("outcome a"), or moderated mediation via efficiency AND significant mediation via efficiency in the benchmark condition ("outcome b").

Results indicated outcome b (See Figure 4). The index of moderated mediation was significant for both personal impact, B = .20, SE = .08, CI_{95} = (.06, .36), and charity efficiency, B = .53, SE = .12, CI_{95} = (.31, .77). In the control conditions, both personal impact, B = -.11, SE = .05, CI_{95} = (-.22, -.02), and charity efficiency, B = -.16, SE = .07, CI_{95} = (-.31, -.03), significantly mediated the effect of overhead ratio on willingness to donate. That is, overhead aversion was mediated not only by personal impact (Gneezy, Keenan, and Gneezy 2014), but also by charity efficiency. More importantly for the present purposes, in the benchmark conditions, charity efficiency significantly mediated the effect of overhead ratio on willingness to donate, B = .37, SE = .08, CI_{95} = (.21, .54), but personal impact did not, B = .09, SE = .05, CI_{95} = (-.01, .19). That is, in the presence of an overhead benchmark, perceived charity efficiency underlies the effect of overhead ratio on willingness to donate. These results thus support both **H2** and **H3**.

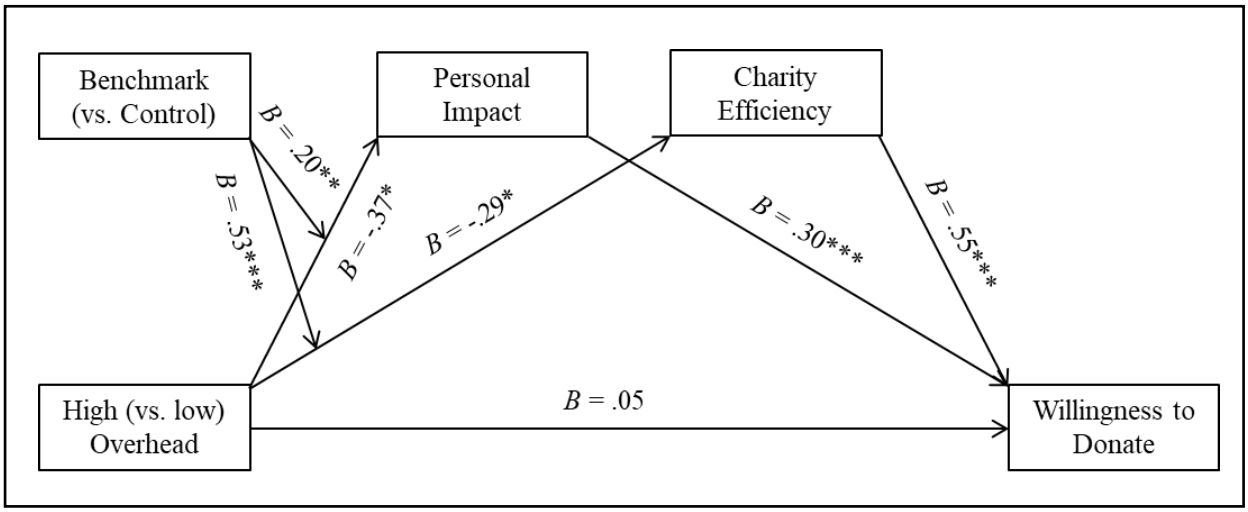

Figure 4. Results of moderated mediation analysis (Study 4).
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Discussion. These results provide evidence that in the absence of a benchmark, low overhead (15%) elicited higher willingness to donate than high overhead (35%) (i.e.,

overhead aversion). However, in the presence of a benchmark, high overhead (with higher benchmark) elicited higher willingness to donate than low overhead (with lower benchmark = 5%; i.e., the reversal of overhead aversion, as shown initially in Study 3). Study 4 also provides additional evidence via mediation that perceived charity efficiency underlies the effect of overhead benchmarks on donation decisions.

GENERAL DISCUSSION

Charitable donations positively impact individual and societal well-being (Oppenheimer 2015). A critical psychological barrier to donating is overhead aversion, whereby donors prefer not to pay for overhead costs with their donations (Caviola, et al., 2014; Charles, Sloan, and Schubert, 2020; Gneezy, Keenan, and Gneezy 2014; Portillo and Stinn 2018). Indeed, overhead-related information is among the most-searched when donors research charities (Hope Consulting 2010). Thus, many individual charities and some charity aggregators report the charities' overhead ratio. On its own, however, a charity's overhead ratio is difficult to interpret. For instance, is 25% a high overhead ratio? In our studies we used overhead benchmarks to provide a relative context for interpreting overhead ratios. For instance, 25% overhead appears high when other similar charities have overhead ratios of 15%, but it appears low when other charities have 35% overhead. Our studies show that such overhead benchmarks can mitigate donors' overhead aversion, thereby increasing donations for charities with relatively low overhead.

Theoretical Contributions

Bowman (2006) first identified overhead aversion as a substantial barrier to charitable giving, but it was Gneezy, Keenan, and Gneezy (2014) who brought the phenomenon to the attention of the broader scientific community. Since then, a rapidly growing literature addressing this problem has emerged. The subsequent studies have been of two main types:

most prior research has sought strategies to mitigate overhead aversion, whereas recently a few researchers have attempted to shed light on what underlies this aversion. The present research provides important theoretical contributions to both these related questions (i.e., process and mitigation), as elaborated below.

First, this research provides a novel explanation of the psychological processes underlying overhead aversion. As stated above, very little prior research has investigated this question. Gneezy, Keenan, and Gneezy (2014) argued that overhead aversion arises from donors' desire for personal impact on the cause. That is, donors want their money to go to the given cause, not to the administration of the charity. Therefore, when the overhead costs are paid by a third-party donor, and hence the target donor's personal impact on the cause increases, donations also increase. We also found clear evidence of this in our Study 4. When we provided no overhead benchmark to participants, perceived personal impact partially mediated the effect of overhead (low vs. high) on willingness to donate.

Of course, like most effects, overhead aversion is likely to have multiple underlying causes. Personal impact is unlikely to be the sole psychological process by which overhead aversion operates. Indeed, the present research reveals an additional process underlying overhead aversion: perceived charity efficiency. In our Study 4, we found that personal impact and charity efficiency both mediated the overhead aversion effect in the absence of a benchmark. No prior research has shown that charity efficiency contributes to overhead aversion effects. In fact, Gneezy, Keenan, and Gneezy (2014) argued that perceived charity efficiency does not contribute to overhead aversion. Notably, however, they did not measure or otherwise control for charity efficiency. Our results reveal for the first time that personal impact and charity efficiency are empirically distinct, and that they both contribute to overhead aversion effects. We consider this to be a fundamental contribution, in that it

explains the basic phenomenon in a way not previously recognized by any of the prior studies of this phenomenon.

A second theoretical contribution is that this research reveals a novel strategy to mitigate overhead aversion in charitable giving. As described above, Gneezy, Keenan, and Gneezy (2014) showed in a large field experiment that using a third-party donor's contribution to pay overhead costs, so that other potential donors could perceive greater personal impact with their donations, significantly increased donations. Qu and Daniel (2021) recently showed that when prospective donors are informed that a relatively high overhead is intended to improve long-term organizational capacity, overhead aversion diminishes. Consistent with Qu and Daniel (2021), we argue that people might not be inherently overhead averse. In fact, we reasoned that overhead could be used strategically to *encourage* potential donors to donate. Assuming that perceived charity efficiency contributes to overhead aversion (Study 4), we hypothesized that providing an overhead benchmark would influence prospective donors' perceptions of the efficiency of a given charity. That is, donors may care more about the charity's *relative* overhead than its absolute overhead. Prospective donors may not know what a typical overhead ratio is. By providing an overhead benchmark, we reveal and emphasize whether the given charity has relatively high or low overhead. And when the charity's overhead is shown to be relatively low, willingness to donate increases. Thus, the present research reveals that benchmarking can be a simple and effective strategy to mitigate overhead aversion and increase donations.

Three further aspects of this mitigation strategy are noteworthy. First, the effect appears to be moderately large. Across the four studies, the average *N*-weighted effect size is $d = .53$. Second, it bears emphasis that benchmarking reliably reversed overhead aversion. In both Studies 3 and 4, we pitted a high-overhead charity with a higher benchmark (e.g., overhead = 35%, benchmark = 45%) against a low-overhead charity with a lower benchmark

(e.g., overhead = 15%, benchmark = 5%). And remarkably, in both studies, participants reliably preferred to donate to the charity with higher overhead. For instance, participants were more willing to donate to the charity with 35% overhead than to the charity with 15% overhead, because the two benchmarks made them appear respectively efficient or inefficient. As stated earlier, if charity websites want to move donors from low overhead to high overhead industries, one strategy may be to provide the average overhead ratio for the industry for each charity.

A final noteworthy point is that this novel mitigation strategy appears to rely more strongly on perceived charity efficiency than on donors' personal impact. In Study 2, both personal impact and charity efficiency mediated the effect of benchmarking on willingness to donate. In the benchmark conditions of Study 4, however, charity efficiency mediated the effect of overhead ratio on willingness to donate, whereas personal impact did not. Collectively, then, we found that benchmarking was strong enough to reverse overhead aversion, and that this mitigation strategy operates primarily via perceived charity efficiency.

Practical Implications

This research has clear implications for charitable organizations: Providing an overhead benchmark can increase donations for charities with low overheads compared to their competition. Moreover, we believe that this overhead mitigation strategy would be especially effective in high-overhead industries. Overhead aversion is a particularly acute problem in charitable industries that naturally have high overhead costs, such as museums, which typically incur high costs to purchase and maintain collections and curations that attract large audiences, and which typically display their artworks and artifacts in expensive city center locations. In such industries, what may appear to prospective donors as high overhead (e.g., 23%) may not signal organizational inefficiency, as donors may infer (Crutchfield and Grant 2008). Thus, by simply informing prospective donors of the relevant

benchmark (i.e., average overhead among arts and culture charities is 25-30%), those prospective donors perceive the given charity as more efficient, and hence they are more likely to convert from prospective to actual donors.

If we view donations as a zero-sum game, then benchmarking would presumably shift a greater share of the donations within any given industry to low-overhead charities, at the expense of high-overhead charities. Upon initial consideration, this may sound unequivocally positive, as favoring low-overhead charities would (a) imply favoring more efficient charities, and (b) incentivize an industry-wide shift toward greater organizational efficiency. However, we urge caution toward this interpretation, because as we have argued, high overhead does not necessarily indicate inefficiency. The Metropolitan Opera House in New York City could have high overhead because it offers superior performances by world-renowned artists in the center of an expensive city; this should not be misinterpreted as organizational inefficiency. Hence, a shift of donations toward low-overhead charities does not necessarily imply increased efficiency. On the contrary, this could even perpetuate the non-profit starvation cycle (Gregory and Howard 2009; Lecy and Searing 2015), whereby cutting overhead costs ultimately reduces the charity's efficiency and donors' impact.

Alternatively, if we view donation amounts as variable rather than fixed (i.e., rather than zero-sum), then benchmarking could increase donations in general. For example, rather than focusing donations on low-overhead charities at the expense of high-overhead charities, benchmarking could increase donations to low-overhead charities without decreasing donations to high-overhead charities. Given that many efficient charities have high overhead, this is our hope. By increasing the total amount of donations, benchmarking could increase societal well-being.

A related but perhaps less obvious practical implication of our research is that benchmarking could also increase the well-being of individual donors and recipients. To the

extent that charitable giving enriches the lives of both donors and recipients (Oppenheimer 2015), benchmarking could increase individual well-being by increasing donations.

Future Directions

Our findings suggest using benchmarks as an effective strategy to mitigate overhead aversion, specifically for otherwise effective charities (in high overhead domains) and points to the role of relative thinking and charity efficiency perceptions. In so doing, our research provides several paths for future exploration.

First, we used as benchmarks the average overhead ratio of "other similar charities". As such, our results suggest that only charities with relatively low overhead in otherwise high-overhead domains would benefit from benchmarking. For instance, the Imagine Children's Museum would benefit from benchmarking because its overhead ratio of 16.6% is lower than the average of 24.3% for all children's museums. However, we see no theoretical reason to assume that the same effect wouldn't also occur in relatively low-overhead domains such as food banks. While it is true that the Shared Harvest Foodbank's overhead ratio of 3.7% (see Table 1) is objectively low and therefore likely to attract donations, further informing potential donors that similar other charities have an average overhead of 6.2% may well still boost Shared Harvest's donations. The present research does not reveal whether there is a lower limit to the effectiveness of benchmarking, and we believe this is an important consideration for future research.

Relatedly, we have chosen to focus on average overhead ratios for the benchmarks in our studies, but other benchmarks are possible. For example, suppose that the average overhead ratio across all charities – regardless of industry – was 20%. Presenting this as a benchmark would render nearly all charities in low-overhead industries as relatively more efficient, and could increase their donations. We do not suggest that this would necessarily be a wise or fair strategy. Rather, we suggest that exactly what is the most appropriate

benchmark is a very important issue for further consideration by charities and watchdogs alike.

Another potentially informative direction for further research is to identify conditions under which benchmarking is unlikely to affect donations. Our research reveals that relative thinking is at the core of benchmarking as an intervention strategy, and our findings suggest charity efficiency as the underlying mechanism. However, several boundary conditions of relative thinking have been documented. For example, relative thinking is reduced when absolute saving is high (Moon, Keasey, and Duxbury 1999) or it is difficult to assess the value of a sales promotion relative to the focal product or its price (Nunes and Park 2003). In the context of overhead aversion, this may translate into the idea that when absolute overhead ratio is low or donor knowledge on overhead is minimal, relative thinking may diminish and benchmarking may be less effective. We leave this to future research.

Finally, we focused on the underlying mechanism and suggested charity efficiency perceptions as a novel explanation of the psychological processes underlying overhead aversion. This implies that charity efficiency perceptions can provide a useful research direction in guiding future intervention strategies aiming at mitigating overhead aversion and increasing donations.

REFERENCES

Aaker, Jennifer, Kathleen D. Vohs, and Cassie Mogilner (2010), "Nonprofits are seen as warm and for-profits as competent: Firm stereotypes matter," *Journal of Consumer Research*, 37(2), 224-237.

Andreoni, James (1990), "Impure altruism and donations to public goods: A theory of warm-glow giving," *The Economic Journal*, 100(401), 464-477.

Anik, Lalin, Michael I. Norton, and Dan Ariely (2014), "Contingent match incentives increase donations," *Journal of Marketing Research*, 51(6), 790-801.

Azar, Ofer H. (2007), "Relative thinking theory," *The Journal of Socio-Economics*, 36(1), 1-14.

Bekkers, René, and Pamala Wiepking (2011), "A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving," *Nonprofit and Voluntary Sector Quarterly*, 40(5), 924-973.

Bowman, Woods (2006), "Should donors care about overhead costs? Do they care?," *Nonprofit and Voluntary Sector Quarterly*, 35(2), 288-310.

Buchheit, Steve, and Linda M. Parsons (2006), "An experimental investigation of accounting information's influence on the individual giving process," *Journal of Accounting and Public Policy*, 25(6), 666-686.

Carroll, Ryall, and Luke Kachersky (2019), "Service fundraising and the role of perceived donation efficacy in individual charitable giving," *Journal of Business Research* 99, 254-263.

Caviola, Lucius, Nadira Faulmüller, Jim A. C. Everett, Julian Savulescu, and Guy Kahane (2014), "The evaluability bias in charitable giving: Saving administration costs or saving lives?," *Judgment and Decision Making*, 9(4), 303-315.

Charity Navigator (2020, May 21). *Financial score conversions and tables*. Retrieved

January 3, 2022, from https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=48

Charles, Cleopatra, Margaret F. Sloan, and Peter Schubert (2020), "If someone else pays for overhead, do donors still care?," *The American Review of Public Administration*, 50(4/5), 415-427.

Coupet, Jason, and Jessica L. Berrett (2019), "Toward a valid approach to nonprofit efficiency measurement." *Nonprofit Management and Leadership*, 29(3), 299-320.

Crutchfield, Leslie R., and Heather McLeod Grant (2012). *Forces for good: The six practices of high-impact nonprofits*. John Wiley & Sons.

Duncan, Brian (2004), "A theory of impact philanthropy," *Journal of Public Economics*, 88(9/10), 2159-2180.

Fetherstonhaugh, David, Paul Slovic, Stephen Johnson, and James Friedrich. (1997), "Insensitivity to the value of human life: A study of psychophysical numbing," *Journal of Risk and Uncertainty*, 14(3), 283-300.

Finucane, Melissa L., Ali Alhakami, Paul Slovic, and Stephen M. Johnson. (2000), "The affect heuristic in judgments of risks and benefits," *Journal of Behavioral Decision Making* 13(1): 1-17.

Gabor, Andre (1964), "Price sensitivity of the consumer," *Journal of Advertising Research*, 4(4), 40-44.

Ghoshal, Sumantra, and D. Eleanor Westney (1991), "Organizing competitor analysis systems." *Strategic Management Journal*, 12(1), 17-31.

Giving USA (2021) *2021 Annual Report***.** Giving USA. Retrieved January 3, 2022, from https://store.givingusa.org/products/2021-annual-report.

Gneezy, Uri, Elizabeth A. Keenan, and Ayelet Gneezy (2014), "Avoiding overhead aversion in charity," *Science*, 346(6209): 632-635.

Goswami, Indranil, and Oleg Urminsky (2016), "When should the ask be a nudge? The effect of default amounts on charitable donations," *Journal of Marketing Research*, 53(5), 829-846.

Gregory, Ann Goggins, and Don Howard (2009), "The nonprofit starvation cycle," *Stanford Social Innovation Review*, 7(4), 49-53.

Grey Matter Research (2018). *The Donor Mindset Study IV*. Grey Matter Research; Op4G.

Hammer, Michael, and James Champy (1993), Business process reengineering. *London: Nicholas Brealey*, 444(10), 730-755.

Hardie, Bruce GS, Eric J. Johnson, and Peter S. Fader (1993), "Modeling loss aversion and reference dependence effects on brand choice," *Marketing Science*, 12(4), 378-394.

Hayes, Andrew F. (2017), *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Hope Consulting (2010) Money for good: the U.S. market for impact investments and charitable gifts from individual donors and investors.

Johnston, Ron, Kelvin Jones, and David Manley (2018), "Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour," *Quality & Quantity*, 52(4), 1957-1976.

Kahneman, Daniel, and Amos Tversky (1979), "Prospect theory: An analysis of decision under risk," *Econometrica*, 47(2), 363-391

Kalwani, Manohar U., Chi Kin Yim, Heikki J. Rinne, and Yoshi Sugita (1990), "A price expectations model of customer brand choice," *Journal of Marketing Research*, 27(3), 251-262.

Lecy, Jesse D., and Elizabeth A. M. Searing (2015), "Anatomy of the nonprofit starvation cycle: An analysis of falling overhead ratios in the nonprofit sector," *Nonprofit and Voluntary Sector Quarterly*, 44(3), 539-563.

McDowell, Evelyn A., Wei Li, and Pamela C. Smith. (2013), "An experimental examination of US individual donors' information needs and use," *Financial Accountability & Management,* 29(3), 327-347.

Meier, Stephan (2007), "Do subsidies increase charitable giving in the long run? Matching donations in a field experiment," *Journal of the European Economic Association*, 5(6), 1203-1222.

Mesch Debra J., Melissa S. Brown, Zachary I. Moore, and Amir D. Hayat (2011). Gender differences in charitable giving. *International Journal of Nonprofit and Voluntary Sector Marketing*, *16*(4), 342-355.

Metzger, Laura, and Isabel Günther (2019), "Making an impact? The relevance of information on aid effectiveness for charitable giving. A laboratory experiment," *Journal of Development Economics*, 136, 18-33.

Miller, Richard L. (1962), "Dr. Weber and the Consumer," *Journal of Marketing,* 26(1), 57-61.

Monroe, Kent B. (2003) Pricing Making Profitable Decisions. 3rd Edition, McGraw-Hill/Irwin, Boston, London.

Moon, Philip, Kevin Keasey, and Darren Duxbury (1999), "Mental accounting and decision making: The relationship between relative and absolute savings," *Journal of Economic Behavior & Organization,* 38(2), 145-153.

Newman, George E., Adam Shniderman, Daylian M. Cain, and Kyle Sevel (2019), "Do the ends justify the means? The relative focus on overhead versus outcomes in charitable fundraising," *Nonprofit and Voluntary Sector Quarterly* 48(1), 71-90.

Nunes, Joseph C., and C. Whan Park (2003), "Incommensurate resources: Not just more of the same." *Journal of Marketing Research,* 40(1), 26-38.

Oppenheimer, Daniel M. (2015), "Increasing donations and improving donor experiences: Lessons from decision science," *Policy Insights from the Behavioral and Brain Sciences,* 2(1), 203-210.

Pieters, Rik (2017), "Meaningful mediation analysis: Plausible causal inference and informative communication," *Journal of Consumer Research* 44(3), 692-716.

Portillo, Javier E., and Joseph Stinn (2018), "Overhead aversion: Do some types of overhead matter more than others?," *Journal of Behavioral and Experimental Economics,* 72, 40-50.

Pruitt, Anna, and Jon Bergdoll (2021, June 17). *Americans gave a record $471 billion to charity in 2020*. Fast Company. Retrieved October 15, 2021, from https://www.fastcompany.com/90647835/americans-gave-a-record-471-billiontocharity-in-2020.

Qu, Heng, and Jamie Levine Daniel (2021), "Is "Overhead" A Tainted Word? A Survey Experiment Exploring Framing Effects of Nonprofit Overhead on Donor Decision," *Nonprofit and Voluntary Sector Quarterly* 50(2), 397-419.

Reiling, Herman T. (1958), "Federal Taxation: What Is a Charitable Organization?," *American Bar Association Journal*, 525-598.

Stout, William D. (2001), "Planning materiality in audits of nonprofit organizations," *The CPA Journal*, 71(12), 47-48.

Thaler, Richard (1980), "Toward a positive theory of consumer choice," *Journal of Economic Behavior & Organization* 1(1): 39-60.

Tian, Yuan, Chiako Hung, and Peter Frumkin (2020), "Breaking the nonprofit starvation cycle? An experimental test," *Journal of Behavioral Public Administration,* 3(1), 1-19.

Tinkelman, Daniel, and Kamini Mankaney (2007), "When is administrative efficiency

associated with charitable donations?," *Nonprofit and Voluntary Sector Quarterly*, 36(1), 41-64.

Tversky, Amos, and Daniel Kahneman (1981), "The framing of decisions and the psychology of choice," *Science*, 211(4481), 453-458.

van der Heijden, Hans (2013), "Charities in competition: Effects of accounting information on donating adjustments," *Behavioral Research in Accounting*, 25(1), 1-13.

Weber, Ernest H. (1834). The sense of touch (H. E. Ross, Trans.). London: Academic Press.

Winer, Russell S. (1986), "A reference price model of brand choice for frequently purchased products," *Journal of Consumer Research*, 13(2), 250-256.