








# Effect size, Experimenter Demand and Inference

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

February, 2025

## 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 to test and understand 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, which in turn compromises 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.<sup>1</sup> 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.<sup>2</sup> 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

---

<sup>1</sup>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.

<sup>2</sup>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 environment  $B$  with a control environment  $A$ . Directional effects are assessed over the decision pair through the difference  $\Delta\mu = \mu_B - \mu_A$ , comparing the average choice in the two conditions. 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 ( $\mu^0$ ), with an induced positive demand ( $\mu^+$ ) and an induced negative demand ( $\mu^-$ ). 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 control  $A$  and treatment  $B$  we attempt to minimize the treatment effect: demanding a decrease in the average choices in  $B$ , while demanding an increase in  $A$ . That is, we examine the demand-minimized treatment-effect  $\Delta\mu^\ominus = \mu_B^- - \mu_A^+$ . In an environment where the literature would predict a positive effect ( $H_A : \Delta\mu > 0$ ), 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  $H_0 : \Delta\mu = 0$ . 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  $\Delta\mu^\oplus = \mu_B^+ - \mu_A^-$ . Similarly, we can assess impact in the opposite direction by evaluating the demand-minimized treatment-effect  $\Delta\mu^\ominus = \mu_B^- - \mu_A^+$ .

In what follows, we examine the potential for experimenter demand to reverse infer-

ence 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.<sup>3</sup> 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.<sup>4</sup> 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.<sup>5</sup>

## 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*

---

<sup>3</sup>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.

<sup>4</sup>See Online Appendix B for detailed calculations.

<sup>5</sup>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.*<sup>6</sup> 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:<sup>7</sup>

**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-

---

<sup>6</sup>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 \(2018\)](#).

<sup>7</sup>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.<sup>8</sup>

### 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  $p \in \{\text{low} = 1/10, \text{high} = 9/10\}$ .

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:<sup>9</sup>

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

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

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:

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

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

---

<sup>8</sup>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.

<sup>9</sup>Our findings are robust to using WTA to perform the same assessment.

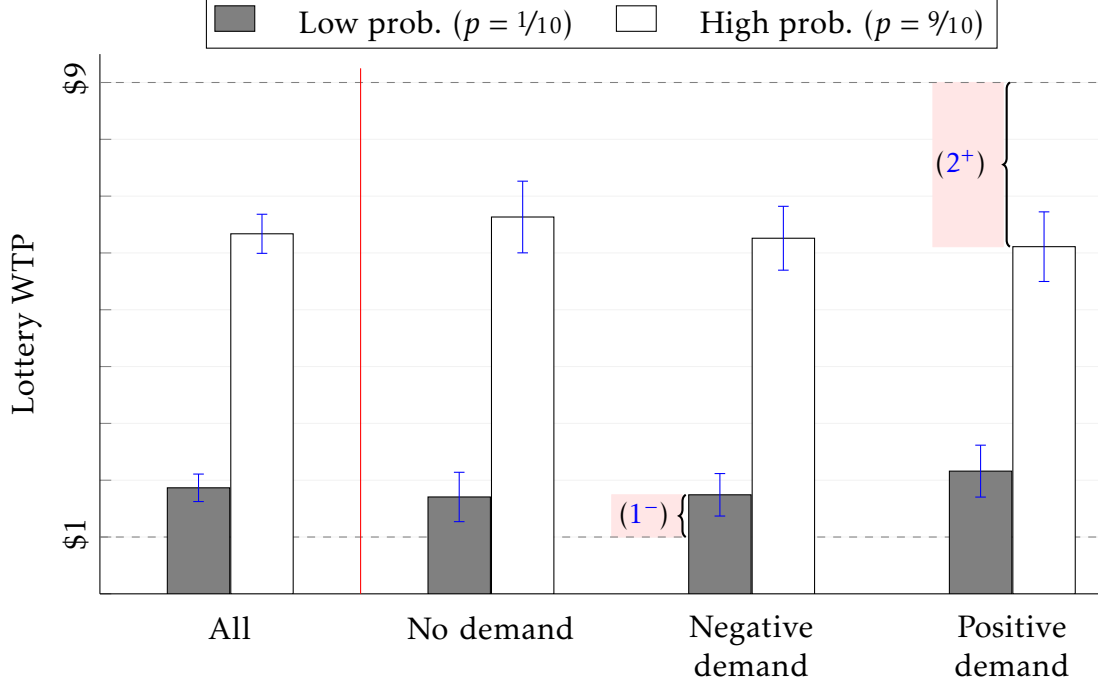


Figure 1: Over- and under-weighting of probabilistic events

*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  $EV_p = p \cdot \$10$ , where Figure A.1 in the Online Appendix illustrates the Excess-values relative to this level.

reduces the rejection region for the alternative.<sup>10</sup>

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.<sup>11</sup> 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

<sup>10</sup>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

<sup>11</sup>Figure A.1 in the Online Appendix illustrates the Excess-valuations and the null hypothesis given in (3).



possible comparisons (lottery probability  $\times$  treatment) is  $p = 0.002$ .<sup>12</sup> 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.<sup>13</sup>

*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:

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

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

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  $\text{Excess-value}_{\text{high}}^+ = -\$2.89$  indicating risk-averse choices ( $p < 0.001$ ). Combining the two for the comparative static No-probability-weighting<sup>⊖</sup> 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

<sup>12</sup>The largest  $p$ -value is found in the low-probability lottery with no demand.

<sup>13</sup>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:

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

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 ( $p < 0.001$ ).<sup>14</sup>

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-

---

<sup>14</sup>While our literature calculations for the high-probability lottery suggested a moderately powered hypothesis ( $\sim 90\%$  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 ( $\sim 30\%$  power). While this would be a poor setup for understanding the endowment effect, this still offers an important opportunity for testing demand effects.

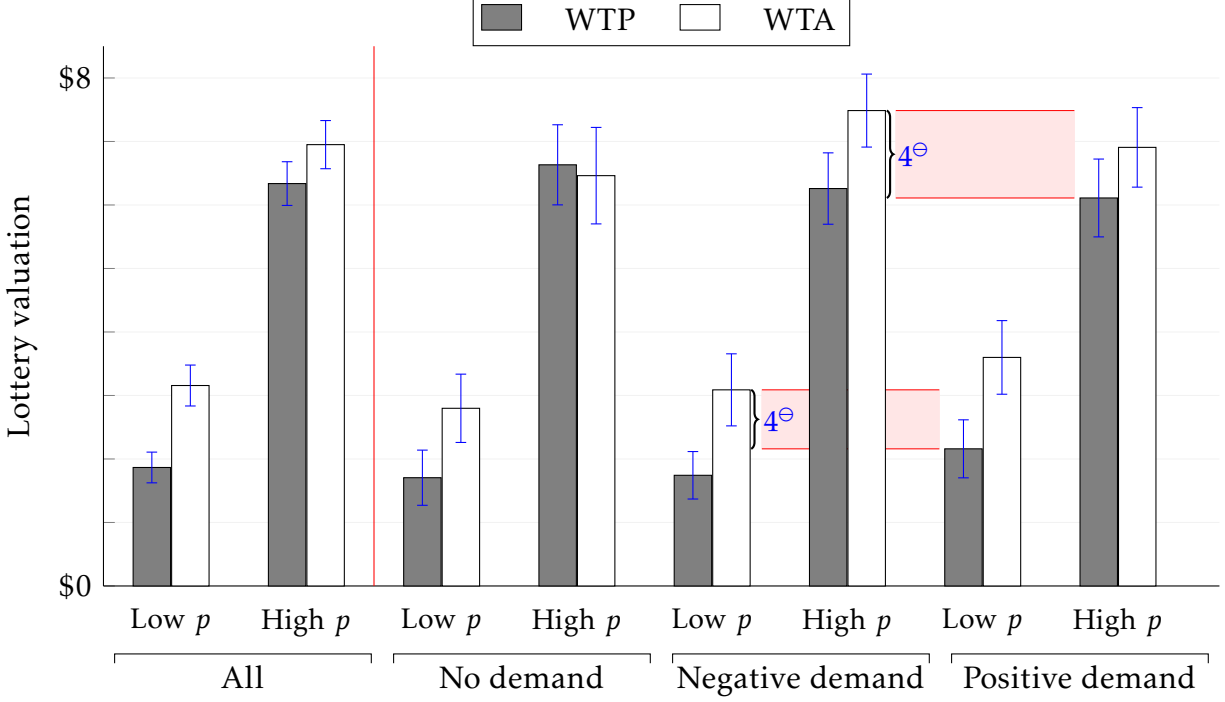


Figure 2: Endowment effect

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 ( $p = 0.731$ ).<sup>15</sup> 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):

<sup>15</sup>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.

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

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.<sup>16</sup> 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-

---

<sup>16</sup>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.<sup>17</sup> 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:

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

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

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

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

---

<sup>17</sup>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.

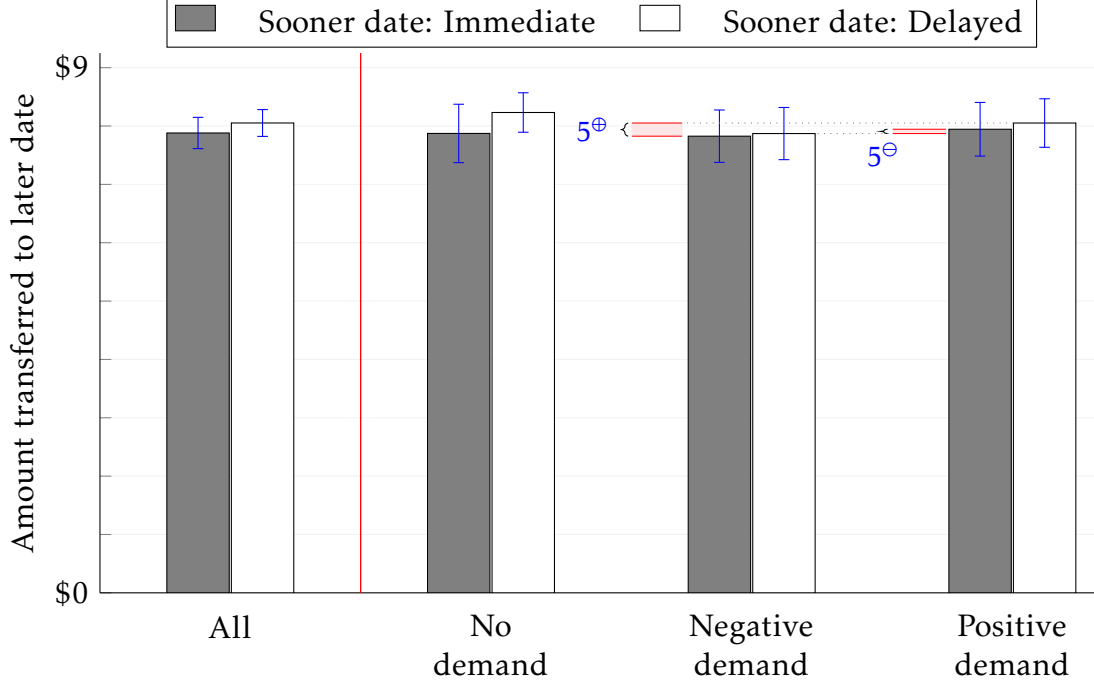


Figure 3: Present bias

*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:

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

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

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,  $c \in \{\text{match} = 0.5, \text{no-match} = 1.0\}$ .

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:

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

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):

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

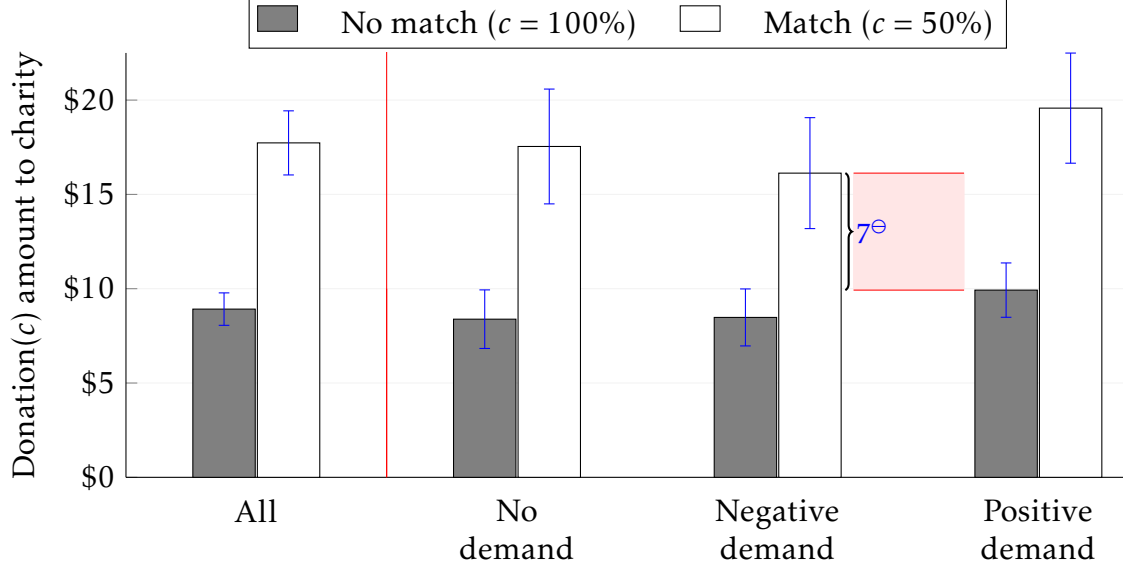


Figure 4: Donation-received response to match

*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  $(7^\Theta)$  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.

While not directly part of our initial hypotheses we can also explore the treatment effect of price-of-giving on Donation-given =  $c \cdot$  Donation-received.. That is, is there a difference in the amount given up by the participant:

$$H_0 : \text{Donor-response} = \text{Donation-given}_{\text{match}} - \text{Donation-given}_{\text{no-match}} = 0 \quad (8)$$

Theoretically, there is no clear reason to expect to reject the null here. Counteracting income and substitution effects mean that theory cannot provide a directional hypothesis without very strong assumptions on preferences. 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). Matching this, 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). Because of this, the treatment allows for an additional exploration of the potential for a false positive around an expected null.

Differentially applying experimenter demand, we fail to reject the null ( $p = 0.217$ ) when trying to maximize the response in

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

However, we do marginally reject the null ( $p = 0.074$ ) when trying to *minimize* the treatment effect here, with

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

So the marginal result here is in the opposite direction from the induced demand treatments. Taken together, the results for the donation given mirror the results for the present-bias null: strong differential demand is not capable of generating a false positive result in our lab sample.

## 4 Population, Significance and Effect size

Having established that strong and differential experimenter demand across conditions does not alter inference in our lab sample, we now ask whether this result also holds when we draw the samples from online populations commonly used in economic papers. In particular, we repeat our experimental design to assess the extent that demand can shift the qualitative conclusions but with samples drawn from Amazon Mechanical Turk (MTurk) and Prolific. Mirroring the results from the lab, we show that differential demand cannot generate false negatives for clear results in the literature. However, unlike the lab sample, we will show that demand *can* generate qualitative false positives in the online samples for comparative statics where the literature would predict a null.

While there is a clear distinction about demand and inference between the lab and online, we will demonstrate that the false positives in our online samples are not inherently driven by a greater response to explicit demand when online. Indeed, isolating the pure effect-size components of demand, we outline a useful positive result across populations: the lab and online samples are actually very similar to one another in terms of the normalized effect size, both in terms of the comparative static effects and the range in demand. Instead, the main force for the demand-driven false positives in our online sam-

ples are driven by the greater inferential precision made possible with the larger sample collected online.

Explicit demand only generates ‘small’ shifts in the assessed effects, for each hypothesis and in each population. However, while these quantitatively small effects do not reach the threshold for significance in our lab sample, in the larger online samples the same-sized demand effects become significant. As such, our findings help to better flag where we might plausibly have concerns over experimenter demand: small effect sizes in larger samples. As the behavioral literature often focuses on the significance of a qualitative effect rather than effect size, in the last section we document effect sizes and samples in economic experiments across six years of papers in the *American Economic Review*.

## 4.1 Inferential effects across Population

We reproducing our lab design with three between-subject treatments (No Demand, Negative and Positive) across eight decisions (four lottery pricing tasks, two convex time budget sets, and two charitable donation decisions). We repeat the experiment using participants recruited from MTurk (756 participants) and Prolific (732 participants) which we will compare to our lab sample (236 participants).<sup>18</sup> Importantly, and as we will show mirroring the pattern in the economic literature, we purposefully recruit larger samples for the two online populations, disciplining this approach across population by balancing the total expenditures across the three populations (see [Rigotti, Wilson and Gupta, 2023](#), for a similar examination across population focused on inferential power rather than demand)).

While we concentrate here on summarizing the qualitative inferential effects across the three populations, in the Online Appendix we provide analogous figures and sum-

---

<sup>18</sup>Our MTurk and Prolific replications differ from the lab-population study as follows: (i) We lower the offered incentives by one-fifth (a lump-sum payment of \$2 and task incentives between \$1-2) to create ecologically valid stake sizes for economic experiments on these platforms. (ii) Randomization of demand treatment occurs at the individual level (as opposed to the session level in the lab). (iii) We shift the charity recipient from a local Children’s hospital in Pittsburgh, to a national level foodbank. Other than accommodating these changes, all questions/framing are identical to the lab treatments. Moreover, the qualitative comparative statics analyzed are the same as the lab: probability weighting (both WTP and WTA), the endowment effect (both low and high probability lotteries), present bias, and charitable giving (both the charity receipt, and donor response).

Table 1: Inferential Results by Population

Comparative static	No demand			Differential demand					
	$\Delta\mu^0$			Deflated, $\Delta\mu^\ominus$			Inflated, $\Delta\mu^\oplus$		
	L	M	P	L	M	P	L	M	P
<i>Literature suggests positive, well-powered</i>									
Prob. weight (Eq.3, WTP)	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
Prob. weight (Eq.3, WTA)	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
Endow. Effect (Eq.4, Low)	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
Charity receipt (Eq.7)	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
<i>Literature suggests positive, under-powered in lab sample</i>									
Endow. Effect (Eq.4, High)	$\sim$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\sim$	$\sim$	$\sim$	$\Rightarrow$	$\Rightarrow$
<i>Literature suggests null</i>									
Present bias (Eq.5)	$\sim$	$\sim$	$\sim$	$\sim$	$\Leftarrow$	$\Leftarrow$	$\sim$	$\Rightarrow$	$\Rightarrow$
Donor response (Eq.8)	$\sim$	$\sim$	$\sim$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\sim$	$\Rightarrow$	$\Rightarrow$
Total observations, $N$	160	500	476	156	473	484	156	473	484

*Note:* Comparative static effects for each population are given in columns: (L)ab, (M)echTurk, (P)rolific. Directions of comparative statics are given by right arrows (in blue) for positive results and left arrows (in red) for negative results. Significance results are indicated by:

(i)  $\Rightarrow$  for 1 percent significance; (ii)  $\Rightarrow$  for 5 percent significance; (iii)  $\rightarrow$  for 10 percent significance; (iv)  $\sim$  for insignificance at 10 percent level.

maries for MTurk and Prolific to those outlined for the lab in Section 3.<sup>19,20</sup> The inferential effects are outlined in Table 1, indicating our seven comparative statics (the table rows), across the three different populations ( $L$  for Lab,  $M$  for MTurk and  $P$  for Prolific), with and without differential demand conditions across the treatment and control.

Each entry in Table 1 indicates the  $t$ -test significance for the treatment-effect null hypothesis  $H_0 : \beta_\Delta = 0$ , assessed via the regression:

$$y_i = \beta_0 + \beta_\Delta \cdot \mathbf{1}(\text{Treatment } B)_i + \epsilon_i. \quad (9)$$

<sup>19</sup>Full results matching our prior qualitative analysis within the MTurk and Prolific replications are included in Online Appendix A. The only real anomaly in the raw results is for Prolific, where we do not find evidence of risk-seeking for the low-probability lottery in the no-demand treatment ( $p = 0.102$  against risk-neutral pricing). However, the Probability-weighting comparative static between the two lotteries is statistically significant, as is the joint test.

<sup>20</sup>In a sensitivity analysis á la dQHR in each separate decision, we show that while demand effects are more consistent in direction for the online populations, the scale of the demand effects at the decision level are not significantly different from the lab.

While specified here as a regression over the decision variable  $y$ , the test is equivalent to a two-sample  $t$ -test between treatment and control, where the estimator  $\hat{\beta}_\Delta$  will be the difference in the means across conditions for the decision variable, so  $\hat{\beta}_\Delta = \Delta\mu = \mu_B - \mu_A$ . Inference is based on the test statistic:

$$T = \frac{\text{Treatment difference}}{\text{Standard error for treatment difference}} \simeq \sqrt{\frac{N}{2}} \cdot \frac{\Delta y}{\hat{\sigma}}. \quad (10)$$

where  $\hat{\sigma}$  is the regression's root-mean-squared error (the pooled standard deviation for  $y$  across conditions in the standard two-sample  $t$ -test), where the approximation on the right assumes a balanced design—with  $\frac{N}{2}$  observations for  $y$  in each of the two conditions,  $A$  and  $B$ , for a total of  $N$  observations.

Each [Table 1](#) entry indicate significance in the intuitive/literature direction for each hypothesis using right arrows (with one/two/three lines in the arrow indicating confidence in the conclusion at the 90/95/99 percent level over a two-sided alternative). In-significance is indicated with a tilde, while significance in the counter-intuitive direction is indicated using left arrows.

In the first three *No demand* columns of [Table 1](#), we indicate the inferential conclusion without explicit demand in either treatment or control (so inference over the demand-neutral comparative static  $\hat{\beta}_\Delta = \mu_B^0 - \mu_A^0$ ). The No demand columns indicate a clear match to our expectations from the wider literature. In the first four rows we indicate the hypotheses where the literature suggests a strong positive result. In fifth row we indicate the results for the Endowment Effect (under the high probability lottery) where we would expect to be underpowered in our lab sample but not in the larger online samples. Finally, in the last two rows we indicate the results from the two hypotheses where we expect a null result (Present bias, and the theoretically ambiguous Donor responses). Without differential demand we replicate the expected literature results, in the lab sample (as discussed above) and in both the MTurk and Prolific samples.

In final set of columns we examine the same hypotheses but where we apply differential demand across the treatment and control. In the *Deflated* columns we use positive-demand in the control and negative-demand in the treatment, attempting to decrease (and potentially reversing) the resulting effect with  $\hat{\beta}_\Delta = \Delta\mu^\ominus = y_B^- - \mu_A^+$ . In the *Inflated* columns we do the converse, using negative-demand in the control and positive-demand in the treatment, attempting to increase the resulting effect via  $\Delta\mu^\oplus = \mu_B^+ - \mu_A^-$ .<sup>21</sup>

---

<sup>21</sup>Note that each result is calculated using the relevant RMSE  $\hat{\sigma}$  and sample-size  $N$  from the selected

Examining the first four rows, we see that differential demand does not lead to false-negatives or reversals. This is true for each of the hypotheses where the literature leads us expect a clear positive result, and for each of the three populations. Despite applying strong and differential demand across conditions, we are incapable of generating an inferential reversal, with significant positive findings across all conditions.

In the fifth row, we examine the hypothesis where our power calculations suggested an underpowered result for the lab sample size. Here, we do see some qualitative reversals. For the lab, as we outline above, we do find a strong positive effect ( $p = 0.001$ ) with differential demand where we found a null without demand ( $p = 0.731$ ); but where the direction in the results here is in the opposite direction from the demand-direction. However, for both MTurk and Prolific we do find that by using differential demand to decrease the treatment effect, we can shift from a positive finding in No demand ( $p < 0.001$  for both) to a null result in the Deflated column ( $p = 0.340$  in MTurk and  $p = 0.232$  in Prolific).

Finally, in the bottom two rows we examine the effects from differential demand on the hypotheses where we expect a null. Here, the results in our online sample do diverge from the lab. In MTurk we can produce both significant negative ( $p = 0.030$  and  $p = 0.044$ ) and positive results ( $p = 0.034$  and  $p = 0.013$ ) using the appropriate demand treatments for the Present-bias and Donor-response hypotheses (respectively). For Prolific we can produce significant positive findings in both hypotheses ( $p = 0.029$  and  $p = 0.004$ ), however the negative effects we generate with demand are only marginally significant ( $p = 0.085$  and  $p = 0.074$ ).<sup>22</sup>

Overall, in terms of significance, we are not able to generate false positives for any of the clear positive results from the literature. Where we do generate differences is in terms of the false positives (and false-negatives for the more-marginal/under-powered hypothesis). For the hypotheses where the literature leads us to expect a null effect—and where we replicate this null in all populations in No demand—when we apply differential demand across the comparative static in the online populations we can generate significant findings.

While Table 1 does document a clear shift in the inferential findings for the online populations, it is not clear that this is driven by the specific population *per se*, or more demand treatments.

---

<sup>22</sup>For the lab, the only reversal in the induced direction that we find is for the Donor response, where we can generate a marginally significant result ( $p = 0.074$ ) for the amount given up without a match (under positive demand) over the amount with a match (under negative demand).

mechanical features of the collected sample. In analyzing the quantitative effects from demand in each of the eight experimental decisions, we certainly do find more systematic directional effects in the online samples.<sup>23</sup> While the online samples do more consistently move in the induced direction of demand, it is unclear that the scale with which we are affecting the result is different online versus the lab. In particular, when we consider the  $t$ -statistic given in (10), the inferential statistic is the product of two separate effects:

$$t = \underbrace{\sqrt{\frac{N}{2}}}_{\text{Precision in estimation}} \cdot \underbrace{\frac{\Delta\mu}{\hat{\sigma}}}_{\text{Effect size}}. \quad (11)$$

A natural question is whether the demand-driven shifts in inference in our online samples are driven by more malleable effect sizes (the treatment difference, normalized by the pooled standard deviation), or instead more mechanically, where the larger samples collected online lead to a significant demand effects through greater estimation precision but the same effect? In the next section we directly examination of how demand affects effect sizes—where we will demonstrate a remarkable similarity across populations for our different hypotheses, both in terms of the overall levels and in range of effect possible through demand.

## 4.2 Normalized Effect Size and Demand

In this section we examine how the normalized effect size  $d = \Delta\mu/\hat{\sigma}$  varies with differential demand across population. The measure of effect—referred to as the Cohen’s  $d$  in the literature—indicates the difference in means between the treatments ( $\Delta\mu$  estimated by  $\hat{\beta}_\Delta$  in (9)) measured in multiples of the pooled standard deviation (the root-mean-squared error  $\hat{\sigma}$  from the regression).<sup>24</sup> Examining the approximation for the  $t$ -statistic in equation (11), we see that the inferential test is the product of two distinct forces: a sample-size

<sup>23</sup>That is, looking at the expected versus realized signs in the difference between the Positive and No demand ( $\mu^+ - \mu^0$ , expected to have a positive sign), and the Negative and No demand treatments ( $\mu^- - \mu^0$ , expected to have a negative sign), we do not find consistent directional effects in the lab ( $p = 0.304$  from a Fisher’s exact test) but do find consistent directional effects from demand online ( $p = 0.020$  for MTurk and  $p = 0.003$  for Prolific).

<sup>24</sup>While similar in spirit to the  $z$ -scores reported in dQHR, variability here is pooled across treatment and control, after adjusting for the level difference in means. While the Cohen’s  $d$  measure is typically motivated by the two-sample  $t$ -test, setting this up over a treatment-effect regression in Equation 9 allows us to generalize to alternative inferential models used in experiments, while using a more-common econometric frame for economists.

effect on the precision of estimation ( $\sqrt{N/2}$ ) and the size of the effect (the Cohen's  $d = \Delta\mu/\hat{\sigma}$ ). As such, if we find that both the levels and range in effect sizes attributable to demand are similar across populations, this will allow to more clearly attribute any difference in inference to sample-size rather than malleability of the population.

In Figure 5, we illustrate the Cohen's  $d$  measure of effect for our seven comparative statics (ordered in the same way as Table 1, by the expected effect size from the literature) and three populations (white for the lab, lighter gray for MTurk and darker gray for Prolific). For each assessed comparative static and population, the plotted diamonds indicate the Cohen's  $d$  in the pooled sample (no differential demand across treatment/control). To show the range in effect attributable to differential demand, the shaded bars indicate the range in Cohen's  $d$  from differential demand across treatment and control, attempting to either minimize or maximize the estimated treatment effect.<sup>25</sup> Overlaid on the figure we indicate a precise null effect ( $d = 0$ , the solid black line) and two rule-of-thumb regions proposed by Cohen, small effects ( $|d| \leq 0.2$ , the red hatched region) and large effects ( $|d| \geq 0.8$ , the blue hatched region).

One of the first takeaways from the figure is that the pure effect sizes are strikingly similar across the three populations. Outside of the figure data, when we examining the Cohen's  $d$  values in each separate demand treatment (so assessing comparative statics in either No Demand, Negative Demand or Positive Demand, without differential application) the correlation in the effect sizes is 0.94 (0.96) between the lab and MTurk (Prolific), while the correlation is 0.99 between the two online populations. However, more than just correlation, the raw scale of the effects sizes are in agreement. Regressing the 21 Cohen's  $d$ 's from one population (7 comparative statics across three fixed demand conditions) on the values from an alternative population, we cannot reject that the slope coefficient is one in any of the six comparisons.<sup>26</sup> Given a Cohen's  $d$  measure of effect from one

<sup>25</sup>For each comparative static (signed using the given equation), we assess the demand-maximized effect size (Cohen- $d^{\oplus}$ ) using data from the treatment  $A^+$  and control  $B^-$ , and the demand-minimized effect size (Cohen- $d^{\ominus}$ ) using data from the treatment  $A^-$  and control  $B^+$ . For the demand range we use the minimum and maximum value across the pooled sample, demand minimized and demand maximized values. If the minimum and maximum value for the range are not given by  $d^{\ominus}$  and/or  $d^{\oplus}$ , then we separately plot those points. In realization, the bottom and top corners of the demand range are always given by  $d^{\ominus}$  and/or  $d^{\oplus}$  in the online samples. In the lab sample there are three cases where the minimum or maximum effect size is not the relevant differential demand value, where in all three the lowest effect size is found in the pooled sample.

<sup>26</sup>The lowest  $p$ -value here is 0.17, with estimated coefficients ranging from 0.94 to 1.02. Shifting the hypothesis, we do reject a value of 0.9 in three of the six regressions (95 percent confidence), and a value of 1.1 in four of the six. Moreover, this relationship for effect size is not a mechanical feature of the experiment, where performing the same exercise over the raw treatment effect difference  $\Delta\mu$ , we reject a slope of 1 in all

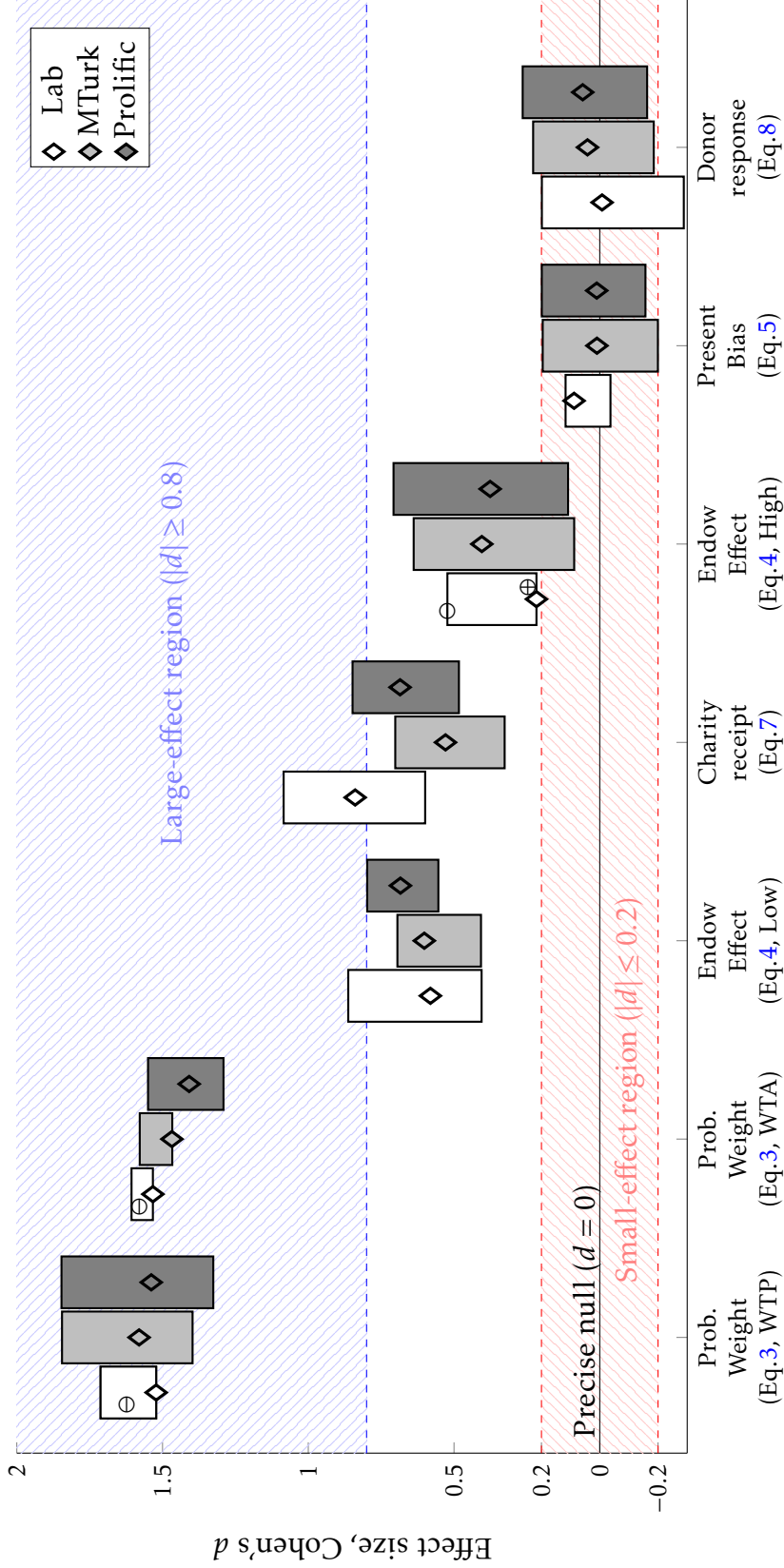


Figure 5: Effect size: overall effects and demand ranges for Cohen's  $d$

Note: Figure shows the directional Cohen's  $d$  (direction chosen for the 'intuitive' direction for the null effects). Diamonds show comparative statics using all data to estimate the treatment effect. Demand ranges are shown using the bars, where we differentially use demand treatments across treatment/control. Bars indicate the value range across Cohen's  $d^\oplus$  and Cohen's  $d^\ominus$ . The small-effect region indicates the effect sizes with a Cohen's  $d$  lower than 0.2. Where the smallest or largest Cohen's  $d$  for the given range is not given by the differential demand value of Cohen's  $d^\ominus$  or Cohen's  $d^\oplus$  (respectively) then we indicate the anomalous value with a  $\ominus$  or  $\oplus$  marker.



population, this would be the cardinal effect size expected in an alternate population.

Looking at the effect sizes from the pooled data (the diamonds in Figure 5), the largest Cohen's  $d$ 's are found in the probability weighting hypotheses, in the endowment effect for the low-probability lottery, and in the charity receipt under a match. The smallest effect sizes are found in our intentional-null over present bias and the theoretically ambiguous result over the donor response (statistically insignificant in all pooled comparisons for all populations). Finally, in the middle we have the endowment effect over the high-probability lottery, where the effect size is small and insignificant in the lab and more moderate in the online samples.<sup>27</sup>

Turning to the range in effect possible with differential demand (illustrated by the bars in Figure 5), our results indicate that the range in the Cohen's  $d$  attributable to differential demand is often smaller in the lab samples, where the average width of the demand ranges is 0.27 for the lab, 0.37 for MTurk and 0.40 for Prolific. The lab also tends to have a more asymmetric response to demand, with stronger effect when using differential demand to increase the Cohen's  $d$ . The pooled  $d$  is on average 0.32 of the way between the  $d^\ominus$  and  $d^\oplus$  values) in the lab, where in the online samples the pooled  $d$  tends to be right in the middle of the demand range (average relative position of 0.47 and 0.48 for MTurk and Prolific). This asymmetric response for the lab somewhat distorts the range comparisons, where the median difference between  $d^\oplus$  and the pooled  $d$  is 0.19 in all three populations. In comparison the median difference between  $d^\ominus$  and the pooled  $d$  is -0.12 in the lab, and -0.20 in both MTurk and Prolific.

In summary, while it is true that demand has less consistent directional effects in the lab sample, the evidence across our experiments suggests that differential demand can move the Cohen's  $d$  by approximately 0.2 in either direction.<sup>28</sup> Given this upper bound—where we again emphasize that the extreme nature of the differential demand we induce should make this a very slack bound for reasonable experimental designed—to achieve a false negative we would need to start with a relatively weak true effect with a Cohen's  $d$  between 0.2 and 0.4. In contrast, if we wanted to create a false positive when the true

---

six comparisons.

<sup>27</sup>Ordinally, the effects sizes are broadly in line with our expectations from the literature across the seven hypotheses. The only substantive difference here is the ordering between the low-probability Endowment Effect and the Charity Receipt in the lab study, with a larger cardinal gap expected between these two hypotheses.

<sup>28</sup>If we instead focus on the worst-case demand range (the high-probability Endowment Effect on MTurk/Prolific, a shift in the Cohen's  $d$  of about 0.33 in either direction is the approximate bound. However, focusing on outliers in this way would seem to be overly conservative.

effect was a precise null, our results suggest that at best we can generate a small effect size, a Cohen's  $d$  with a magnitude of approximately 0.2.

However, because the only false-positives we can create with demand have very small effect sizes, this is where the sample size becomes important, and potentially perverse in the interaction with demand. For a small demand-generated effect size of 0.2 to be statistically significant, we require moderate to large sample sizes ( $N \geq 200$ , so more than 100 observations in each of the two treatments). The collected sample sizes therefore explain why we do not detect false positives for present bias or the donor response in the lab ( $\approx 75$  participants for each decision), while we do detect significant effects in the two online populations ( $\approx 250$  participants for each decision).

### 4.3 Effect and Sample Size in Economic Experiments

To contextualize our findings for effect and sample size, we examine recent experimental papers in the *American Economic Review* (AER) over the past six years (2019–24).<sup>29</sup> Selecting experimental work that has either a treatment-control or difference-in-difference design, we found 28 articles published in the AER, using a combination of laboratory, online, representative, or lab-in-the-field samples.

From reading the 28 articles we extracted 308 distinct experimental comparisons (an average of 11 per article) where a qualitative treatment effect is assessed over an economic outcome  $y$  through experimental treatment variation.<sup>30</sup> While the specific article comparisons possibly had additional econometric controls (subject fixed effects, demographic controls, etc.) our focus was on understanding the raw effect size between treatments. To do this we characterize the effect size as follows:

1. Is the treatment effect identified via a single treatment/control comparison, or is it instead identified via difference-in-difference?
2. Does the dependent variable  $y$  have intensity or is it binary?

---

<sup>29</sup>While we might expect effect sizes to potentially be larger in a top general-interest journal like the AER than a field journal, focusing the analysis on this journal allows us to understand the dangers from demand-effects for the most-important behavioral papers. Moreover, the AER's clear data policy allows us to download the data and code repository to re-examine each article

<sup>30</sup>In terms of process, the 28 articles in the sample were divided across the coauthors according to topic. Each coauthor extracted experimental comparisons from the assigned article that used experimental treatment variation. Comparisons were then coded according to: (i) the dependent variable (either binary or not); (ii) the identification design (either treatment/control or difference-in-difference); and (iii) whether the result was characterized in the article as a null.

For the majority of our comparisons (61 percent) we have a dependent variable with intensity across a simple treatment/control comparison. For these comparisons we run the regression in Equation 9 to calculate the Cohen’s  $d$  as  $d = \hat{\beta}_\Delta / \hat{\sigma}$ . For a minority of comparisons (30 percent) while the identification is still via a simple treatment/control comparison, the dependent variable  $y$  is binary, so the standard deviation is a pure function of the treatment levels. For these comparisons we compute the Cohen’s  $h$  measure which has a similar quantitative interpretation to the Cohen’s  $d$ .<sup>31</sup> For 9 percent of the comparisons, identification is via difference-in-difference, with exogenous treatments in a  $2 \times 2$  design over  $\{A, B\} \times \{C, D\}$ . For these comparisons we run the regression

$$y_i = \beta_0 + \beta_{\Delta_A} \cdot \mathbf{1}(\text{Treatment}(A, \cdot))_i + \beta_{\Delta_C} \cdot \mathbf{1}(\text{Treatment}(\cdot, C))_i + \beta_{\Delta\Delta} \cdot \mathbf{1}(\text{Treatment}(A, C))_i + \epsilon_i. \quad (12)$$

We then compute a pseudo-Cohen’s  $d$  for the interaction effect via  $d = \hat{\beta}_{\Delta\Delta} / \hat{\sigma}$  (the difference-in-difference treatment effect measured relative to the decision heterogeneity).<sup>32</sup>

Figure 6 illustrates the results from our effect-size meta-analysis (where Table 2 in the Online Appendix) provides summary statistics. On the vertical axis of the figure we indicate the measured effect size (the absolute value of the Cohen’s  $d$  or  $h$ ) and on the horizontal axis we plot the total number of participants. Each of the 308 points in the figure represents a different experimental comparison (from the 28 AER articles). Marker shapes in the figure indicate the population participants were drawn from: squares for laboratory samples with undergraduates, diamonds for online populations (including both MTurk and Prolific), circles for representative samples, and triangles for lab-in-the-field studies. We shade each marker to delineate between each comparison was presented in the original article we shade each point as white to indicate results presented as nulls, and gray to indicate results presented as significant.

Overlaid on the figure we indicate three key regions: (i) small effect sizes, the region

<sup>31</sup>The measure is double the difference in arcsin-root of the different treatment proportions. So  $h = 2 \left[ \arcsin \left( \sqrt{\hat{\beta}_0 + \hat{\beta}_\Delta} \right) - \arcsin \left( \sqrt{\hat{\beta}_0} \right) \right]$ . The Cohen’s  $h$  has a similar quantitative interpretation to the Cohen’s  $d$ , with absolute values of 0.2 and below characterized as small, 0.5 a medium effect, and larger than 0.8 as a large effect.

<sup>32</sup>For a small number of the difference-in-difference comparisons (5 comparisons from 27 with diff-in-diff designs, 1.6 percent of the total comparisons) the dependent variable is binary. Given the additional degrees of freedom in the linear-probability model (so that the standard deviation is not pinned down by the interaction) we indicate the pseudo-Cohen’s  $d$  in the same manner as the standard diff-in-diff comparisons.

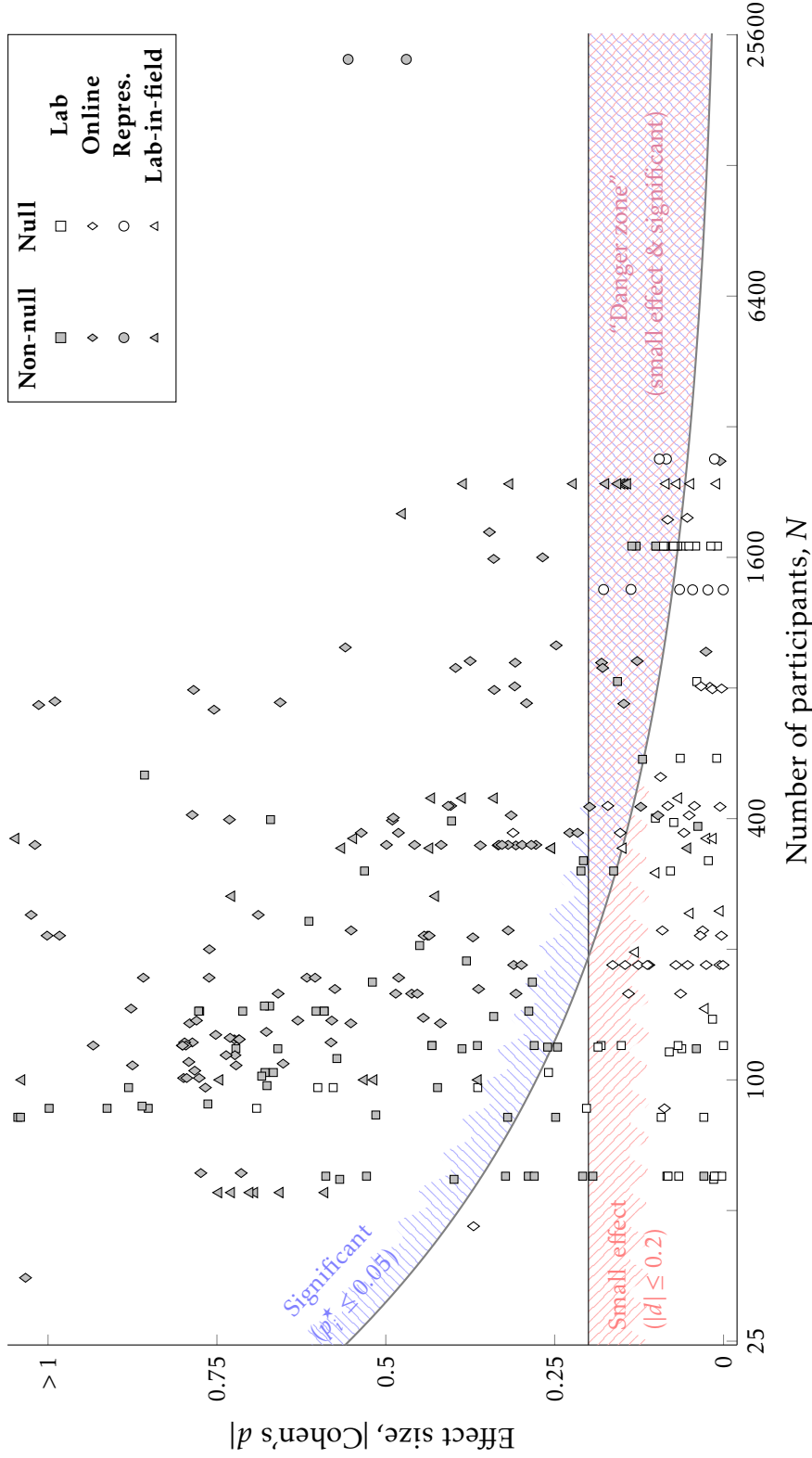


Figure 6: Effect and sample sizes for experiments in the *American Economic Review*

Note: Figure illustrates the number of participants (on a log scale) against the Cohen's- $d$  effect size for experimental articles appearing the *American Economic Review* between 2019–24. White points indicate results characterized as nulls in the article (per coder's assessment), where shaded points are characterized as results. Where dependent and independent variables are binary, we replace the Cohen's  $d$  with the Cohen's  $h$  difference in arcsines (though we cannot do this for difference-in-difference designs). Shaded regions indicate: (i) small effect sizes ( $|d| < 0.2$ ); (ii) significant effects based on participant observations in a balanced between-subject design ( $p_N^* \leq 0.05 \equiv \sqrt{N/2} \cdot |d| \geq 1.96$ ); (iii) intersection of small and significant is completely shaded.

Table 2: Summary statistics for AER comparisons: Move to ONLINE APPENDIX

<b>Comparisons</b>	<b>All (308)</b>	<b>Lab (107)</b>	<b>Online (145)</b>	<b>Repres (11)</b>	<b>Lab-in-field (45)</b>
Effect size ( $ d $ )	0.34	0.26	0.42	0.08	0.32
Participants ( $N$ )	214	120	215	1,350	342
Characterized as null effect	31%	37%	23%	82%	29%
Small effect	37%	43%	28%	82%	42%
Significant effect	70%	62%	77%	55%	73%
Danger zone	10%	12%	6%	36%	16%
<b>Non-null comparisons</b>	<b>Overall (213)</b>	<b>Lab (67)</b>	<b>Online (112)</b>	<b>Repres. (2)</b>	<b>Lab-in-field (32)</b>
Effect size ( $ d $ )	0.49	0.43	0.55	0.51	0.43
Participants ( $N$ )	172	119	215	22,476	342
Small effect	13%	18%	8%	0%	19%
Significant effect	92%	84%	96%	100%	97%
Danger zone	8%	10%	4%	0%	16%

*Note:* Table indicates summary stats across the 308 experimental comparisons in our 28 articles published in the *American Economic Review* in the last six years (2018–2024). Figures in parentheses give the number of experimental comparisons for each experimental comparison. Figures for the effect size and participants indicate medians across the relevant comparison samples, all other figures are given as proportions.

below the  $|d| = 0.2$  line, the upper bound on a demand-generated false-positive from our experiments, which is also the level described as a “small effect” by Cohen; (ii) statistically significant results, the region above the curve defined by  $\sqrt{N/2} \cdot |d| = 1.96$ , where we here assume a significance level of 5 percent and participant-level observations in a between-subject design; (iii) a “danger zone,” the intersection of the previous two regions, comparisons with both a small effect size and a large enough sample for statistical significance.

Across all the comparisons, the median effect size is 0.34 and the median participant count is 308. Comparisons from online populations tend to have significantly larger published effect sizes (0.42 vs. 0.21) and larger participant counts (215 vs. 120) than the comparisons drawn from the laboratory population.<sup>33</sup> However, a lot of the difference in effect size across population is driven by a different proportion of comparisons characterized as nulls by the authors (the white points in Figure 6, 37 percent for lab comparisons,

<sup>33</sup>Using a median test, we find that the median effect size and participant counts for the online comparisons are significantly greater than the lab, with  $p = 0.041$  and  $p = 0.001$ , respectively.

but only 23 percent for online). Examining only the comparisons where the authors characterize it as a significant finding, the median lab effect size is 0.43 while for the online populations it is 0.55 (insignificantly different with  $p = 0.535$ ).

For Figure 6, the proportion of comparisons in each of the three labeled regions are: (i) 37 percent of the comparisons have small effect sizes (13 percent when we exclude the white points characterized as nulls);<sup>34</sup> (ii) 70 percent of the comparisons are in the significant region (92 percent when we exclude the results characterized as nulls);<sup>35</sup> and (iii) 10 percent of comparisons are in the intersecting danger zone (8 percent for non-nulls).

## 4.4 Discussion/Conclusion for Effect Sizes

What then should we conclude from the above exercise? Focusing only on results characterized by the original-article authors as non-null, the effect-size glass then is about seven-eighths full, or one-eighth empty, depending on your point of view. The large majority of experimental results have effect sizes that our results indicate are simply too big to have been generated with demand. That is, in 87 percent of experimental comparisons where the authors believed they had identified a clear positive effect, an appeal to the substantial effect size would have helped to assuage concerns that an effect was demand driven. Moreover, the majority of experimental comparisons with small effect sizes we find are cases where the original authors explicitly label their results as nulls.

For the remaining one-eighth of papers with a small effect size and significant finding, what sticks out from our meta-analysis is not the population the results are drawn from (with similar proportions across population) but the larger sample size. Looking only at non-null comparisons with small-effects the median participant count is 903 (466 for the lab, 890 for online and 2,364 for lab in the field). In contrast, the median participant

---

<sup>34</sup>Having a small effect size is a strong predictor of the author's characterization, with a 88 percent agreement between the two variables. Of the 95 comparisons the authors characterize as null, 86 are in the small effects region, so 9 mismatches (3 percent of the total comparisons). A larger proportion of the mismatches (28 comparisons, 13 percent of the total) are instead those with a small effect but where the author's do not characterize the result as a null.

<sup>35</sup>DELETE: Despite the strong assumptions in forming the significant region (participant-level observations with no other control variables in the regression), a comparison being in the significant region for effect size is again a strong predictor of the author's result characterization. Of the 308 comparisons we find an 88 percent agreement rate between the significant region and the author's result characterization. The mismatches are approximately balanced between results that the authors categorize as nulls that our measure indicates significance for (20 comparisons, 6 percent of the total) and results the authors characterize as non-null that our measure indicates insignificance (16 comparisons, 5 percent of the total).

count for non-null comparisons with larger effect sizes is 158 (118 for the lab, 200 or online and 304 for lab in the field).

Obviously, having a large sample size is not a smoking gun, but it does put more onus onto authors to clearly outline the size of their effect alongside their  $p$ -values. While we would always advocate for articulating an effect size in a way that makes economic sense within the precise context of the paper, we do think that authors should also make a clear effort to report effect size in an externally valid way. In particular, by reporting clear measures of effect size like the Cohen's  $d$ , readers can clearly separate between large/small effect sizes and precise/imprecise estimation due to the sample size.

It is important to reiterate that we are an extreme case for experimenter demand, where strong experimenter demand is differentially induced across treatments. Similar to dQHR, we stress that we are identifying an upper bound on the potential effect from demand. Having a Cohen's  $d$  below 0.2 certainly does not mean a result is generated by demand, particularly if a clear mechanism for an interaction across treatments cannot be articulated. Acknowledging a small but significant effect size does place some onus on authors to articulate why the effect matters, and how it has been identified. Other options exist for ruling out demand (see [de Quidt, Vesterlund and Wilson, 2019](#), for details) through both appeals to design features and explicit demand treatments.

## 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 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 reversing 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 reverse a directional hypothesis. Although upholding the clear evidence against false neg-



atives, 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 the 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 bounding 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 Alj  z 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   stling, and Erik Wengstr  m.** 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.

- Rigotti, Luca, Alistair Wilson, and Neeraja Gupta.** 2023. "The Experimenters' Dilemma: Inferential Preferences over Populations." University of Pittsburgh working paper.
- 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.