Smith ScholarWorks

Economics: Faculty Publications

Economics

3-14-2023

The Effect of Experimenter Demand on Inference

David Danz

Marissa Lepper
University of Pittsburgh

Guillermo Lezama

Priyoma Mustafi

Lise Verterlund
University of Pittsburgh

See next page for additional authors

Follow this and additional works at: https://scholarworks.smith.edu/eco_facpubs

Part of the Economics Commons

Recommended Citation

Danz, David; Lepper, Marissa; Lezama, Guillermo; Mustafi, Priyoma; Verterlund, Lise; Wilson, Alistair; and Winichakul, K. Pun, "The Effect of Experimenter Demand on Inference" (2023). Economics: Faculty Publications, Smith College, Northampton, MA.
https://scholarworks.smith.edu/eco_facpubs/102

This Article has been accepted for inclusion in Economics: Faculty Publications by an authorized administrator of Smith ScholarWorks. For more information, please contact scholarworks@smith.edu

Authors

David Danz, Marissa Lepper, Guillermo Lezama, Priyoma Mustafi, Lise Verterlund, Alistair Wilson, and K. Pun Winichakul

This article is available at Smith ScholarWorks: https://scholarworks.smith.edu/eco_facpubs/102

The Effect of Experimenter Demand on Inference

K. Pun Winichakul ⓡ Guillermo Lezama ⓡ Priyoma Mustafi ⓡ
Marissa Lepper ⓡ Alistair Wilson ⓡ David Danz ⓡ
Lise Vesterlund ⓡ

August, 2024

Abstract

To assess the threat of experimenter demand, we ask whether a hypothetical 'ill-intentioned' researcher can manipulate inference. Four classic behavioral comparative statics are evaluated, and the potential for false inference is gauged by differentially applying strong positive and negative experimenter demand across the relevant decision pair. Evaluating three different subject pools (laboratory, Prolific, and MTurk) we find no evidence of experimenter demand eliminating or reversing directional effects. The response to experimenter demand is very limited for all three subject pools and is not large enough to generate false negatives, though we do find evidence of false positives when testing precise nulls in larger online-subject pools.

1 Introduction

Experiments provide an essential tool for testing and understanding economic phenomena. By directly controlling the decision environment the experimenter can isolate and identify causal relationships that would be hard to assess with observational data. However, concerns have been raised that participants distort their behavior to align with their perception of the experimenter's hypothesis, in turn compromising inference.

While procedures to mitigate concerns for experimenter demand are widely adopted, careful experimental procedures may not be enough to defend against the critique that a result is driven by experimenter demand.[1] A defense proposed in de Quidt, Haushofer and Roth (2018) (henceforth dQHR) is to bound the potential effect on the decision estimate by deliberately inducing experimenter demand, in both positive and negative directions.[2] Using online-subject pools, dQHR demonstrate their approach across a series of small-stake economic decisions and show substantial and significant movement in response to induced demand.

Critical in assessing the potential distortions induced by experimenter demand, however, is not only the quantitative response for individual decisions, but more importantly the qualitative inference. Kessler and Vesterlund (2015) argue that the emphasis in experimental studies is on identifying the direction or sign of an effect, rather than the precise magnitude, where the communication of experimental findings centers on causal inference. Further, experimenter demand concerns often point to participants wanting to confirm a comparative-static hypothesis (see e.g., Orne, 1962).

To assess the qualitative impact of experimenter demand, we use the dQHR procedures to pose a worst-case hypothetical: can an 'ill-intentioned' experimenter manipulate both treatment and control to change an inference? We consider extreme distortions of the expected effect by (i) differentially exposing one treatment to positive demand and another treatment to negative demand, and by (ii) using what dQHR refer to as strong demand where participants are asked to do the experimenter 'a favor' by taking a higher or lower action than they normally would. Using three commonly studied populations

[1] Surveying published experimental papers de Quidt, Vesterlund and Wilson (2019) find that the vast majority of studies rely on designs that mask the hypothesis (abstract frames, between-subject designs, sequential revelation of treatments) and focus participant attention on the decision environment of interest (incentivized and anonymous decisions), while also making detailed instructions and procedures available for replication and assessment of undue influence.

[2] See also Bischoff and Frank (2011) where a professional actor (unsuccessfully) aims to induce high or low contributions and Tsutsui and Zizzo (2014) where individual measures of demand-susceptibility fail to correlate with participant decisions.

(laboratory, MTurk, and Prolific) we assess four classic behavioral comparative statics: (i) probability weighting, (ii) endowment effect, (iii) present bias, and (iv) tradeoffs between payoffs to self and others. With three of these predicting a directional effect and one a null effect we can assess the potential for both false negatives and false positives.

For the laboratory population of undergraduates we find that the quantitative decision estimates are largely insensitive to experimenter demand. Qualitatively, we confirm the expected treatment effects in all four comparative statics when we hold constant the demand environment (no demand, positive demand, and negative demand). Moreover, qualitative statistical conclusions in each of the four comparative statics are unchanged when we differentially select the demand conditions in treatment and control to purposefully weaken inference. That is, even with an extreme-inferential distortion (differential and strong experimenter demand) of the expected effect, the *qualitative* inferences are not affected.

Expanding the analysis to the online MTurk and Prolific populations we again find a small quantitative response to strong experimenter demand, where the quantitative effects are so small that when differentially applied they do not eliminate directional effects nor generate false negatives. However, when applied to the knife-edge case of a precise null we do find that extreme experimenter demand can generate false positives in the larger online samples.

The summary takeaway is largely a positive message: When applying the dQHR design to examine four classic comparative statics in three important subject populations, we find that experimenter demand is small in magnitude. With differential and strong demand, we do not find any evidence that we can change directional economic inference. A reversal of the literature comparative static is only seen in the larger online samples in the knife-edge case of a precise null.

The remainder of the paper is organized as follows. Section 2 describes our design for the laboratory population and Section 3 discusses the impact of experimenter demand on both quantitative decision estimates and on qualitative comparative statics. In Section 4, we report on results from replications on the online populations of MTurk and Prolific, which we compare to our laboratory population. Finally, in Section 5 we conclude.

2 Design

We explore the effect of experimenter demand over four classic comparative statics in behavioral economics: (i) probability weighting; (ii) endowment effect; (iii) present bias; and (iv) tradeoffs between self and others. For each of these four cases we examine a qualitative result, derived by comparing a treatment and control, A and B. Directional effects are assessed over the decision pair through the difference $\Delta x = x_A - x_B$, comparing the average choices in the two treatments, x_A and x_B, respectively. The four canonical cases are selected to vary over the qualitative effects we would expect from the literature. To assess whether experimenter demand can affect the qualitative treatment effect, we consider the actions of a hypothetical ill-intentioned experimenter and induce strong and differential demand at the decision level (cf. dQHR) to distort inference. For each decision setting, we use between-subject variation to measure a real-valued average choice without demand (x^0), with an induced positive demand (x^+) and an induced negative demand (x^-). Where positive demand results from encouraging participants to take a higher action than they normally would, and negative demand from taking a lower action. We then examine the impact of experimenter demand on qualitative treatment effects, by looking for both false negatives (when the literature would lead us to expect a directional response) and false positives (when the literature predicts a null).

For all but one of our comparative statics we examine whether strong experimenter demand can generate *false negatives* (or potentially effect reversals). By inducing *differential* demand across the A and B treatment pair we attempt to minimize the treatment effect: demanding a decrease in the average choices in A, while demanding an increase in B. That is, we examine the demand-minimized treatment-effect $\Delta x^{\ominus} = x_A^-$ - x_B^+. In an environment where the literature would predict a positive effect ($H_A : \Delta x > 0$), sufficiently large experimenter demand could lead to a failure of rejecting a null effect. On the flip side, the literature would lead us to expect a null effect for the present-bias comparative static, despite an intuitive directional prediction. For this knife-edge case we use the demand treatments to explore the possibility that experimenter demand can generate a *false positive* by rejecting the null $H_0 : \Delta x = 0$. That is, we use experimental demand to maximize the treatment effect, increasing (decreasing) the average response in the A (B) treatment through induced demand and examining the difference $\Delta x^{\oplus} = x_A^+$ - x_B^-. Similarly, we can assess impact in the opposite direction by evaluating the demand-minimized treatment-effect $\Delta x^{\ominus} = x_A^-$ - x_B^+

In what follows, we examine the potential for experimenter demand to reverse inference across our four comparative statics, drawing samples from: (i) the standard laboratory-subject population of undergraduate students, as well as the online-subject populations (ii) MTurk and (iii) Prolific. For comparison, experiments on all three populations were conducted online and over the same eight tasks. While the stakes in the MTurk and Prolific samples are scaled down to reflect ecologically valid differences for these populations, the core decision tasks are similar across the three populations.[3] Laboratory sample sizes were selected to obtain 90 percent power for all directional comparative statics using effect sizes reported in the literature, while online samples were selected to balance budgets across subject pools and to reflect the larger samples commonly seen in online studies.[4] Below we introduce the details and results from the lab sample and in Section 4 we compare results across the three populations.

2.1 Laboratory Sample

Our laboratory sample consists of 236 undergraduates recruited from the Pittsburgh Experimental Economics Laboratory (PEEL) subject population. We conducted 12 online sessions that follow the virtual laboratory procedures outlined in Danz et al. (2021) to mimic standard lab procedures. Participants make eight within-subject decisions, divided into four tasks, before completing a short demographic survey. They move through the study at their own pace but are required to listen to pre-recorded audio instructions prior to each decision. Payments were made electronically using Venmo and consisted of a $10 lump-sum and a payment based on one randomly selected decision.[5]

2.2 Demand Treatments (Between Subject)

Three between-subject treatments (randomized at the session level) manipulate the experimenter demand: (i) *no demand* (80 participants); (ii) *positive demand* (77 participants); and (iii) *negative demand* (79 participants). To bound the experimenter demand, we induce the *strong* form of experimenter demand in dQHR. That is, the three demand treatments are identical except for an additional sentence in the demand-treatment instructions. For the positive (negative) treatments the sentence is: *"You will do us a favor if you*

[3] Our experiments were pre-registered at AsPredicted for the lab (#53869), MTurk (#54625) and Prolific (#99884) samples. See Online Appendix C for reviewer links and details.

[4] See Online Appendix B for detailed calculations.

[5] All lump-sum payments and decision payments from tasks without intended delay occurred immediately after the session. We also paid Venmo fees for instant bank transfer (the maximum of 1.75% or $0.25).

take a higher (lower) action *than you normally would*."[6] The sentence appeared in red on the decision screen and was read aloud on the recorded instructions.

2.3 Task Pairs (Within Subject)

Each of the four tasks in our experiment is composed of an A/B treatment pair:[7]

Task 1 Participants are endowed with $10 and we use the Becker, DeGroot and Marschak (1964) mechanism to elicit their willingness-to-pay (WTP) for two lotteries for winning a $10 prize, one with a low (1/10) probability of winning, the other with a high (9/10) probability of winning.

Task 2 Participants are endowed with $10 and the Task-1 lotteries, and we elicit the willingness-to-accept (WTA) for the two lotteries.

Task 3 Participants are endowed with $10 in a "sooner" period and $1 in a "later" period one week later, where they can redistribute up to $9 from sooner to later, earning 20 percent interest on any delayed amount. The task pair switches the sooner date, either that day (today) or the next day (tomorrow).

Task 4 Participants are endowed with $20 and are asked to decide how much to donate to a local food bank. The treatment pair varies whether the donation is or is not matched dollar-for-dollar.

3 Laboratory Results

Our results focus on four behavioral comparative statics that can be examined with the four decision pairs described above. For each comparative static we follow an identical analysis: first outlining the expected result from the literature, and whether our experimental finding in the pooled data replicates the result. Second, we assess the quantitative response to experimenter demand. Third, given the sensitivity, we explore whether differentially applied experimenter demand is large enough to affect inference on the com-

[6] This strong-demand language from dQHR dates back to Binmore, Shaked and Sutton (1985)'s instructions in an ultimatum game, and is intended to generate demand effects that exceed those possible with more subtle wording. See also Ellingsen, Östling and Wengström (2018).

[7] Tasks 1 and 2 appeared in an individually randomized order, but were always followed by Task 3 then Task 4. Within task, the order of the decision pair was randomized at an individual level.

parative static. Unless otherwise stated, reported p-values are derived from two-sample T tests against a null effect.[8]

3.1 Probability Weighting

Our first comparative static examines whether participants overweight low-probability events and underweight large ones. For the Task-1 decision pair we endow participants with \$10 and ask for reports of their willingness-to-pay for two separate lotteries with a chance of winning an additional \$10, with $p \in \left\{\text{low} = {}^{1}\!/_{10}, \text{high} = {}^{9}\!/_{10}\right\}$.

Evidence of probability weighting is commonly seen in valuations that exceed the expected value (EV) for low-probability-of-winning lotteries and fall short of the EV for high-probability-of-winning lotteries. That is, probability weighting is revealed in our lottery valuations by risk-seeking choices at the low probability of winning and risk-averse choices at the high probability (Kahneman and Tversky, 1979). While the literature often examines more-structured models of probability weighting (Prelec, 1998), we focus on the prediction that the inferred risk attitude shifts from risk-seeking to risk-averse as we move from the low- to high-probability lotteries (see e.g., Harbaugh, Krause and Vesterlund, 2010).

Literature comparative static: We expect risk-seeking choices for the low-probability lottery (WTP in excess of the \$1 EV) and risk-averse choices for the high-probability lottery (WTP beneath the \$9 EV), anticipating rejection of the nulls in favor of the alternatives the Probability-weighting hypothesis is that:[9]

$$H_A : \text{Excess-value}_{\text{low}} = \text{WTP}_{\text{low}} - \text{EV}_{\text{low}} > 0, \tag{1}$$

$$H_A : \text{Excess-value}_{\text{high}} = \text{WTP}_{\text{high}} - \text{EV}_{\text{high}} < 0. \tag{2}$$

While the alternative is directional and multivariate—where the proper null would be a failure of either of these two conditions—we use the more-expansive two-sided hypothesis:

$$\text{No-probability-weighting} = \text{Excess-value}_{\text{low}} - \text{Excess-value}_{\text{high}} = 0. \tag{3}$$

This null looks for similar excess-valuations across the two lotteries, and as it is two-sided,

[8] Because our demand comparative statics are identified using between-subject treatments, we maintain consistency across all inferential tests by not using within-subject identification across task pairs.

[9] Our findings are robust to using WTA to perform the same assessment.

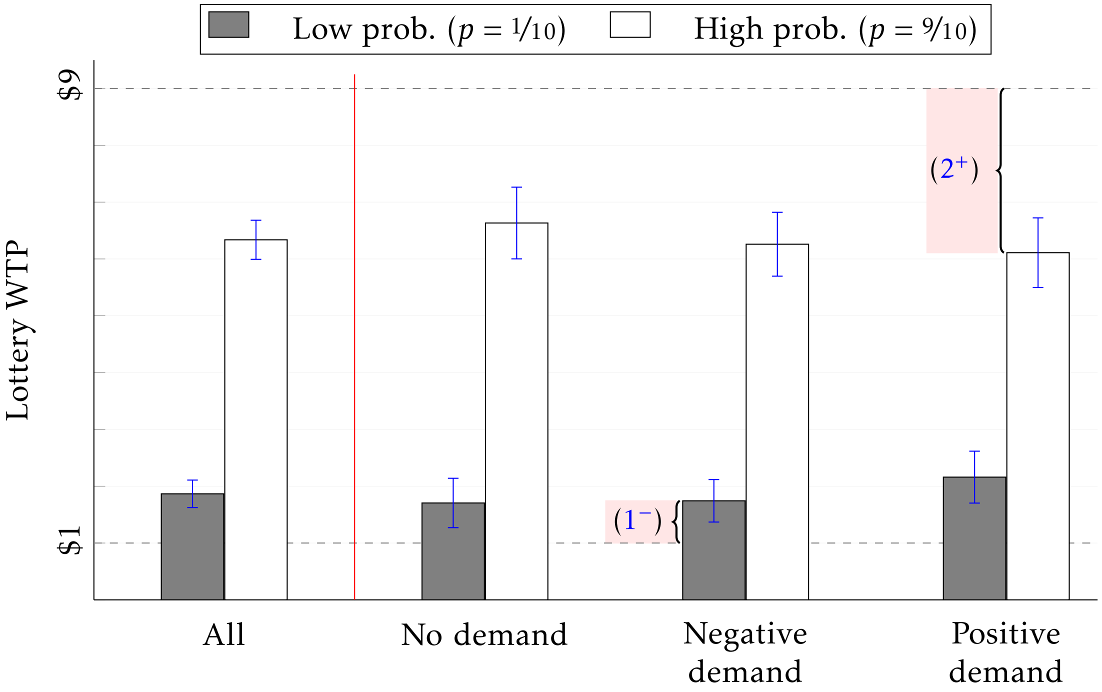


Figure 1: Over- and under-weighting of probabilistic events

Note: Average WTP for a lottery with a Low (p = 1⁄10) or High (p = 9⁄10) chance of winning \$10, both pooled and separated by demand treatment. Solid blue lines represent 95 percent confidence intervals. Dashed lines demarcate $\text{EV}_p = p \cdot \$10$, where Figure A.1 in the Online Appendix illustrates the Excess-values relative to this level.

reduces the rejection region for the alternative.[10]

Figure 1 shows the average WTP for low-probability (gray bars) and high-probability (white bars) lotteries, where the dashed gray lines at \$1 and \$9 indicate the respective EVs.[11] In the farthest left bars, we provide the pooled average WTP by lottery probability across all three treatments, while the three sets of bars on the right present the averages by treatment. Starting from the pooled results in All we see a full replication of the literature finding: the low-probability lottery is valued significantly above the \$1 EV (indicating risk seeking), while the high-probability lottery is valued significantly below the \$9 EV (indicating risk aversion). While the average valuation for the low-probability lottery is \$1.87 (significantly different from \$1 with $p < 0.001$), the high-probability lottery is \$6.33 (significantly different from \$9 with $p < 0.001$). This result from the pooled data is mirrored in each of the three treatments: where the largest p-value across the six

[10] As all of our experimental results will satisfy the directional part of the probability weighting hypothesis, this simpler null make it easier for demand to generate a false-negative

[11] Figure A.1 in the Online Appendix illustrates the Excess-valuations and the null hypothesis given in (3).

possible comparisons (lottery probability×treatment) is $p = 0.002$.[12] Joint tests of no difference between the WTP of the lotteries and the EV are rejected with high confidence ($p < 0.001$ in all comparisons) in favor of the behavioral comparative static from probability weighting in all three treatments.[13]

Tests for false negative & comparative-static reversal: A key concern for experimental inference here is whether we can eliminate the risk reversal by differentially inducing strong experimenter demand across the decision pair. To that end we examine whether demand can eliminate evidence of probability weighting (or even reverse it). We therefore examine WTP for the low-probability lottery under negative demand (asking participants to decrease their reported valuations) and for the high-probability lottery under positive demand (asking participants to increase reported valuations). The two shaded areas in Figure 1 assess this extreme distortion:

$$\text{Excess-value}^{-}_{\text{low}} = \text{WTP}^{-}_{\text{low}} - \text{EV}_{\text{low}}, \tag{1^-}$$

$$\text{Excess-value}^{+}_{\text{high}} = \text{WTP}^{+}_{\text{high}} - \text{EV}_{\text{high}}, \tag{2^+}$$

where the greatest chance of finding a null for the change in risk attitude is over the differentially distorted null that

$$H_0 : \text{No-probability-weighting}^{\ominus} = \text{Excess-value}^{-}_{\text{low}} - \text{Excess-value}^{+}_{\text{high}} = 0.$$

Looking across the demand treatments, when attempting to reduce WTP with negative demand we find $\text{Excess-value}^{-}_{\text{low}}$ =\$0.74 , and so still find risk-seeking choices for (1^-) with $p < 0.001$, while with positive demand the difference between the high-probability lottery WTP and its EV is Excess-value $^{+}_{\text{high}}$=-\$2.89 indicating risk-averse choices ($p < 0.001$). Combining the two for the comparative static $\text{No-probability-weighting}^{\ominus}$ hypothesis, we see that even with differentially applied experimenter demand, we reject the null hypothesis in favor of the literature finding, that the inferred risk attitude moves from risk-seeking to risk-aversion across the two lotteries ($p < 0.001$). The classic evidence of probability weighting is not even attenuated, much less reversed, under strong and

[12] The largest p-value is found in the low-probability lottery with no demand.

[13] Joint test p-values are from the harder-to-reject null of same difference,

$$H_0 : \text{WTP}_{\text{low}} - \text{EV}_{\text{low}} = \text{WTP}_{\text{high}} - \text{EV}_{\text{high}},$$

where the easier-to-reject null that *both* differences are zero (risk neutrality) leads to qualitatively similar results.

opposing experimenter demand.

3.2 Endowment Effect

For our second comparative static we assess the endowment effect, that the minimum price an agent willing-to-accept to sell an item (WTA) exceeds the maximum price they are willing-to-pay to buy the same item (WTP). While studies often examine the endowment effect over physical items such as mugs or pens (Knetsch, 1989), we instead follow the literature that assesses it over lotteries (see e.g., Sprenger, 2015). That is, we use the Task-1 and Task-2 assessments to determine whether, as in previous studies, participants' WTA exceed WTP for a given lottery (Knetsch and Sinden, 1984; Harbaugh, Krause and Vesterlund, 2010; Sprenger, 2015). While the endowment effect is a general phenomenon the literature suggests differences in power across our two probabilities. Using Sprenger (2015) to formulate power estimates, our lab study is (within each treatment) well-powered for uncovering the comparative static for the low-probability lottery ($1/10$), but has lower power for the high-probability lottery ($9/10$).

Literature comparative static: For each of the two probabilities of winning p, we expect to reject the null in favor of the alternative hypothesis below:

$$H_A : \text{Endow-effect}_p = \text{WTA}_p - \text{WTP}_p > 0 \tag{4}$$

As before, we illustrate raw averages across treatments, where Figure 2 shows the average WTA (white bars) and WTP (gray bars) for each lottery. We find evidence of the endowment effect using the pooled data (first four bars). Pooled across all three demand treatments participants require more to sell their lotteries ($3.15 and $6.95 for the low- and high-probability lotteries, respectively) than they are willing to pay to acquire the exact same lotteries ($1.86 and $6.33, respectively). These differences are significant both individually (low: $p < 0.001$; high: $p = 0.019$) and jointly ($p < 0.001$).[14]

Mirroring the probability-weighting results, experimenter demand does not significantly move the average responses by participants in our laboratory sample. Comparing average WTA to average WTP by demand treatment for the well-powered low-

[14] While our literature calculations for the high-probability lottery suggested a moderately powered hypothesis (~90% power) that we might try to turn into a null via demand, resampling from the pooled All-task data instead suggests much lower power in our actual implementation (~30% power). While this would be a poor setup for understanding the endowment effect, this still offers an important opportunity for testing demand effects.

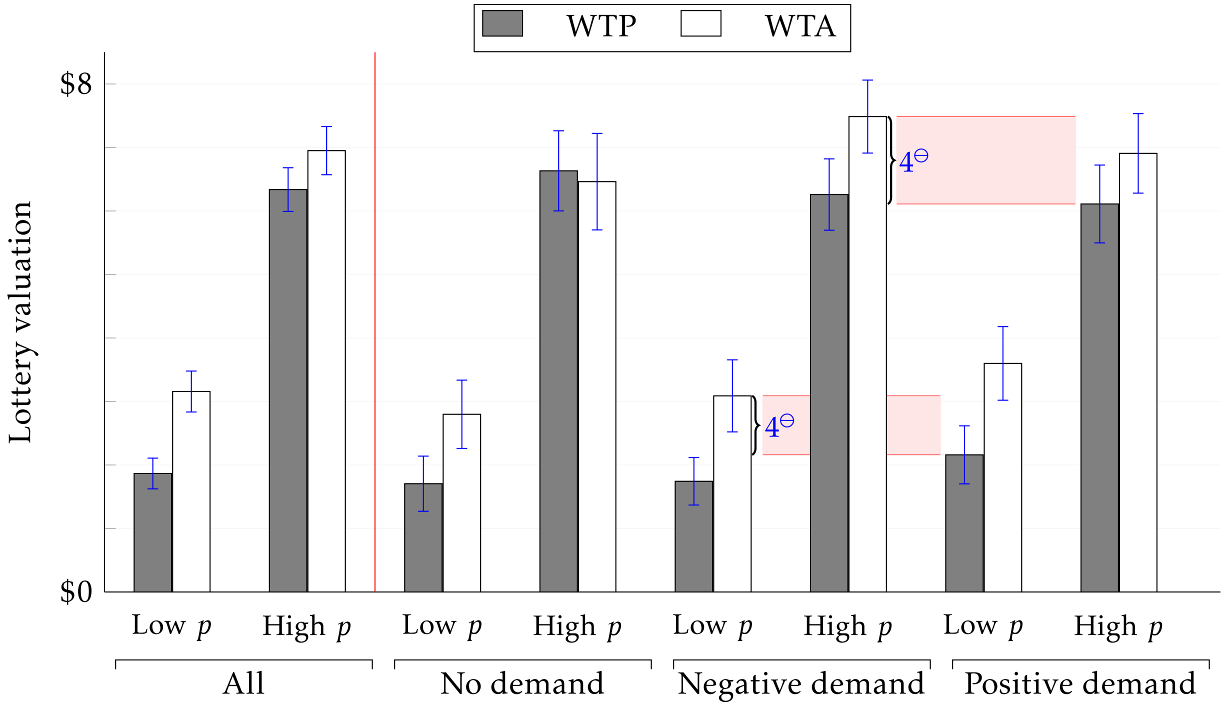


Figure 2: Endowment effect

Note: Average WTP and WTA for lotteries with a Low-p ($1/10$) or High-p ($9/10$) chance of winning \$10, both pooled and separated by demand treatment. Blue bars represent 95 percent confidence intervals.

probability lottery we find significant differences in all three cases ($p < 0.002$). However, for the under-powered high-probability lottery we find across treatment a significant difference with negative-demand ($p = 0.001$), a marginal difference with positive-demand ($p = 0.071$), and an insignificant difference with no-demand ($p = 0.731$).[15] Our experimental data in each separate demand treatment therefore mirrors the power calculations: strongly significant evidence of the endowment effect for the low-probability lottery, and more variable results for the high-probability lottery. Pooling over the low- and high-probability lottery for a test of (4) we find significant evidence of the endowment effect in each treatment .

Tests for false negative & comparative-static reversal: To examine if we can eliminate evidence of an endowment effect, or reverse the comparative static, we compare WTA decisions under negative demand to WTP decisions under positive demand. Specifically, for both lottery types, we explore the following relationships (the shaded comparisons in Figure 2):

[15] Looking at the joint-null of no difference in both lotteries, we can reject for each separate demand treatment with a maximal p-value of 0.037.

$$\text{Endow-effect}_p^{\ominus} = \text{WTA}_p^{-} - \text{WTP}_p^{+} \tag{$4^{\ominus}$}$$

We do not find that differential demand creates false negatives. While the request to undervalue a lottery being sold and overvalue the same lottery being bought attenuates the gap between WTA and WTP for the low-probability lottery ($\text{Endow-effect}_{\text{low}}^{\ominus} = \$0.93$) the difference is still highly significant ($p = 0.012$). Looking at whether differential demand can affect an under-powered study, the valuation gap for the high-probability lottery, which could plausibly have moved in the opposite direction given the reduced power, actually increases (where $\text{Endow-effect}_{\text{high}}^{\ominus} = \$1.38$, $p = 0.001$). This movement is in the *opposite direction* from the induced demand, though it is not separable from what we might expect due to sampling variation with lower power.[16] Our inability to generate a false negative for a joint test over both lotteries ($p < 0.001$ when assessing the low- and high-probability lotteries together) leads us to conclude that, even with strong differential experimenter demand, we cannot remove nor reverse the endowment effect.

3.3 Present Bias

Our third comparative static examines a behavioral feature of intertemporal decision making: present bias. Participants are asked to transfer up to \$9 from a sooner payment date (*Immediate*, or with a *Delay* of one day) to a later payment date seven days after the sooner date, where any amount moved to the later date earns 20 percent interest.

Neoclassical models of exponential discounting predict that a constant temporal distance between payment dates (a week's delay) would lead to the same amount transferred when the sooner date is today versus tomorrow. However, a large behavioral literature has examined impatience and present bias in which decision makers discount immediate benefits less than those with small delays (Laibson, 1997; O'Donoghue and Rabin, 1999). As such, participants with present-biased preferences are predicted to be less impatient when the sooner date is tomorrow (since both are delayed) rather than when the sooner date is today (only the later date is delayed).

While present bias is confirmed when examining work allocations over time (Augen-

[16] Another interesting assessment is whether differential demand can increase the significance for the under-powered test. However, attempting to maximize the comparative static $\text{Endow-effect}_{\text{high}}^{\oplus} = \text{WTA}_{\text{high}}^{+} - \text{WTP}_{\text{high}}^{-}$ we do not increase the significance, with a p-value of just 0.127.

blick, Niederle and Sprenger, 2015), the behavioral hypothesis is not confirmed when allocating money over time (Andreoni and Sprenger, 2012). To explore the potential for experimenter demand to generate a false positive we implement the assessment over monetary payments using the Andreoni and Sprenger (2012) implementation of a convex time budget set.[17] In selecting their methodology we expect to not find evidence to reject the null.

Literature comparative static: The present-biased comparative static predicted by the literature over monetary allocations is therefore the null:

$$H_0 : \text{No-Present-Bias} = \text{Transfer}_{\text{delay}} - \text{Transfer}_{\text{immediate}} = 0. \tag{5}$$

Our design choices here allow us to examine whether differential demand can generate false positives in favor of the directional present-biased hypothesis that

$$H_A : \text{Present-Bias} = \text{Transfer}_{\text{delay}} - \text{Transfer}_{\text{immediate}} > 0. \tag{6}$$

We illustrate the results in Figure 3, where we show the average amount transferred to the later payment date when the sooner date is either immediate (gray bars) or delayed (white bars). The pooled data ($N = 236$) on the left of the figure shows that on average \$7.88 is transferred with an immediate sooner payment date, versus \$8.05 when there is a delay. Although the results move in the direction of present bias, the difference is small and insignificant in the pooled sample ($p = 0.339$). So, despite the increase in power over Andreoni and Sprenger (2012) in the pooled sample, our results replicate the original finding.

Looking separately at each demand treatment on the right of Figure 3 we see the same pattern in each treatment: slight evidence of present bias, but with no significant treatment effect. The largest difference is in the no-demand treatment, where participants transfer \$0.36 more to the later date when the sooner payment is delayed ($p = 0.239$), but the smaller differences under negative and positive demand are much further from significance ($p = 0.888$ and $p = 0.733$, respectively).

Tests for a false positive & comparative static reversal: We now examine evidence for present bias under extreme demand by asking whether we can create a false positive

[17] The authors summarize their finding on present bias as a null effect (from their conclusion "*[a]dditionally, we find no evidence of present bias.*"), where they attribute this to clearer methodological control when using delay over monetary payments.

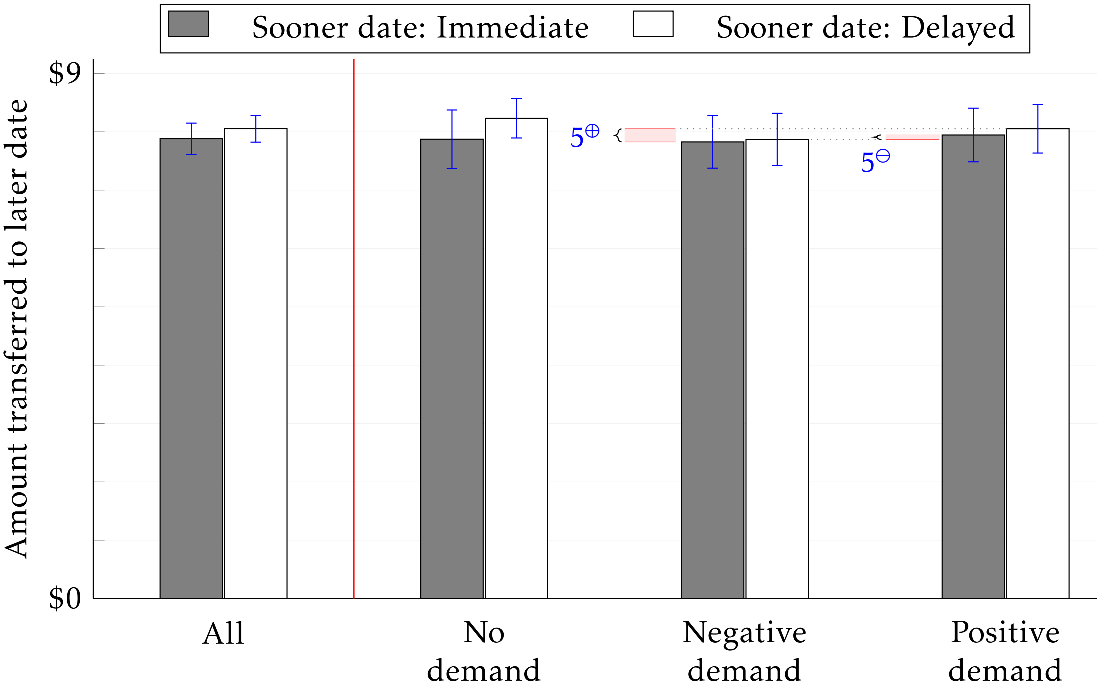


Figure 3: Present bias

Note: Average amount transferred to the later date when the sooner date is Immediate or has a Delay, pooled and separated by demand treatment. Blue lines represent 95 percent confidence intervals.

when differentially applying demands for the immediate and delayed conditions. Because the literature led us to expect a null result, we examine deviations from the null in both directions, where the first would indicate present bias:

$$\text{Present bias}^{\oplus} = \text{Transfer}^{+}_{\text{Delay}} - \text{Transfer}^{-}_{\text{Immediate}} \tag{$5^{\oplus}$}$$

$$\text{Present bias}^{\ominus} = \text{Transfer}^{-}_{\text{Delay}} - \text{Transfer}^{+}_{\text{Immediate}} \tag{$5^{\ominus}$}$$

Inspecting these comparisons in Figure 3 makes clear that we cannot generate substantive effects by selectively applying demand between the immediate or delayed sooner-payment conditions. Directionally, we can widen the present-bias gap of $0.17 found in the pooled result to $0.23 in the ($5^{\oplus}$) comparison, though the difference is still far from being significant (p = 0.465). Similarly, we can change the direction of the effect to a -$0.07 gap (a small delay bias) by reversing the demand conditions per equation ($5^{\ominus}$). However, the reversed direction is again insignificant (p = 0.819). Despite considering the knife-edge case of a true null, our extreme experimenter-demand manipulations cannot alter the qualitative inferences for this intertemporal question, fully replicating the

Andreoni and Sprenger (2012) result.

3.4 Tradeoffs between Self and Others

Our final comparative static examines the tradeoff between money to self and money to others. Specifically, we examine the comparative static of how charitable giving responds to a decrease in the price-of-giving. Participants are endowed with \$20 and asked to allocate the money between themselves and a donation to a local food bank, for a decision pair where the donation given is or is not matched one-for-one. That is, we vary the price-of-giving a dollar to the charity, $c \in \left\{\text{match} = 0.5, \text{no-match} = 1.0\right\}$.

The law of demand predicts that as the price-of-giving falls the donation-received by the charity increases (the total amount including the match). Indeed, this is consistent with empirical evidence (Andreoni and Miller, 2002; Eckel and Grossman, 2003; Huck and Rasul, 2011; Karlan and List, 2007) showing an inverse relationship between the price-of-giving and donation-received.

Literature comparative static: Across our decision pair, we expect that increasing the price-of-giving from low (one-for-one match) to high (no match) decreases the donation-received. That is, we expect to reject a null in favor of the following comparative static:

$$H_A : \text{Charity-receipt} = \text{Donation-received}_{\text{match}} - \text{Donation-received}_{\text{no-match}} > 0. \tag{7}$$

Figure 4 shows the average donation-received by the charity when the donation is unmatched (gray bars) and matched (white bars), for the pooled data (left) and then separated by demand treatment (right). The figure indicates a clear and significant comparative static in the pooled data. As expected a one-for-one match significantly increases average donations-received by the charity (from \$8.92 to \$17.54, $p < 0.001$), a directional response that is replicated in each of the three demand treatments (with a one-for-one-match increasing average donations-received by \$9.16, \$7.65, and \$9.65 for the negative-demand, no-demand, and positive-demand treatments, respectively, all $p < 0.001$).

Tests for a false negative & comparative static reversal: For differential demand, we compare donations-received by the charity under the one-for-one match when exposed to negative demand (pushing participants to reduce giving) to donations-received by the charity without a match when exposed to positive demand (to increase giving):

$$\text{Charity-receipt}^{\ominus} = \text{Donation-received}^{-}_{\text{match}} - \text{Donation-received}^{+}_{\text{no-match}} \tag{$7^{\ominus}$}$$

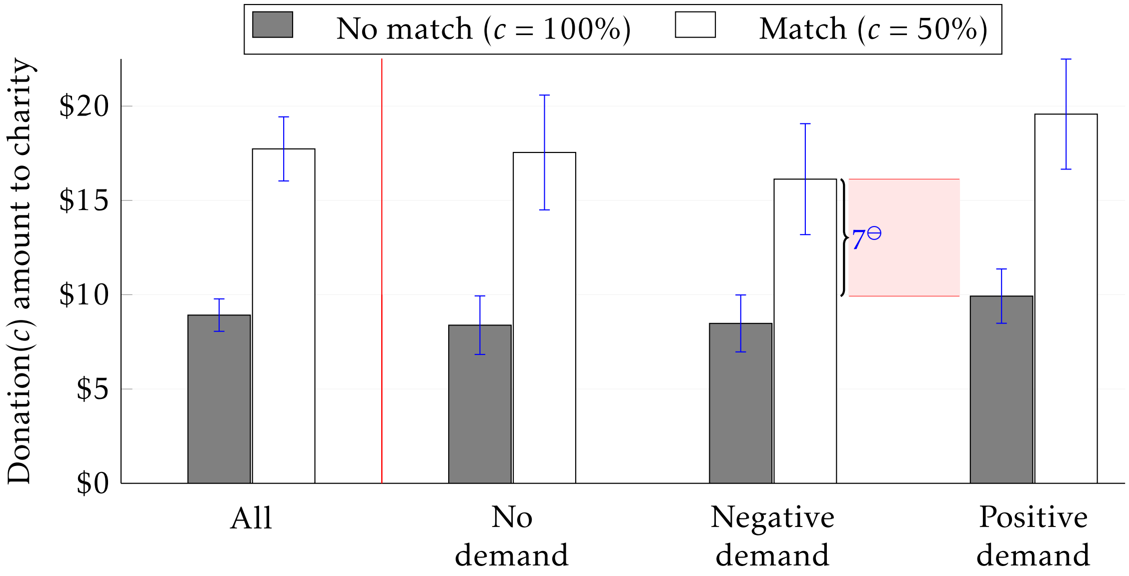


Figure 4: Donation-received response to match

Note: Average donation-received by the charity when matched (low price-of-giving $c = 50\%$) or unmatched (high price-of-giving $c = 100\%$), both pooled and separated by demand treatment. Solid blue lines represent 95 percent confidence intervals.

The assessment is whether experimenter demand can create a false negative on the donation-received by attenuating the response to the decreased price-of-giving, where the differential demand effect in ($7^{\ominus}$) is illustrated in Figure 4 as the shaded area. While differentially applying experimenter demand reduces the assessed effect to \$6.20, the result is still in the expected direction and highly significant ($p < 0.001$). Thus, extreme experimenter demand does not give rise to false negatives for the donation-received by the charity.[18]

[18] While not part of our initial hypothesis we can also explore the treatment effect of price-of-giving on: Donation-given $= c \cdot$ Donation-received. Counteracting income and substitution effects means theory cannot provide a directional hypothesis, where results from the literature are also mixed showing either no or a small positive effect (Huck and Rasul, 2011; Karlan and List, 2007; Karlan, List and Shafir, 2011). Our data indicates a null-effect (in both the pooled data, $p = 0.931$, and the three separate demand treatments, smallest p-value of 0.696), so we can further explore the potential for a false positive around another knife-edge null. Differentially applying experimenter demand, we fail to reject the null ($p = 0.217$) when trying to maximize the response in

$$\text{Donor-response}^{\oplus} = \text{Donation-given}^{+}_{\text{match}} - \text{Donation-given}^{-}_{\text{no-match}} = \$1.31,$$

but we marginally reject the null ($p = 0.074$) when trying to minimize the treatment effect

$$\text{Donor-response}^{\ominus} = \text{Donation-given}^{-}_{\text{match}} - \text{Donation-given}^{+}_{\text{no-match}} = -\$1.86.$$

.

4 Demand Effects Across Subject Populations

In the previous section, we found that strong experimenter demand gives rise to very small quantitative responses. The estimates for each of the eight experimental decisions are reported in Figure 5(a), following the dQHR methodology. Each bar is the z-scored difference between the average decision in the positive- or negative-demand treatments and with the no-demand treatment as the control baseline. We also estimate the pooled effects across the eight decisions in the first "All tasks" bars.

For *All tasks* we show that the pooled impact of the positive-demand treatments (0.10σ) is only marginally significant ($p = 0.077$), while the pooled impact of the negative-demand treatments is not significantly different from baseline. Looking across decisions we cannot reject that the direction of the response is independent of the demand treatment (one-sided Fisher exact test p=0.304). Similar conclusions follow for each decision—where the only significant quantitative effect we find in the predicted direction is in the low-p WTA elicitation ($p = 0.045$).[19] The limited lab response to experimenter demand caused us to not find false negatives or false positives.

Next we explore whether the results documented in the lab extends to two common online subject pools. In particular, we replicate our design using 756 participants recruited on Amazon Mechanical Turk and 732 participants recruited on Prolific.[20] Mirroring common laboratory and online studies we examine larger samples online (balancing total subject payments across samples). This broadens our understanding of how susceptible qualitative inference is to experimenter demand across studies. The quantitative results from the MTurk and Prolific replications are reported in Figure 5(b) and (c).[21]

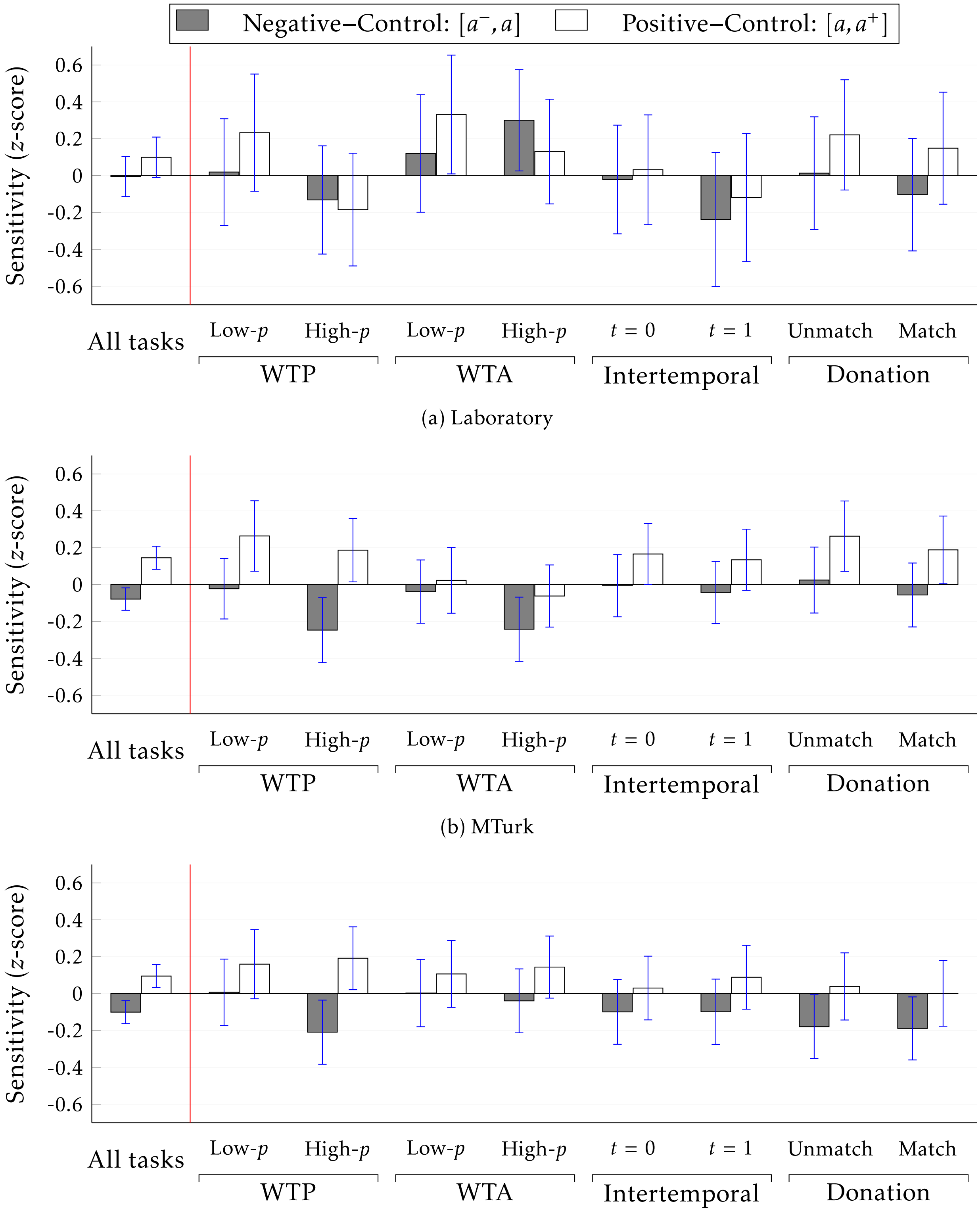
Like the lab sample, the quantitative impacts of experimenter demand are small in our online samples, but the larger sample-size means these effects are estimated with greater precision.[22] Unlike the lab, the positive- and negative-demand treatments on MTurk

[19] We find a significant demand effect on the high-p WTA lottery, however, in the *opposite* direction of the induced demand ($p = 0.033$).

[20] Our MTurk and Prolific replications differ from the lab-population study as follows: (i) We lower the incentives by one-fifth (a lump-sum payment of $2 and task incentives between $1-2) to create ecologically valid stake sizes. (ii) Randomization of the demand treatment occurs at the individual level.

[21] We include full qualitative results from our MTurk and Prolific replications in Online Appendix A. On Prolific, we do not find evidence of probability-weighting for the low-probability lottery in the no-demand treatment ($p = 0.102$ against risk-neutral pricing); however, the comparative static test across the Excess-values between the two lotteries is statistically significant, as is a joint test.

[22] The response to experimenter demand seen in our MTurk sample is smaller than that seen in dQHR. This difference in the quantitative response may result from changes in the subject pool between 2016/2017 (dQHR) and the end of 2020 (our study), from differences in design, or from the quality-controls applied



(a) Laboratory

(b) MTurk

(c) Prolific

Figure 5: Sensitivity Analysis

(Prolific) both generate statistically significant pooled effects of 0.15σ (0.10σ) and -0.08σ (-0.10σ) in the induced directions.[23] Further, we can reject that the directional response is independent of demand treatment (one-sided Fisher's exact $p = 0.020$ for MTurk and $p = 0.003$ for Prolific).

Similar to our laboratory sample, the small quantitative demand effects in our online samples leave qualitative inference on directional hypothesis unaffected. There is no evidence of false negatives or comparative static reversals from differential demand. In contrast to the lab, in the knife-edge case of a precise null (present bias), we find evidence that experimenter demand can create false positives in both online populations.[24] That is the larger sample affordable online has a perverse interaction with demand, creating significant false positives despite the quantitatively small shifts.

Finally, it is important to reiterate that we examine an extreme case for experimental demand to influence inference, where strong experimenter demand is differentially induced. Therefore, we expect the magnitude of demand effects found in our study to be much larger than what one might expect in a standard experiment.

5 Conclusion

Our study tests whether experimenter demand can distort key inferences drawn from experiments. We use de Quidt, Haushofer and Roth (2018) to bound the quantitative impact of strong experimenter demand on decisions made within four classic behavioral phenomena. We then use these bounds to explore whether the most extreme instances of experimenter demand can reverse comparative statics, threatening the qualitative inference in experimental studies.

Using a laboratory population with college students, we find surprisingly little response to experimenter demand. The quantitative effects are small, insufficient for reverting any qualitative inference. Our laboratory results show no signs of false negatives and no more than marginal evidence of false positives when testing the knife-edge case of a null hypothesis. The response to experimenter demand for online populations (MTurk and Prolific) remains small, and we find no evidence that experimenter demand can

in our study not being available in 2016/2017.

[23] The corresponding p-values are $p < 0.001$ ($p = 0.003$) and $p = 0.011$ ($p = 0.002$), respectively.

[24] See Table A.4 for inferential tests attempting to minimize and maximize each comparative static via demand. A reason for the false positives here is that the online populations both show a higher demand sensitivity, and our budget-matched samples are larger for the online samples.

reverse a directional hypothesis. Although upholding the clear evidence against false-negatives, the online samples do demonstrate the sensitivity to experimenter demand when testing a precise null. With extreme experimenter demand, strong and differentially applied across decisions, it is possible to generate false positives in the online samples.

Although most experimental designs eliminate or mitigate the impact of experimenter demand (de Quidt, Vesterlund and Wilson, 2019) our results demonstrate limited effect on inference of deliberate and extreme experimenter demand. Requesting that participants select a high or low action causes only slight movement in the decision estimates, a movement that gives rise to small changes in the treatment effect, and in turn is insufficient for reverting a directional inference. Our laboratory and online samples show no evidence that a hypothesized ill-intentioned experimenter will succeed in generating a false negative result, however differentially moving demand around a precise null can result in false positives in larger online samples.

References

Andreoni, James, and Charles Sprenger. 2012. "Estimating time preferences from convex budgets." *American Economic Review*, 102(7): 3333–56.

Andreoni, James, and John Miller. 2002. "Giving according to GARP: An experimental test of the consistency of preferences for altruism." *Econometrica*, 70(2): 737–753.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. "Working over time: Dynamic inconsistency in real effort tasks." *Quarterly Journal of Economics*, 130(3): 1067–1115.

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak. 1964. "Measuring utility by a single-response sequential method." *Behavioral Science*, 9(3): 226–232.

Binmore, Ken, Avner Shaked, and John Sutton. 1985. "Testing noncooperative bargaining theory: A preliminary study." *American Economic Review*, 75(5): 1178–1180.

Bischoff, Ivo, and Björn Frank. 2011. "Good news for experimenters: Subjects are hard to influence by instructors' cues." *Economics Bulletin*, 31(4): 3221–3225.

Danz, David, Neeraja Gupta, Marissa Lepper, Lise Vesterlund, and K. Pun Winichakul. 2021. "Going virtual: A step-by-step guide to taking the in-person experimental lab online." Available at SSRN: http://dx.doi.org/10.2139/ssrn.3931028.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and bouding experimenter demand." *American Economic Review*, 108(11): 3266–3302.

de Quidt, Jonathan, Lise Vesterlund, and Alistair J Wilson. 2019. "Experimenter demand effects." In *Handbook of research methods and applications in experimental economics.* , ed. Arthur Schram and Aljǎz Ule, 384–400. Edward Elgar Publishing.

Eckel, Catherine C, and Philip J Grossman. 2003. "Rebate versus matching: does how we subsidize charitable contributions matter?" *Journal of Public Economics*, 87(3-4): 681–701.

Ellingsen, Tore, Robert Östling, and Erik Wengström. 2018. "How does communication affect beliefs in one-shot games with complete information?" *Games & Economic Behavior*, 107: 153–181.

Harbaugh, William T, Kate Krause, and Lise Vesterlund. 2010. "The fourfold pattern of risk attitudes in choice and pricing tasks." *Economic Journal*, 120(545): 595–611.

Huck, Steffen, and Imran Rasul. 2011. "Matched fundraising: Evidence from a natural field experiment." *Journal of Public Economics*, 95: 351–362.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263–292.

Karlan, Dean, and John A. List. 2007. "Does price matter in charitable giving? Evidence from a large-scale natural field experiment." *American Economic Review*, 97(5): 1774—93.

Karlan, Dean, John A List, and Eldar Shafir. 2011. "Small matches and charitable giving: Evidence from a natural field experiment." *Journal of Public Economics*, 95(5-6): 344–350.

Kessler, Judd, and Lise Vesterlund. 2015. "The external validity of laboratory experiments: The misleading emphasis on quantitative effects." In *Handbook of Experimental Economic Methodology.* , ed. Guillaume R Frechette and Andrew Schotter, 392–405. Oxford University Press.

Knetsch, Jack L. 1989. "The endowment effect and evidence of nonreversible indifference curves." *American Economic Review*, 79(5): 1277–1284.

Knetsch, Jack L, and John A Sinden. 1984. "Willingness to pay and compensation demanded: Experimental evidence of an unexpected disparity in measures of value." *The Quarterly Journal of Economics*, 99(3): 507–521.

Laibson, David. 1997. "Golden eggs and hyperbolic discounting." *Quarterly Journal of Economics*, 112(2): 443–478.

O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing it now or later." *American Economic Review*, 89(1): 103–124.

Orne, M. T. 1962. "On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications." *American Psychologist*, 17: 776–783.

Prelec, Drazen. 1998. "The probability weighting function." *Econometrica*, 497–527.

Sprenger, Charles. 2015. "An endowment effect for risk: Experimental tests of stochastic reference points." *Journal of Political Economy*, 123(6): 1456–1499.

Tsutsui, Kei, and Daniel John Zizzo. 2014. "Group status, minorities and trust." *Experimental Economics*, 17: 215–244.