Experimenter Demand Effects[1]

Jonathan de Quidt
Lise Vesterlund
Alistair J. Wilson

September 2024

1. Introduction

Experimenter demand effects are one of the classic critiques of empirical social science and have been discussed for over one hundred years.[2] The general concern is that, knowing they are participating in a study or investigation, the participant tells the researcher what they think the researcher "wants to hear," biasing the research conclusions.

For demand effects to arise, two conditions must be satisfied. First, participants must have some belief (quite possibly incorrect) about what it is the researcher potentially expects or wishes them to do. Second, participants must desire to "help" the researcher in this way, rather than to simply truthfully reveal their preferences, beliefs, or personal information.[3]

While we are not aware of any experimental study where inference on the direction of the comparative static is driven by experimenter demand, the profession has taken, and continues to take, such concerns very seriously. Indeed the absence of demand-driven results may be the many steps taken to minimize such concerns. The goal of this chapter is to provide a simple roadmap to

[1] de Quidt: Queen Mary University of London and Institute for International Economic Studies, Stockholm, j.dequidt@qmul.ac.uk. Vesterlund: University of Pittsburgh, vester@pitt.edu. Wilson: University of Pittsburgh, alistair@pitt.edu. We thank many people for discussions that contributed to the ideas in this paper, especially our coauthors David Danz, Johannes Haushofer, Marissa Lepper, Guillermo Lezama, Priyoma Mustafi, Chris Roth and K. Pun Winichakul.

[2] "[The participant's] general attitude of mind is that of ready complacency and cheerful willingness to assist the investigator in every possible way by reporting to him those very things which he is most eager to find." (Pierce, 1908, in the context of psychological investigation). "The subject's performance in an experiment might almost be conceptualized as problem-solving behavior... he sees it as his task to ascertain the true purpose of the experiment and respond in a manner which will support the hypotheses being tested." (Orne, 1962).

[3] Such motivations might arise from reciprocity toward the experimenter (who is typically financially rewarding their participants); a misguided desire to help with scientific progress; or perceived social pressure. However, it is far from clear that participants would have such preferences. Lab and online participants largely see their participation in experiments a participation in a marketplace for "experimental labor", they participate to make money, not to be "scientific do-gooders" (Camerer, 2015). The cognitive load of a study may well make speculation and concerns for the experimenter second order, and even if present, 'professional participants' are repeatedly informed that for science to work they should do what they (rather than the experimenter) prefer.

a researcher who is planning to conduct a study and wants to engage seriously with concerns about experimenter demand. We do four things.

First, we outline a simple formal economic model of a participant in a research study who is potentially influenced by demand effects. This model gives us a framework around which to structure the rest of the discussion.

Second, we summarize a body of methodological tools that have become "best practices" in experimental economics, which are widely believed to be effective ways to mitigate or minimize biases due to experimenter demand and are adopted in most experimental studies. In many cases, adhering to these best practices will be sufficient to allay your, or a potential reviewer's concerns about demand effects. The discussion and evidence that we review is mainly focused on laboratory, online, and lab-in-the-field experiments, but most of these practices are applicable in randomized-control-trial field experiments as well.

Third, we outline a set of tools for bounding potential demand biases. These are most useful for cases where, despite adopting best practices, the researcher or reviewer has lingering concerns about demand effects, for stress-testing of a design and findings, and for generating knowledge about demand effects as an empirical phenomenon as has been done in the studies we cite. Bounding methods are equally applicable in lab/online/field settings but typically require extra data collection which can be costly. We review several ways that researchers can minimize these costs.

Fourth, and finally, we review evidence on demand effects. Studies using the bounding approach mentioned above have shown that experimental participants can be induced to change their behavior by deliberately and explicitly telling them about a desired direction of choice, confirming that we are right to address concerns about experimenter demand in our own studies. But the good news is that across studies the effect of deliberately inducing experimenter demand is generally modest. Several attempts to use demand effects to reverse the direction of an experimental treatment effects (i.e., change a qualitative inference), have failed.

Our chapter draws heavily on three papers by some subset of us. De Quidt, Haushofer and Roth. (2018) developed the theoretical framework and bounding approach and provide evidence on magnitudes with MTurk participants. De Quidt, Vesterlund and Wilson (2019) sets out our views on design "best practices" for demand-effect mitigation and provides an extensive review of the literature demonstrating that these practices are widely adopted. Winichakul et al. (2024) provide new experimental evidence on the magnitude of demand effects, showing that they do not change inferences from a sequence of canonical tasks, and that this conclusion is robust across several different participant pools including laboratory and online participants.

We give a high-level overview of those prior works, updating with relevant recent studies but trying to avoid too much repetition or detail. Readers who want to adopt the approaches that we advocate should take a close look at those papers for fuller details.

2. Theoretical framework

De Quidt et al. (2018) propose a belief-based model of experimenter demand effects that we summarize here. The model is useful for conceptualizing where potential demand effects could be coming from, and how to address them.

In the model, the experimenter is interested in "natural" behavior, whereby experimental participants choose an *action* a (e.g., a donation in a dictator game, an effort choice, an auction bid, a game strategy, a belief report) to maximize their utility. Here we will consider actions that are real valued (e.g., the amount given in a dictator game), but the same constructions will work for ordinal data (e.g., a decision to cooperate/defect in a prisoner's dilemma).

The optimal action depends on *decision-relevant design features* ζ, which captures all aspects of the design that matter for decision making in the absence of experimenter demand bias. Examples are the rules of the game being played, the incentives, the decision-relevant information provided to participants, etc. Given ζ, the optimal, "natural" action is denoted by $a(\zeta)$. This is the behavior that the researcher would like to observe and understand.

The experimenter might want to measure a particular action $a(\zeta)$, such as how much money people donate given a certain budget and recipient. Or they might want to measure a treatment effect $a(\zeta_1) - a(\zeta_0)$, such as how giving changes when the donation is or is not matched or how it changes when the framing of the decision changes.

When making their choices, experimental participants may also form a belief about what the experimenter *wants* them to do. This is modelled as a binary state $h \in \{-1,1\}$ where a negative value corresponds to certainty on “the experimenter wants low actions” and a positive value is certainty on “the experimenter wants high actions.” For instance, consider a participant that is naturally quite generous. However, when a dictator game is framed as a “taking game” this participant guesses that they are expected to be selfish ($h = -1$, meaning a low a is expected). Whereas framing the same dictator as a “giving game” might signal they are expected to be more ($h = 1$).

The participant’s belief might be influenced by the decision-relevant design features in ζ, but also depend on other *decision-irrelevant* features that we denote by ρ.[4] We therefore model the belief as a conditional expectation $E[h|\zeta, \rho]$. So, in the above-mentioned example, we would expect $E[h|\text{Giving game}] \geq E[h|\text{Taking game}]$. Complete uncertainty about the direction of the hypothesis is therefore given by $E[h|\zeta, \rho] = 0$, whereas $E[h|\zeta, \rho]$ at the extremes of -1 or 1 correspond to complete certainty about a negative or positive demand, respectively.

Such beliefs are irrelevant if the participant does not care what they perceive the experimenter wants, but those concerned for experimenter demand bias may argue that participants care about following the experimenter’s perceived demands.[5] This is modelled by a preference parameter $\phi(\zeta, \rho)$ whose sign captures the desire to comply with ($\phi(\zeta, \rho) \geq 0$) or defy ($\phi(\zeta, \rho) < 0$) the experimenter’s perceived wishes. The magnitude of ϕ captures the strength of this preference.[67]

Formally the model is set up as follows. The participant is assumed to maximize an objective function that is separable in “natural” preferences $v(a, \zeta)$ (the preferences the experimenter wants to learn about), and a linear-separable demand-driven preference $a \cdot \phi(\zeta, \rho)E[h|\zeta, \rho]$:

$$U(a, \zeta) = v(a, \zeta) + a \cdot \phi(\zeta, \rho)E[h|\zeta, \rho]$$

[4] De Quidt et al. (2018) use ζ to represent all design features, we partition them into decision-relevant and decision-irrelevant to highlight that only the decision-relevant features matter for the natural action $a(\zeta)$. De Quidt (2024) follows the same formulation.

[5] Such influences could be conscious or subconscious/implicit.

[6] The magnitude of ϕ might depend on decision-relevant features (conceivably, large incentives could increase feelings of reciprocity) or decision-irrelevant features (e.g., if the participant learns that the experiment is being conducted by their professor). Some subpopulations might be more susceptible than others as we discuss in sections 3.2.3. and 5.5. A “natural field experiment” (Harrison and List, 2004) where the participant is unaware of the presence of an experimenter can be represented by $\phi = 0$.

[7] De Quidt et al. (2018) present within-subject evidence suggesting that defiers (with $\phi < 0$) are rare.

The “natural action” $a(\zeta)$ is the maximizer of v, the action that would be taken if there was no experimenter demand bias, either because $\phi = 0$ (no desire to please the experimenter) or because $E[h|\zeta, \rho] = 0$ (the participant is completely unsure about the direction of experimenter demand).

We will call the *latent-demand* action, which is the action observed by the experimenter and potentially distorted by demand effects, $a^L(\zeta, \rho)$.[8] If v is concave, we predict that when participants are motivated to comply with the experimenter’s perceived wishes ($\phi > 0$), their behavior will be distorted in the direction of their belief about h, with the magnitude of distortion related to the magnitude of $\phi E[h|\zeta, \rho]$.

The bias due to experimenter demand for a particular action is therefore:

$$a^L(\zeta, \rho) - a(\zeta, \rho)$$

while the bias in a treatment effect estimate will be:

$$[a^L(\zeta_1, \rho_1) - a(\zeta_1)] - [a^L(\zeta_0, \rho_0) - a\ (\zeta_0)]$$

This framework suggests two approaches to dealing with demand effects. The first is to try to minimize the bias terms above. “Mitigation” approaches that (i) dampen participants’ motives for pleasing the experimenter (shrinking ϕ), (ii) weaken their inferences about what is expected of them (shrinking $|E[h|\zeta, \rho]|$) will accomplish this.[9] We cover these in the next section. “Bounding” approaches instead seek to *amplify* demand effects in a controlled way so as to construct credible upper and lower bounds on natural behavior (i.e., partially identify $a(\zeta)$). We cover those in the following section.

3. Mitigating experimenter demand effects through design

An assessment of the potential role of experimenter demand effects starts with a review of the experimental instructions and procedures. Careful experimental design along with detailed documentation of the participant-facing interaction helps reduce experimenter demand concerns. The expectation in the profession is that when submitting a paper, you provide access to all

[8] “L” denotes “latent” (unobservable to the experimenter) demand effects.

[9] When estimating treatment effects the bias will tend to be smaller when participants in the treatment and control groups hold similar beliefs about experimenter demand: $E[h|\zeta_1, \rho_1] - E[h|\zeta_0, \rho_0] \approx 0$. Many of the design features that we discuss have the additional benefit of reducing variation in beliefs between treatment groups, reducing the size of the bias. De Quidt et al. (2018) discuss under what conditions harmonizing beliefs can eliminate bias altogether.

material participants see, this includes instructions, computer screen shots, survey measures, etc. All treatment-specific changes in instructions must be included and treatment changes should be presented side-by-side for easy comparison.[10] Further, we strongly encourage you to write and include a script detailing all procedures and verbal statements made during the study.[11] Collectively these materials not only allow for an assessment of experimenter demand, but also of your identification strategy, and are critical for subsequent replication efforts. As evidence that it is standard practice to provide access to the experimental instructions, we found in reviewing the published experimental literature from 2012-2017 that 80 percent of papers provide access to instructions, and that the majority does so for all treatments (de Quidt et al., 2019).[12]

But what are reviewers looking for when assessing the experimental instructions and procedures? What designs might reduce concerns for experimenter demand? While we do not know of any economic studies where experimenter demand drives the qualitative inference on comparative statics, the profession has nonetheless taken such concerns seriously and has responded by developing and adopting a set of best-practice designs that reviewers expect to see implemented when possible. Aiming to learn about participants' true preferences ($v(a, \zeta)$) best-practice designs for reducing experimenter demand concerns center on two main objectives: one is to minimize participant speculations on the experimenter's hypothesis (aiming to push the expectation $E[h|\zeta, \rho]$ to zero), and the second is to minimize participant willingness to deviate from their preferred choice to confirm a potential hypothesis (aiming for participants to have no desire to comply with a perceived hypothesis, thus pushing $\phi(\zeta, \rho)$ to zero).[13] For most best practices there

[10] For example, the instructions in Bracha and Vesterlund (2017) denote four treatment variations as follows, "During the study, [T1: we will tell you how much each member of your group earned, and how much each member donated to the child he or she is paired with]. [T2: we will tell you how much each member of your group earned, but we will not tell you how much each member donated to the child he or she is paired with] [T3: we will tell you how much each member of your group donated to the child he or she is paired with, but we will not tell you how much each member earned] [T4: we will not tell you how much each member of your group earned, nor will we tell you how much each member donated to the child he or she is paired with]."

[11] This practice of writing detailed scripts is commonly adopted for laboratory experiments, and while there is less interaction with online participants, the practice should be encouraged for all experimental interventions, in particular for field experiments.

[12] Included in the review were published lab and online experiments, as well as lab-in-the-field experiments. Randomized-control-trial field experiments were not included in the review.

[13] The ideal aim is to eliminate speculation such that $E[h|\zeta, \rho] = 0$, however note that for inference on treatment effects it is sufficient that beliefs are held constant across treatments. Side-by-side comparisons of treatment instructions are helpful for such an assessment.

is at most scant evidence that their adoption impacts experimenter demand, however their wide adoption suggests that they are perceived to work (de Quidt et al., 2019).

We will review best-practice designs for reducing experimenter demand, first presenting designs that aim to make the hypotheses less salient and more difficult to guess, and second presenting designs and procedures that aim to reduce the distortion from the participant's 'natural' choice $a(\zeta)$. Our emphasis is on practices adopted for experiments in the lab and online, and when conducting lab-in-the-field studies, however, we also discuss how these practices apply to randomized-control-trial experiments in the field.

3.1. Concealing the hypothesis

Most economic experiments are designed to test hypotheses about causal effects on economic behavior. That is, the aim of the experiment is to assess how changes in some independent variable affect a dependent variable. In most laboratory and online studies the interest is one of determining the qualitative response, that is, what is the direction of the response and is the response economically meaningful, whereas field experiments more often are concerned about the precise magnitude of the effect.[14] A varied set of design choices aim to reduce experimenter demand by concealing the hypothesis under investigation ($E[h|\zeta, \rho] = 0$) or at a minimum holding beliefs constant across treatments. While some changes are as small as hiding the names of the researchers involved in a study to eliminate speculations on the potential purpose of a study, others are more substantial. Most significant are abstract framing of the decision environment and efforts to conceal the variables of interest. We will discuss these in order.

3.1.1. Abstract framing

In modeling decision making in the real-world, economists simplify the environment of interest, focusing on what is seen as key drivers and ignoring details that are not seen as critical to

[14] Kessler and Vesterlund (2015) argue that more so than the precise magnitude of an effect, most lab experiments aim to determine the direction or sign of an effect, and whether the response is economically meaningful. Where communication of experimental findings in the laboratory centers on causal inference, noting "Indeed, the non-parametric statistical methods commonly used to infer significance rely solely on qualitative differences. Few experimental economists would argue that the precise magnitude of the difference between two laboratory treatments is indicative of the precise magnitude one would expect to see in the field or even in other laboratory studies in which important characteristics of the environment have changed.' Similar assessments should be expected for both online and field experiments.

behavior.[15] Abstract framing carries over when exploring decision making in an experiment. Rather than describing decisions as efforts to improve the environment, purchase insurance, apply for schools, or donate organs etc., experimental economists instead present participants with choices over actions that have associated payoffs mirroring the incentives of those decisions. Abstract framing is seen as uncovering fundamental characteristics of preferences and behavior that can be applied more generally, and as potentially concealing the experimenter's hypothesis and in turn reducing experimenter demand. While context is important in some examinations, a naturally framed experiment may cause participants to anchor on the environment and speculate on what is seen as desirable behavior. The hope is that abstract frames focus attention on the provided information and incentives.

While there is evidence that framing can influence behavior, it is often challenging to determine why behavior changes, and we are not aware of studies that demonstrate that abstract framing reduces the potential for experimenter demand. For example, labeling the prisoner's dilemma as a "community game" rather than a "Wall-Street game" or "stock-market game" substantially changes cooperation rates (Kay and Ross, 2003; Liberman et al., 2004; Ellingsen et al., 2012), however, rather than resulting from experimenter demand, changes in labeling simultaneously change beliefs about other participants' choices and in turn change the resulting equilibrium (Ellingsen et al., 2012).[16] Nonetheless abstract framing is often heralded as muting the perception that one decision is more appropriate than another, hence reducing the potential for experimenter demand. Another advantage of abstract framing is that it makes it easier to introduce treatment variations that hold all other characteristics constant, similar changes may seem less natural in context-rich experiments and require more substantial changes in framing.

As evidence of best practice, we find in reviewing the experimental literature from 2012 to 2017 that abstract framing is widely adopted. Abstract framing describes the designs used by 96.1 percent of published papers in the top field journal *Experimental Economics* and for 89.4 percent of the experimental studies published in the *Top-Five Journals* (de Quidt et al., 2019).[17]

[15] The simplicity aim is not unique to modeling in economics "Everything should be made as simple as possible, but no simpler" applies throughout academia.

[16] Dreber et al. (2012) show that dictator-game giving is unaffected by neutral and non-neutral frames.

[17] Included as the *Top-Five Journals* are the *American Economic Review, Econometrica, the Journal of Political Economy, the Quarterly Journal of Economics,* and the *Review of Economic Studies*.

3.1.2. Concealing variables of interest

An important design choice in experimental economics is whether inference on causal impact is done by manipulation or subtraction. Design-by-manipulation is a case where treatment variation is used to exacerbate the effect the independent variable has on the behavior of interest, whereas design-by-subtraction refers to designs where a treatment eliminates the potential role for the independent variable, holding all else constant. More recent experimental studies have expanded to draw inference from design-by-correlation, where an external measure of the independent variable is used to assess correlation with the behavior of interest. For example, in the context of evaluating if risk aversion drives behavior, design-by-correlation would mean separately eliciting a measure of risk preferences and determining whether it correlates with behavior, design-by-manipulation would explore treatment variation where risk aversion is predicted to further distort deviations, and design-by-subtraction would introduce a treatment, where holding all else constant the potential for risk aversion to drive behavior is removed. A weakness in both design-by-correlation and in design-by-manipulation is that other factors may be causing the change in the dependent variable, leaving design-by-subtraction as the gold standard of the profession.

Regrettably experimenter demand throws a wrench in the perception that design-by-subtraction provides clean identification of causal effects. In particular, participants who are aware of treatment variation may not only respond to changes in the independent variable, but also to the inferred hypothesis on how such treatment changes impact behavior ($E[h|\zeta,\rho] \neq 0$). This experimenter demand confound has caused scholars to limit the use of **within-subject designs,** where participants make decisions for all relevant independent variables and are fully aware of the independent variable of interest. When possible, scholars instead rely on **between-subject designs**, where each participant only experiences one treatment, and thus is unaware of the independent variable.[18] In concealing the independent variable, participants cannot guess what

[18] Between-subject design hinges on random assignment to treatment and requires that participant characteristics are balanced across treatment. Charness et al. (2012) provides an in-depth discussion of the tradeoffs involved between within- and between-subject designs, across multiple dimensions. They emphasize the lack of order effects and the reduced potential for demand effects in between-subject designs and the greater statistical power in a within-subject design.

part of the experimental environment is varied between treatments, nor what the underlying hypothesis might be.[19]

Between-subject designs are seen as best practice for reducing experimenter demand, as noted by Camerer (2003, p.41) they have been adopted as the "norm in experimental economics." In reviewing the published experimental literature, we find that this certainly is the case for papers in the profession's field journal *Experimental Economics* where 89 percent of studies rely on between-subject designs, by comparison the share is 59 percent of experimental studies in the *Top-Five Journals* (de Quidt et al., 2019).

Although within-subject designs carry some challenges, there are steps that can be taken to both control and assess the potential for experimenter demand. Inference on the independent variable can be muddied when **progressively revealing** treatments, where treatment changes are provided as the experiment progresses, and initial information only describe a 'series of decisions' that can impact their earnings.[20] When progressively revealing treatments keep in mind that experimental economists adhere to a strict "no-deception rule." Participants must be informed if initial decisions impact future opportunities, choices, or earnings. Assessments of experimenter demand effects may also be ensured by reversing treatments, such that some participants first make decisions in treatment A and then B, while others first decide in treatment B and then A. In reversing treatment order, you can explore treatment effect between-subjects, as well as within-subject, and similarity in the two may be seen as evidence for the absence of experimenter demand (note however, that factors other than experimenter demand can give rise to order effects, e.g., learning, wealth effects, etc.). When reversal of order is not possible, the selected order of revelation should be one that

[19] Levati et al. (2011) examine behavior in either trust or dictator games (with between-subject assignment) but inform participants of the alternative treatment that other participants will receive, in what they call a "hybrid design." They find substantially different behavior from standard between-subject results in the literature, and argue that the provision of information on other treatments in a between-subject design helps subjects interpret the environment and question relatively (thereby increasing the validity of the *relative* results). However, this reasoning seems confounded with experimenter demand.

[20] As evidence that this can have an effect, Burks et al. (2003) find that trusting behavior is lower in trust games when participants know in advance that they will play both game roles during the experiment. An option for concealing the independent variation is to insert time gaps or filler/decoy tasks to make the within-study comparison between treatments less salient. A risk of doing so is that these seemingly innocent decoy tasks may influence the participant's perception of the purpose of the study. Roux and Thöni (2015) find in the context of a Cournot oligopoly experiment no influence of control questions on choices, and this holds independent of participants beint told that the control questions are randomly generated (and therefore less informative about the hypothesis).

minimizes sensitivity to experimenter demand, where important and/or more sensitive decisions are elicited first, and less important and/more robust measures are elicited later.[21]

Progressive revelation of treatment is however infeasible in some within-subject designs. For example, strategy-method elicitations, where participants make decisions contingent on the choices of others, challenge the ability to conceal the independent variable and its impact on choices. For example, Zizzo (2010) argues that the rate of "conditional cooperation" in public-good games can increase when using the strategy method to ask participants to condition their choices on others, and Echenique et al., (2016) worry enough about demand effects in a centralized-market game that they move the environment away from the ecologically valid strategy method (stating a preference) toward a direct method of eliciting a sequence of choices. A step that can help assess the potential for experimenter demand in strategy-method experiments is to simultaneously run a direct-response treatment, where participants in real-time respond to their opponent's actual choice, and compare results from the two. Intriguingly such comparisons suggest that choices are the same under the two elicitations (for example, Muller et al., 2008; Brandts and Charness, 2011).

While experimenter demand effects are thought to be smaller in between-subject than within-subject designs, there is limited evidence of experimenter demand driving the responses in within-subject designs. Lambdin and Shaffer (2009) replicate three classic experiments (the child custody experiment of Shafir, 1993; the Asian Disease experiment of Tversky and Kahneman, 1981; and the marbles lottery of Tversky and Kahneman, 1986) using both between- and within-subject designs, and asking participants to guess the experimental hypothesis at the end of each study. Choices were similar across implementations, and participants failed to guess the hypotheses in both.[22]

[21] For example, measures on demographic characteristics are seen as robust and should be elicited at the end of a study. Studies that examine how men and women differ in behavior go to great lengths to remove references to gender in the study (see for example Niederle and Vesterlund, 2007). This becomes particularly tricky when there is a need to reveal the gender of an opposing player, perhaps best masked by showing a photo of the opponent (Babcock et al., 2017) or by presenting a recorded greeting by the opponent (Bordalo et al., 2019).

[22] See also Alcott and Taubinsky (2015) who find "substantial dispersion in perceived intent" when using a post-survey questionnaire to ask subjects what they thought the intent of the study was, many participants do however guess the correct hypothesis. We caution against eliciting beliefs on expected hypothesis in the lab, where potential contamination across sessions may increase experimenter demand.

The focus in hiding the experimental hypothesis has been on concealing the independent variable, however a few studies also take steps to conceal the dependent variable. This is typically done by including decoy tasks along with the task of interest, thus leaving the participant in doubt on what is the primary variable. For example, in examining a "joy of destruction" game where participants could destroy the endowments of others' Abbink and Sadrieh (2009) embed the choice in a decoy real-effort task, such that participants may be uncertain on which choice is of primary interest. The risk of adding decoy tasks is that these may detract attention from the task of interest, and that they on their own may influence behavior and cause speculations on potential experimenter demand. In contrast to the substantial efforts taken to conceal the independent variable, best practices have yet to be explored and codified for concealing the dependent variable.

3.2. Reducing response to experimenter demand

Some interactions may make participants more prone to experimenter demand. Participants may feel more inclined to deviate from their preferred choice to confirm a perceived experimental hypothesis, when there are no consequences to their actions, when their actions are observed, when there is limited social distance to the experimenter, or when they feel pressured by the experimenter. A series of best practices for reducing experimenter demand aim to leave participants with no desire to respond to a perceived hypothesis, these include using designs where decisions are incentivized and anonymous, and where potential social pressure from the experimenter is minimized by avoiding certain participants and limiting interactions with the experimenter.

3.2.1. Incentives

Choices in experiments are generally incentivized to center participant focus on the decisions at hand and to induce preferences. The hope is that incentives increase attentiveness, reduce noise in decision making, and make it costly to deviate from a 'natural' preferred choice, thus reducing the potential for experimenter demand. While there is evidence that incentivized decisions help reduce noise and improve performance, there is limited evidence of the interaction with experimenter demand.[23] De Quidt et al. (2018) is to our knowledge the only study to explore the interaction

[23] In reviewing the literature Camerer and Hogarth (1999) finds that stakes improve performance and decrease noise in some tasks, but notes that the impact on mean behavior is limited and argues that the response should not be over-emphasized. Amir (2012) replicates a number of classic results online at low stakes using Amazon Mechanical Turk. Camerer (2015) argues that insensitivity to stakes suggests that concerns about demand effects are overblown. Ariely

between incentives and experimenter demand. Comparing decisions with no incentives to those with small incentives, they find at best a small reduction in the response to intentional experimenter demand. It is nonetheless the case that incentivized decisions are used by close to all experimental studies in the *Top-Five Journals* (91 percent) and in *Experimental Economics* (99 percent).

3.2.2. Anonymous decisions

Anonymity helps minimize the interaction between the experimenter and the participant and is thought to reduce the participant's potential sense of pressure to deviate from their natural choice. Anonymity impacts both how participants make decisions, how their decisions are recorded, how they interact with others in the study, and how they interact with you.

To reap the benefits of anonymous decisions it is important that participants are informed of this prior to making any decisions. The institutional review board (IRB) will commonly ask that participants are informed of their anonymity (or lack thereof) in their consent forms. You will typically be required to store and share your data in a deanonymized way and need to inform participants of this fact. Further we recommend that your instructions make clear precisely what anonymity means in your study.

Participant identifiers can be used to track individual decisions without revealing to you or others what decisions were made by whom. For strategic interactions you may not need to reveal any information about what individual made which decision, but if you do, you should ideally use no more than the personal identifier. Online experiments (e.g., *MTurk* or *Prolific*) offer a high degree of anonymity, since usually participants cannot be personally identified and decisions are made far from the experimenter and other participants. For sensitive outcomes, researchers have gone further, deliberately inserting noise into measurement to make it impossible to identify a given participant's response with certainty (e.g., List et al, 2004; Karlan and Zinman, 2012; Fischbacher and Föllmi-Heusi, 2013). When conducting experiments in a setting where multiple participants are gathered (in the lab or in the field) take steps to ensure the promised anonymity. Provide participants with a setting where their decisions and earnings are not seen by others. When

et al. (2009) finds changes in behavior when presented with large incentives. In reviewing the literature Gneezy et al. (2011) demonstrate that behavior sometimes is sensitive to the magnitude of the incentives. See also Enke et al. (2023) for evidence that although high incentives can increase response time, the magnitude of incentives (no, standard, high) do not influence cognitive biases.

possible, seat them in screened booths where others can not see their decisions and pay them in private, so others cannot attempt to infer their decisions.

There is some evidence that anonymity may influence behavior in a manner that is consistent with experimenter demand. Hoffman et al. (1994) found that giving decreased in a double-blind dictator game experiment where the experimenter could not identify how much a participant gave. There were however substantial procedural differences between the two treatments. Holding other characteristics constant across treatments and exploiting more subtle anonymity treatments Barmettler et al. (2012) finds no effect of anonymity on dictator, ultimatum or trust-game behavior.[24] Key in drawing inference on the potential role of experimenter demand is of course whether we are interested in understanding other-regarding behavior that arises in a completely anonymous setting, or in a setting with greater observability.

We find in reviewing the literature of published papers from 2012 to 2017 that anonymous decisions are widely accepted, accounting for 94 percent of publications in *Experimental Economics* and 83 percent in the *Top-Five Journals*.[25]

3.2.3. Participants

Initial experimental studies often relied on convenient samples, with early studies being conducted in classrooms (e.g., Kahneman, Knetsch, and Thaler, 1990) or with staff members (e.g., Dresher and Flood's experiments using staff at RAND, see Roth, 1993). Convenient samples may however be more susceptible to experimenter demand either because participants have knowledge about the underlying hypothesis ($E[h|\zeta,\rho] \neq 0$), or because they care more about confirming the hypothesis ($\phi(\zeta,\rho) \neq 0$). It is advisable to rely on participant pools that mute experimenter demand concerns, that is avoiding participants who are colleagues, friends, students who currently are enrolled in your classes or students who have advanced knowledge of economics. Indeed, the vast majority of

[24] Further, Loewenstein (1999) suggests that the strong emphasis on anonymous decisions in Hoffman et al.'s experiments may have suggested to participants that selfishness was expected.

[25] The form of anonymity is typically limited to being single blind. In experimental economics single- and double-blind, distinguishes between who can identify what decision was made by whom, where the participant being unaware is referred to as single-blind, and both the participant and experimenter being unaware is referred to as double-blind. While potentially possible, for most single-blind economic experiments it would require a challenging investigation to link an individual's choice to their identity. In the medical literature single- and double-blind, instead distinguishes between who can identify the assignment to treatment, where the participant being unaware is referred to as single-blind, and both the participant and experimenter being unaware is referred to as double-blind.

published experimental papers are conducted in experimental laboratories, online, or in the field, with less than 5 percent of studies being conducted in the classroom (de Quidt et al., 2019).

In reducing the potential for experimenter demand effects, it is also important to consider the prior experience of participants in your study. For example, in a between-subject study, you need to ensure that participants are in only one treatment, and you may want to recruit only participants who have no prior experience with related studies.[26] Further some in-person laboratories go to great lengths to avoid participants who previously have been subjected to deception and debriefings, because they may fail to focus on the actions and incentives at hand, and instead speculate on the underlying hypothesis. While there is evidence that the response to treatment differs by participant experience, there is no evidence that it interacts with the response to experimenter demand.[27]

3.2.4. Limiting experimenter-participant interaction

For online experiments the interaction between experimenter and participant is generally reduced to that of instructions and screen shots, and the assessment of undue influence is limited to a review of those written interactions. In-person experiments however present a number of avenues for experimenter influence, and you should aim for a design that minimizes the presence and role of the experimenter. An ex-ante and detailed protocol of experimenter-participant interaction will ensure session and treatment comparisons and reduce the potential for experimenter demand.

Your experimental protocol should describe when, what is said and done, and by whom. It should include details on communication with participants before they get to the study, procedures for checking them in and seating them, details on how instructions will be distributed, and potentially read out loud, a full description of how questions will be answered and what commentary is

[26] Good in person experimental labs have extensive procedures in place to secure that there is a full and precise account of a participant's study history (controlling for multiple emails, variations on names, etc.). Similar participation histories are not available when conducting experiments online or in the field.

[27] For online studies, de Quidt et al. (2018) find very similar responses to explicit experimenter demand across experienced (Amazon MTurk workers) and less-experienced populations (respondents to an online political panel survey). See also Section 5.5 for a discussion of response to experimenter demand across populations in the lab and online (MTurk, Prolific).

permissible, details on how to minimize presence during the experiment, and a plan for how participants exit the study and receive their payment.[28]

Options with less potential bias include having the individuals conducting the experiment be unfamiliar with the purpose of the study, having video-recorded instructions, or having participants read instructions on their own. A potential drawback of the latter is however that instructions will not be common information for participants in a session.

While there is concern that in-person experiments make it possible to influence participant behavior through facial expressions, gestures, or tone of voice (Ortmann, 2005), there is little evidence to support such concerns. For example, Bischoff and Frank (2011) find that behavior is robust to incentivized actors reading instructions to induce treatment effects in a solidarity game (Selten and Ockenfels, 1998).[29]

3.3. Relevance for field experiments

Much of the above discussion carries naturally over to randomized-control-trial field experiments with some nuance. For instance, abstract framing is not commonly employed in these field experimental settings partly because the researcher is usually interested in behavior in the specific context they are studying. But concealing the hypothesis, incentives, anonymity, careful participant pool selection, and limiting interaction are all relevant.

It is quite common in field experiments to employ different organizations for program implementation and for data collection, or, increasingly, to use administrative data for outcome measurement. These touch on several of the above ideas: they can be thought of as a way to limit experimenter-participant interaction, or to conceal the independent or dependent variable. Importantly, field experiments also offer the possibility to conceal the presence of an experimenter altogether (in a "natural field experiment", Harrison and List, 2004).

[28] To ensure that session and treatments are comparable, we recommend responding to questions by using statements from the instructions, and only responding to questions in private, potentially followed up with a public statement on the question asked.

[29] There are however cases where the mere presence of an experimenter affects behavior, for example, Cilliers et al. (2015) show more generous behavior in a lab-in-field dictator game in Sierra Leone when a "silent white foreigner" is present.

3.4. Reviewing and new developments

Our emphasis throughout has been on how you as an author can make design choices to reduce concerns regarding experimenter demand, however it is important to also reflect on how experimenter demand concerns should be considered when you are reviewing a paper. In doing so, keep in mind that while there is evidence that intentional, transparent, and strong experimenter demand can move decision estimates, there is no evidence that it has confounded inference.[30] This is not to say that experimental economic studies are without flaws, but rather that experimenter demand concerns have been second order. As for any study, assessment of identification and relevance should be the first-order concern. Is the inference sound and the causal effect clearly identified? Can we rule out alternative explanations for the observed effect? Is the observed comparative static likely to extend not only to other environments, but also to other parameters? If concerned about manipulation, it may be of interest to get assurance that the results of all related treatments and pilot studies are reported (e.g., Roth, 1994), and to assess whether the study's sample size was set ex ante (Simmons, Nelson, and Simonsohn, 2011). With that said, in evaluating the potential for experimenter demand, start by reviewing the design, read the instructions to assess if participants might have been subjected to undue influence. Pay particular attention to differences in language across treatments. Consider yourself as a participant in the study, would you have been able to make a clear guess over what the experimenter's hypothesis was? Would it have affected your behavior in a meaningful way? And if so how? When raising concerns that the results are affected by experimenter demand, state clearly what you have in mind. What effects or statements in the protocol or instructions are likely to have affected behavior? What direction would the demand effect be in? Is there a treatment or test that would put your concerns to rest, if so, propose it, or offer sufficient detail so the authors can design tests to respond to your concerns.

Our review of the literature made clear that experimental economists have taken concerns for experimenter demand very seriously, and a series of best practices have been developed and adopted to minimize such concerns. Central for these practices are incentivized choices in an abstract frame, and the reliance on between subject-designs with anonymized choices. We found that these procedures consistently are applied in 84 percent of papers published in the top field

[30] For a summary of the response to experimenter demand across domains see section 5.

journal *Experimental Economics*, and very frequently applied (46 percent) in the *Top-Five Journals* (de Quidt et al., 2019).

Further, the profession is developing new procedures for assessing the potential impact of experimenter demand. For example, DellaVigna and Pope (2022) assess the effect of eliminating the consent form and potentially hiding to MTurk participants that they are in a study, and as we discuss in the upcoming section de Quidt et al. (2018) derive and develop procedures for bounding potential experimenter demand effects. Haaland and Roth (2020) show that an obfuscated follow-up can be used to gauge the potential for experimenter demand in survey experiment when comparing information treatments, where participants a week after treatment are resurveyed to assess lasting treatment effects. Also looking at survey experiments, Mummolo and Peterson (2018) explore the effect of (1) directly informing participants of the experimental hypothesis, (2) providing information on the direction of the hypothesis; and (3) paying participants to confirm the hypothesis.[31] Neither Haaland and Roth (2020) or Mummolo and Peterson (2018) finds evidence that results are sensitive to experimenter demand.

Finally, there have been recent efforts to assess the impact of experimenter demand using design-by-correlation, where a survey measure is used to capture a participant's potential sensitivity to experimenter demand, and the measure in turn is used to assess differential treatment effects. For example, Dhar et al.(2018) include measures of the Crowne-Marlowe social desirability scale to assess whether treatment-driven changes in attitudes toward gender equality result from experimenter demand.[32] They find that while participants with high scores show more support for gender equality, this holds for both treatment and control, indicating that participant scores do not drive the treatment effect. Similarly, Alcott and Taubinsky (2015) use the Snyder (1974) "Self-Monitoring Scale" to measure subjects' responsiveness to experimenter demand and find no correlation between individual scores and treatment effects.[33]

While there are many new and clever assessments of experimenter demand, we caution against resorting to design-by-correlation assessments. Not only do we not know that the chosen scale

[31] Both Mummolo and Peterson (2018) and de Quidt et al. (2018) aim to exaggerate demand effects.

[32] The Crowne-Marlowe social desirability scale gathers individual measures on several too-good-to-be-true personality traits (e.g., never intensely disliked anyone, like to gossip, always practice what I preach).

[33] The Snyder scale self-monitoring scale aims to assess an individual's will and ability to modify how they are perceived by others (e.g., I can only argue for ideas which I already believe, I'm not always the person I appear to be. I would probably make a good actor, etc.)

captures sensitivity to experimenter demand, but even if it did, we have no way of assessing whether differential responses to treatment result from experimenter demand. It may well be that a high score correlates with a participant's underlying preferences and that these, rather than a sensitivity to experimenter demand, gives rise to larger treatment effects. In other words, it is difficult to interpret the findings of these approaches, whether or not, it is found that the elicited score significantly interacts with treatment.

4. Bounding demand effects and assessing robustness

The first line of defense against demand effects is always to try to mitigate them, following the advice in the previous section. But sometimes you or a reviewer might have a legitimate concern about unmitigated demand biases (i.e., you are still concerned that ϕ or $E[h|\zeta,\rho]$ are large) and want to assess the robustness of your findings to their influence or provide convincing evidence that any bias must be small. This section explains a bounding approach, developed by de Quidt et al. (2018), that you can follow to allay these concerns.[34]

In essence, the approach here is the opposite of mitigation. The aim is to deliberately *amplify* demand effects in a structured way, to assess their magnitude and *partially identify* the natural actions or treatment effects contained within the bounds.

The bounding approach entails adding positive and negative "demand treatments" to an existing experimental design. Demand treatments are *explicit* signals that seek to directly manipulate participants' beliefs about the direction of the researcher's hypothesis, h. The key idea is that by manipulating these beliefs, we can elicit an upper and lower bound containing the natural action. Positive demand treatments try to push up $E[h|\zeta,\rho]$, while negative treatments try to push it down. We label the actions under such treatments $a^{+}(\zeta,\rho)$ and $a^{-}(\zeta,\rho)$ respectively.

The goal is to find demand treatments such that the natural action lies in-between $a^{+}(\zeta,\rho)$ and $a^{-}(\zeta,\rho)$. We need this to hold irrespective of the initial belief $E[h|\zeta,\rho]$. The model tells us that this is possible if, irrespective of their initial belief, under a positive demand treatment, the participant is persuaded that high actions are more likely desired than low actions ($E[h|\zeta,\rho,\text{positive demand}] \geq 0$), and under a negative demand treatment they are persuaded that

[34] See also the concurrent paper by Mummolo and Peterson (2019) who use a similar approach albeit not expressed in terms of bounds.

low actions are more likely desired than high actions ($E[h|\zeta, \rho, \text{negative demand}] \leq 0$). This is possible if the participant perceives the demand treatments to be more persuasive than the latent demand signal that comes from ζ and ρ. In other words, the demand treatments must be sufficiently convincing that they can reverse the sign of the participant's belief about h.[35]

If that assumption holds, then the bounds on the natural action are:

$$a(\zeta) \in [a^{-}(\zeta, \rho), a^{+}(\zeta, \rho)].$$

Combining bounds on actions we can construct bounds on treatment effects:

$$a(\zeta_1) - a(\zeta_0) \in [a^{-}(\zeta_1, \rho_1) - a^{+}(\zeta_0, \rho_0), a^{+}(\zeta_1, \rho_1) - a^{-}(\zeta_0, \rho_0)]$$

4.1. Choosing appropriate demand treatments

To construct demand treatments, de Quidt et al. (2018) recommend adding additional text to experimental instructions, typically just one extra sentence. Crucially, a demand treatment should send a signal about the experimenter's wishes and nothing else (in particular, we do not want to change the decision-relevant features of the design, ζ). They propose two general-purpose phrasings, but these can be adapted to suit your setting:

Weak: *We expect that participants who are shown these instructions will [work, invest, ...] more/less than they normally would.*

Strong: *You will do us a favor if you [work, invest, ...] more/less than you normally would.*

The sentences are designed to fit the theoretical motivation. In particular, the phrase "more than you normally would" is chosen to explicitly make reference to choices higher or lower than the

[35] Formally, de Quidt et al. (2018) model updating via two signals. The first $h^L(\zeta, \rho) \in \{-1,1\}$ is what the participant infers from the decision-relevant and decision-irrelevant features, and is believed to be true with probability $p^L(\zeta, \rho)$, and the second $h^T \in \{-1,1\}$ is the demand treatment (negative or positive), believed to be true with probability p^T. To obtain valid bounds, we require that $E[h|h^L(\zeta, \rho), h^T = -1] < 0$ and $E[h|h^L(\zeta, \rho), h^T = 1] > 0$, i.e. the posterior following a demand treatment has the same sign as the demand treatment, whatever was $h^L(\zeta, \rho)$. This is possible if the participant perceives the demand treatments to be more persuasive than the latent demand signal $h^L(\zeta, \rho)$, i.e. $p^T \geq p^L(\zeta)$.

natural action.[36] They are carefully phrased so as to avoid deception.[37] The key assumption is that such a statement about expected behavior sends a stronger signal than whatever the participants are inferring explicitly or implicitly from other features of the experimental design. If so, valid bounds are obtained.

When choosing a demand treatment you face a tradeoff. Stronger language conveys a stronger signal about experimenter demand and so is more likely to give valid bounds—it is more robust if the latent demand bias is, itself, large. But this comes at the cost of generally giving rise to *wider* bounds because behavior responds more, and is therefore costly for power.

We believe that in most cases, when bounding an action or treatment effect using a pair of demand treatments, "weak" language will be more than sufficient for valid bounds. Typical experiments are unlikely to send implicit signals even as strong as "we expect that…" and therefore this language should be more than sufficient for bounding purposes. Strong treatments can still be useful though. If one can show that even strong treatments give narrow bounds, we should be highly confident that demand is a non-issue in this setting—see the evidence from Winichakul et al. (2024) discussed below. Further strong treatments can be useful for structural estimation of the demand effect (see de Quidt et al., 2018), and may be a better choice when using the approach outlined in de Quidt (2024) to bound a treatment effect using a single demand treatment (see below).

De Quidt et al. (2018) and Winichakul et al. (2024) implement the proposed bounding approach in a large number of applications using online and laboratory participant pools. We discuss those below, along with other applications, in Section 5. A broad summary of these findings is that the approach "works" in the sense that demand treatments mostly shift behavior in the anticipated direction, with limited defier behavior, and the bounds obtained look reasonable on various dimensions.

[36] One could imagine other types of signal, such as "will work **a lot/a little**." The downside of such an approach is that the direction relative to the natural action is sometimes ambiguous. For instance, a participant who planned to exert zero effort but who is told they are expected to work a little has arguably received a positive, rather than a negative, demand treatment.

[37] The idea here is that stating a general expectation about behavior ("we expect participants to work more than they normally would") cannot be true for both the positive and negative demand treatments. But the sentence above, by self-referentially invoking the instructions that are being read, is truthful because we do believe that this sentence itself will change behavior. See de Quidt et al. (2018) for discussion on this point.

For the remainder of this section we highlight a couple of important considerations when using the bounding approach in practice.

4.2. Reducing the cost of data collection

An important cost of the bounding approach is that it entails additional treatment arms. For example, to obtain bounds on a treatment effect in a simple two-arm between-subjects experiment entails four additional treatment arms (one positive and one negative demand treatment per main treatment arm). That is obviously quite costly.

To reduce data collection costs there are three main shortcuts that can be used.

4.2.1. Lower-bounding treatment effects

Normally the researcher (or reviewer) has a guess or concern about demand effects in a particular direction. For instance, consider a survey experiment where participants report their prior on, say, crime rates in their city, and then are told that the true crime rate is much lower. The concern would normally be that participants guess they are supposed to update in the direction of the new information, exaggerating the true treatment effect by exaggerating by how much their beliefs have changed. In that case we are probably mostly interested in a *lower bound* on the treatment effect which can be obtained by applying a positive demand treatment to the control group, and a negative demand treatment to the treatment group. That cuts the number of additional treatments required in half.

4.2.2. Within-subject demand bounds

De Quidt et al. (2018) show that in some cases it is possible to use demand treatments in a within-subjects design. The basic idea is that for every participant we would first elicit their action under the standard treatment conditions. Then, at the end of the study, the task is repeated but with a demand treatment added. There are a number of significant advantages to this approach. First, it obviates the need to recruit additional participants for bounds estimation since these can be estimated using the main sample respondents. Second, it increases the power in estimating bounds by exploiting the within-subject change in behavior between neutral and demand treatments. Third, it enables classifying participants into types: those who do not respond to demand, those who comply, and those who defy the experimenter's wishes. The bounds can then be easily corrected for defier behavior. The main downside of using within-subject variation is that it makes the demand-treatment variation transparent to participants, which as discussed in the previous section might amplify the demand effects that it measures. Fortunately the evidence we have suggests this is not too problematic in practice: De Quidt et al. (2018) compared the bounds obtained from within- and between-subject application of demand treatments, applied to dictator-game giving and risky choice. The estimated bounds were quantitatively very similar in both approaches, suggesting that the within-subject approach worked well.

4.2.3. Using a single strong demand treatment to bound a treatment effect

De Quidt et al.(2018)'s bounding approach always requires two demand treatments to construct a bound (one positive, one negative). De Quidt (2024) asks under what conditions we can accomplish similar goals using a single demand treatment. The setting of interest is as follows. You find a treatment effect, say, positive: $a^L(\zeta_1, \rho_1) - a^L(\zeta_0, \rho_0) > 0$. You want to check if this could be "explained by a demand bias," i.e., can you reject the null $a^L(\zeta_1) - a^L(\zeta_0) = 0$. Intuitively, one might check whether it is possible to reproduce the observed effect by adding positive demand to the control group (or negative demand to the treatment group), i.e., reject the null if we also find $a^L(\zeta_1, \rho_1) - a^+(\zeta_0, \rho_0) > 0$. de Quidt (2024) shows that this does not work under the identification assumptions of de Quidt et al. (2018), but it is a valid approach under stronger assumptions. In particular a "strong" positive demand treatment applied to the control group might reasonably be expected to satisfy this stronger condition. Applying a single demand treatment to one arm is usually logistically easier than the full positive/negative treatment pair approach.

5. Application of the methods

In this section, we will discuss applications of the de Quidt et al. (2018) methods for bounding experimenter demand effects. We outline the domains where demand effects have been studied and discuss evidence on the response to experimenter demand. The news is mostly positive. In most settings, the "weak" form of explicit demand has only a minor effect. The more-extreme "strong" form of demand has moderate effects in the de Quidt et al. (2018) MTurk sample. However much tighter bounds are uncovered using similar populations (MTurk and Prolific) in Winichakul et al. (2024) suggesting more minor effects. Moreover, even tighter bounds are found when using a standard laboratory population of undergraduate students.

Table 1: Summary of Existing Applications

Paper	Tasks	Demand	Direction	Populations
Uncertainty:				
De Quidt et al.(2018)	Investment game	Strong & weak	Both	MTurk, Online panel (U.S.)
Winichakul et al.(2024)	Lottery pricing	Strong	Both	Lab (U.S.), MTurk, Prolific
Incentives & Information:				
De Quidt et al.(2018)	Real effort	Weak	Both	MTurk
Roth & Wohlfart (2020)	Expectation formation	Weak	Positive	MTurk
Gao & Tavoni (2024)	Information intervention	Weak	Both	Online panel (China)
Mummolo & Peterson (2019)	Political science surveys	~Weak & ~Strong	Both	MTurk, Qualtrics
Time Preferences:				
De Quidt et al. (2018)	Convex time budget	Strong & weak	Both	MTurk
Winichakul et al. (2024)	Convex time budget	Strong	Both	Lab (U.S.), MTurk, Prolific
Altruism & Morality:				
De Quidt et al. (2018)	Dictator & ultimatum game	Strong & weak	Both	MTurk, Online panel (U.S.)
De Quidt et al. (2018)	Lying game	Strong & weak	Both	MTurk
Winichakul et al. (2024)	Charitable giving	Strong	Both	Lab (U.S.), MTurk, Prolific
Haushofer et al. (2023)	Dictator game	Weak	Positive	Lab (Kenya)

Below we break down what decisions have been studied across four broadly defined domains, summarized in **Table 1**: (i) Preferences over uncertainty; (ii) the response to incentives and information; (iii) time preferences; and (iv) altruism and morality. Our goal is to direct readers to the relevant papers that examine demand treatments and provide an understanding of the magnitude of the bounds on demand effects in each setting. We refer to these magnitudes as demand ranges, presented as z-scores by measuring the difference across the decisions under positive and negative demand ($a^{+}(\zeta,\rho) - a^{-}(\zeta,\rho)$), expressed in multiples of the latent decision's standard deviation.[38] We will then briefly summarize evidence of differences in response across

[38] De Quidt et al. (2018) call this object "sensitivity."

populations. Finally, we will report on research exploring demand effects on inference over treatment effects.

To assist in summarizing the experimental results, Table 1 reports for each domain the papers that have used the de Quidt et al. (2018) procedure to bound decision estimates, indicating the type of demand induced, the directions of the demand treatments used (positive, negative, both), and the populations studied.

5.1. Uncertainty

Two studies have assessed the impact of experimenter demand on decisions under uncertainty. In the first, de Quidt et al. (2018) examine the risky investment game by Gneezy and Potters (1997) when participants have a $1 endowment and are asked what portion of their endowment they want to invest in a risky project (yielding a 40 percent chance of tripling their investment). The environment is studied with and without ambiguity (implemented by not informing participants of the 40 percent chance),[39] and using both weak (*"We expect that participants who are shown these instructions will invest more in the project than they normally would."*) and strong demand (*"You will do us a favor if you invest more in the project than you normally would."*) in the positive and negative directions (increasing and decreasing the amount invested, respectively). The demand range under strong demand is moderate at just over half a standard deviation (0.53σ). However, when examining the weak-demand treatments there is only a marginally significant effect (0.16σ). A similar story is found for the ambiguous investment task, where the demand range under strong demand is moderate (0.46σ), and the effect of the weak demand treatments is small, again only a marginally significant effect (0.17σ).

Winichakul et al. (2024) use the strong form of experimenter demand to examine the WTP and WTA for two lotteries over a $10 prize: one with a low chance of winning (10%) and one with a high chance of winning (90%). Their study looks at three subject populations: an undergraduate laboratory population, MTurk and Prolific.[40] The response to strong experiment demand is small for the lab population (an average range of 0.11σ across the four decisions) with no consistent

[39] The risky decision was framed as choosing an amount to invest in a project with 40% chance of success. The ambiguous decision was framed as choosing a ball from an urn with unknown mix of two colors, the participants picked a color and an amount to invest.

[40] For the MTurk and Prolific studies, incentives are scaled down to a lottery over a $1 prize instead of $10.

direction of the demand-treatment effects. In all four decisions, the directional effect of strong demand on the decision estimate is independent of demand being positive or negative. Thus, for the lab population there is little systematic response to experimenter demand in the uncertainty domain. The effects are also small when using either MTurk and Prolific populations (an average range of 0.21σ for both populations). Overall, the results suggest little potential impact of experimenter demand when examining decisions under uncertainty.

5.2. Incentives and information

Understanding how incentives and information affect behavior is a key task in many economic analyses. De Quidt et al. (2018) study how demand affects real effort decisions using the DellaVigna and Pope (2018) real-effort task, where participants need to alternate keystrokes between *"a"* and *"b,"* yielding an experimental point for each alternation. There are two treatments: one with no monetary reward and one with a \$0.01 reward for every 100 experimental points (see section 5.6 for a comparison). Demand is induced in both directions using both weak- and strong-demand language. Strong demand has a substantial effect in the no-incentive treatment (0.78σ) but a much smaller effect when participants have monetary incentives to perform (0.2σ). This supports the idea that incentives diminish the impact of demand (reducing the intensity of $\phi(\zeta, \rho)$ relative to the intended incentives $v(a, \zeta)$ the experimenter is studying). However, under the weak-demand treatment, there is no difference in response to demand when effort is and is not incentivized, indeed the response to demand is not significant, and the sign of the effect is independent of the demand treatment being positive or negative.

Demand treatments have also been used to bound the effects of demand in information interventions. Roth and Wohlfart (2020) examine how information on an expert's probabilistic forecast of a future recession impacts personal expectations and plans over consumption and investment. Demand effects are assessed in a single demand treatment where weak positive demand is used (in comparison to a no-demand control in an additional set of questions). An insignificant effect from the demand treatment is used to argue that demand is not a substantive driver of the results.[41] Gao & Tavoni (2024) examine how information on the benefits of an environmentally friendly lightbulb (either monetary or environmental benefit information) affects

[41] See our discussion in section 4 for how to interpret bounds from single demand treatments.

lightbulb purchasing decisions over the next 10 months. Gao & Tavoni (2024) implement negative and positive weak demand treatments across the two information conditions and the control.[42] The effects from weak demand are small in all three treatments (a demand range of 0.15σ for the control, and with inconsistent directional effects in the environmental and monetary treatments). The authors conclude that demand effects are small in their two information treatments.

Demand treatments have also been deployed in information survey experiments in political science. Mummolo and Peterson (2019) outline several demand treatments to examine survey responses though with a different methodology from de Quidt et al. (2018). They replicate five political science studies on MTurk and Qualtrics samples using three distinct methods for inducing demand. The methods vary from providing a hint of the experimenter's hypothesis or an explicit description of it, a stronger version of the weak demand treatment where they provide a statement on the expected direction of an effect, and an incentivized version of strong demand where they directly pay participants for helping the researcher confirm their hypothesis (cf. Table 2 in Mummolo and Peterson, 2019). While explicit demand ranges are not provided, the effects from demand are small.

5.3. Time preferences

Many experimental papers measure preferences over the timing of streams of income and consumption (see Cohen et al., 2020, for a survey). The earliest experiments tended to treat money and consumption as the same thing, offering participants a "money earlier or later" choice, varying the monetary amounts at the earlier and later points to identify a temporal preference. However, a critique of these experiments was that money is a storable good, and that the models being tested are over delayed consumption. In response to this, papers added different reward media such as food (Reuben et al., 2010) or switched to a costly real-effort task (Augenblick et al., 2015). However, in terms of demand treatments, the only experiments with induced demand are over convex-budget sets (cf. Andreoni and Sprenger, 2012), a paradigm where the core result is a null effect over present bias for monetary bundles.

Both de Quidt et al. (2018) and Winichakul et al. (2024) examine the impact of experimenter demand when using convex monetary budget sets in the Andreoni and Sprenger (2012) paradigm.

[42] Here the population is recruited from a large online platform in China, with a nationally representative sample (on age and gender, non-representative on income and education).

Agents are given an endowment of money on a date t and can move part of the endowment to a date a week later, $t + 7$,[43] where anything moved to the later date earns 20 percent interest. Both papers examine the case of now versus later with t=0, where Winichakul et al. (2024) additionally examine the case of $t = 1$ (tomorrow versus a-week-from-tomorrow).[44]

The demand ranges are for the most part small. When exploring strong experimenter demand Winichakul et al. (2024) finds for the lab population inconclusive evidence of any substantial demand effect across the two demand treatments (0.05σ in one treatment and mis-signed effects in the other) and finds small demand-effect ranges for the Prolific and MTurk samples (approximately 0.18σ). Using an Mturk sample de Quidt et al. (2018) find a moderate demand-effect range under strong demand (0.35σ) and a smaller and insignificant range under weak demand (0.12σ).

5.4. Altruism and morality

De Quidt et al. (2018) examines three decision tasks in this domain. Giving to others in the dictator and ultimatum games (including the responder decision) and making transfers in a trust game (both roles). Outside of the strategic setting, they also look at a lying game ($0.10 for every head in ten self-reported coin flips). The demand-effect range is found to be large under strong demand (0.56-0.69σ) and smaller under weak demand (often insignificantly different from zero, 0.04-0.24σ).[45] Across domains and decision tasks, only the response under the no-incentive real-effort task has a similarly large demand effect.

Winichakul et al. (2024)'s assessment of tradeoff between the self and others in their lab sample also reveals responsiveness to experimenter demand. They examine a dictator game between the participant and a charity, varying the price of giving with either no match or a one-for-one match. They find that strong demand gives rise to a more consistent response to experimenter demand for

[43] In de Quidt et al. (2018), the participant is allowed to move the entire amount to the future date.

[44] For the online populations in both papers, the initial time t allocations are $1, while for the lab population in Winichakul et al. the initial allocation is $10.

[45] Larger effect sizes are found for the conditional second-mover responses, where strong demand moves the total range by 1.06σ in the trust game (0.75σ in the ultimatum game). This is clearly an extreme effect, where even the weak demand treatments have a moderate demand effect range (0.29σ and 0.28σ in the trust and ultimatum games, respectively).

the lab population (0.21-0.25σ), and find similarly small demand ranges for the MTurk and Prolific populations (0.20-0.22σ), and at a similar size to the uncertainty tasks.

5.5. Population comparisons

Winichakul et al. (2024) examine the effect of strong demand over the same set of tasks when using samples from three different populations: an undergraduate lab sample (with stakes of $10-$20 involved in the decisions) and online samples from Prolific and MTurk (with stakes of $1-$2 in the decisions). What do we learn about the response to experimenter demand across these populations?

For their lab sample, they fail to find evidence that the strong demand treatments move decisions. Looking at eight decisions across both positive and negative demand treatments, and looking at the sign of the effect (relative to the latent no-demand control) they fail to reject independence. That is, the directional movements of the strong demand treatments are no different from a coin flip ($p = 0.304$ from a one-sided Fisher exact test). Further, the average range of the lab demand effects are small (0.10σ) relative to natural variation in the tasks across participants. This null effect from demand treatment reinforces the idea that even with extreme demand (i.e., the strong form), the uncovered bounds are so small in statistical terms that they are unlikely to affect qualitative inferences.

In contrast, for both Prolific and MTurk, when using the larger sample sizes common for online studies, the detected responses to positive or negative experimenter demand are significantly different from zero, and we can reject independence in favor of demand pushing the decision estimate in the intended directions ($p = 0.003$ and $p = 0.020$, respectively). Still, even with strong demand the estimated average demand ranges are small (0.20σ and 0.21σ, respectively).

The two studies can also be used to explore whether certain subsets of the populations are more sensitive to experimenter demand. However, we find no systematic or substantial differences. For example, de Quidt et al. (2018) find a slight increase in the demand range for female participants (0.15σ higher than men), while Winichakul et al. (2024) find essentially identical demand ranges for men and women in the evaluated tasks.[46]

[46] Similar null effects in a comparison of the demand ranges are found for: (i) race (caucasian/non-caucasian), (ii) income (above or below US $70k); (iii) education (above/at or below HS). However, they do find an effect from (iv)

5.6. Qualitative inferences

While the above focuses on the demand ranges for specific isolated decisions, most experimental work examines qualitative differences between a treatment and a control (Kessler and Vesterlund, 2015). It is possible that a comparison of two measures can amplify the total effect of demand, where a critical reviewer may argue that demand effects in the treatment are positive while they are negative in a control, resulting in a false positive. Such hypotheticals are extreme, and as we describe above the onus here would be on the critical reviewer to make a clear argument for how such differential demand has been generated. Compelling arguments for differential demand will be particularly challenging in between-subject designs where participants are unaware of their assigned condition, especially if instructions have only minimal changes. However, assuming a reviewer can reasonably argue that demand will increase the level in one condition and decrease it in another, what can be done? Well, a bounding approach is again possible, where experimenters can assess the worst-case scenario for inference by imposing differential strong demands for treatment and control, and observing the effects on inference.

The main aim of Winichakul et al. (2024) is to examine the effects of demand over such qualitative inferences. Their selected tasks are treatment-control decision pairs across four canonical behavioral comparative statics: (i) probability weighting; (ii) the endowment effect; (iii) present bias; and (iv) a price response to charitable giving. While three of these are expected to exhibit a significant response, in their present-bias task pair a null response is expected.[47]

For their lab results, even strong differential demand across treatment and control is unable to change any of the four qualitative inferences. Strong demand in opposed directions does not negate the positive findings on probability weighting, the endowment effect, and charitable giving. Nor are they able to generate a false positive on the present-bias result. Very similar findings also emerge from their online populations when considering the expected positive results, where strong demand does not qualitatively alter the effects in either of the three results. However, they do find that strong demand can generate a significant finding in the online populations when the expected

being older (above 33, increases range by 0.11σ, p=0.010); and (v) migrant status (not being born in the US increases the range by 0.10σ, p=0.081).

[47] The task pair for the present-bias task uses the convex-budget set over monetary payments at different points in time from Andreoni and Sprenger (2012, see above for details). For monetary payments, their paper indicates a null response when comparing the now vs future and tomorrow vs future decisions.

result is a knife-edge null (the present-bias hypothesis). The reasons for this are the combination of two factors: (i) slightly larger effects of demand in the online populations, more consistently in the intended direction than the lab; and (ii) larger samples in the online populations. That is, the larger samples that researchers tend to gather in online settings because observations are both cheaper and easier to collect (cf. Rigotti et al. 2023), create a better environment for false positives driven by demand. That is, even the small demand ranges detected in their experiments can become significant for a true null with a large enough sample.

In the incentives domain, de Quidt et al. (2018) also examine the effects of demand on inference, here with an MTurk sample. Examining a real-effort task pair (piece-rate payment versus no incentive at all in the '*a*' to '*b*' keystroke alternation task) they examine whether incentives affect effort/production. They do this both in the latent decision with no explicit demand, as well as with both the strong and weak forms of demand. Mirroring the larger demand range for the no-incentive decision, they find large economic effects under strong demand, where the underlying treatment-effect from incentives moves from a 40 percent effect with no demand, to an 11 percent increase when minimized with differential demand (a 93 percent increase if using demand to maximize the effect). While these economic movements are large under the more-extreme strong demand, there is no shift in the statistical inferences, where the qualitative conclusion in all comparisons is the intuitive conclusion that a piece-rate increases effort.

Much smaller economic effects are found with the weak form of demand. Attempting to minimize (maximize) the real-effort incentive effect with differential weak demand essentially replicates the latent measure, with a 41 (42) percent increase in effort. Just as the strong-demand treatment, there is again no effect on the qualitative inferences.[48] Similarly, Gao and Tavoni's (2024) weak-demand implementations also allow for an examination of differential demand across the information treatment vs control comparison. Using their replication data, weak demand does not alter the conclusions on their stronger monetary-benefit intervention, though differential weak demand leads to an insignificant estimate on their smaller environmental-benefit intervention.

[48] While unexamined in their paper, two further comparative-statics are possible in the de Quidt et al. replication data (comparing uncertainty and ambiguity, and dictator- and ultimatum-game giving). The data here again indicates relatively minor inferential effects under the weak version of demand.

The main findings across the qualitative inference comparisons with explicit demand treatments broadly indicate that demand-effects over comparative statics are second-order. The more reasonable weak-demand treatments mostly have small effects, where shifts in statistical inference only occur for either more-marginal economic findings, or for precise nulls with much larger samples (typically on online populations). In particular, the evidence from Winichakul et al. (2024) in the laboratory samples (with typical lab stakes and sample sizes) indicates no qualitative effects from even the extreme strong form of demand.

6. Summary of recommendations and open questions

Our recommendations can be summarized as follows:

1. Adopt best practices wherever possible. In many cases this will be sufficient to allay concerns about demand bias.
2. Where direct evidence on demand effects exists from prior studies on bounding, this may help support the case that your design is robust, but we caution against over-inferring from this evidence as your setting and design features may differ significantly.
3. Where concerns remain, or where one wants to convincingly demonstrate robustness, adopt a bounding approach as outlined above. Where to save on implementation costs you may want to consider one-sided, within-subject, or "single-treatment" bounds.

It is important to emphasize that by their nature demand bounds on the decision estimates are designed to *exceed* any bias due to demand. So, when bounding does not change your findings, the conclusion is clear, but when the bounds show substantial sensitivity, it does not mean that your results are driven by experimenter demand, but rather that more investigation is warranted. A promising check could be to measure whether it is plausible that participants' beliefs about the research hypothesis differ significantly between treatments. This could be done by eliciting beliefs about the research hypothesis from third parties that have been shown the same information as the study participants. Alternatively you may wish to modify your design to assess whether your results are robust in a design where potential demand confounds are removed.

As we have argued throughout, the profession takes concerns about experimenter demand seriously. In our view, the guidance above provides a robust tool kit to allay these concerns. We are not arguing that all experimental results should simply be taken at face value. Critical review of all threats to relevance and identification is central to the scientific process, but demand biases in well-designed experiments should rarely be prominent among those.

There are some areas where more evidence would be welcome. Bounding methods have been applied to a range of canonical games and subject pools, but for the most part have not been used to systematically study the relative importance of different design features that we have highlighted. In cases where following all the design recommendations is costly, it would be valuable to have a more precise sense of the tradeoffs involved.

7. References

Abbink, Klaus, and Abdolkarim Sadrieh. "The pleasure of being nasty." *Economics Letters* 105, no. 3 (2009): 306-308.

Allcott, Hunt, and Dmitry Taubinsky. "Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market." *American Economic Review* 105, no. 8 (2015): 2501-2538.

Amir, Ofra, David G. Rand, and Ya'akov Kobi Gal. "Economic games on the internet: The effect of $1 stakes." *PloS One* 7, no. 2 (2012): e31461.

Andreoni, James, and Charles Sprenger. "Estimating time preferences from convex budgets." *American Economic Review* 102, no. 7 (2012): 3333-56.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar. "Large stakes and big mistakes." *Review of Economic Studies* 76, no. 2 (2009): 451-469.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. "Working over time: Dynamic inconsistency in real effort tasks." *Quarterly Journal of Economics* 130, no. 3 (2015): 1067-1115.

Babcock, Linda, Maria P. Recalde, Lise Vesterlund, and Laurie Weingart. "Gender differences in accepting and receiving requests for tasks with low promotability." *American Economic Review* 107, no. 3 (2017): 714-747.

Barmettler, Franziska, Ernst Fehr, and Christian Zehnder. "Big experimenter is watching you! Anonymity and prosocial behavior in the laboratory." *Games and Economic Behavior* 75, no. 1 (2012): 17-34.

Bordalo, Pedro, Katherine Baldiga Coffman, Nicola Gennaioli, and Andrei Shleifer. "Beliefs about Gender." *American Economic Review* 109, no. 3 (March 2019): 739–773

Bracha, Anat, and Lise Vesterlund. "Mixed signals: Charity reporting when donations signal generosity and income." *Games and Economic Behavior* 104 (2017): 24-42.

Brandts, Jordi, and Gary Charness. "The strategy versus the direct-response method: a first survey of experimental comparisons." *Experimental Economics* 14 (2011): 375-398.

Burks, Stephen V., Jeffrey P. Carpenter, and Eric Verhoogen. "Playing both roles in the trust game." *Journal of Economic Behavior & Organization* 51, no. 2 (2003): 195-216.

Camerer, Colin F. “Behavioral Game Theory: Experiments in Strategic Interaction.” Princeton University Press, 2003.

Camerer, Colin F. “The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List.” *Handbook of Experimental Economic Methodology* 18 (2015).

Camerer, Colin F., and Robin M. Hogarth. "The effects of financial incentives in experiments: A review and capital-labor-production framework." *Journal of Risk and Uncertainty* 19 (1999): 7-42.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn. "Experimental methods: Between-subject and within-subject design." *Journal of Economic Behavior & Organization* 81, no. 1 (2012): 1-8.

Cohen, Jonathan, Keith Marzilli Ericson, David Laibson, and John Myles White. "Measuring time preferences." *Journal of Economic Literature* 58, no. 2 (2020): 299-347.

DellaVigna, Stefano, and Devin Pope. "What motivates effort? Evidence and expert forecasts." *Review of Economic Studies* 85, no. 2 (2018): 1029-69.

DellaVigna, Stefano, and Devin Pope. "Stability of experimental results: Forecasts and evidence." *American Economic Journal: Microeconomics* 14, no. 3 (2022): 889-925.

De Quidt Jonathan. "Low-cost, One-sided Bounds for Experimenter Demand Effects." *Mimeo* (2024).

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. "Measuring and bounding experimenter demand." *American Economic Review* 108, no. 11 (2018): 3266-3302.

De Quidt, Jonathan, Lise Vesterlund, and Alistair J. Wilson. "Experimenter demand effects." in *Handbook of research methods and applications in experimental economics*, pp. 384-400. Edward Elgar Publishing, 2019.

Dhar, Diva, Tarun Jain, and Seema Jayachandran. "Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in India." *American Economic Review* 112, no. 3 (2022): 899-927.

Dreber, Anna, Tore Ellingsen, Magnus Johannesson, and David G. Rand. "Do people care about social context? Framing effects in dictator games." *Experimental Economics* 16 (2013): 349-371.

Ellingsen, Tore, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. "Social framing effects: Preferences or beliefs*?." Games and Economic Behavior* 76, no. 1 (2012): 117-130.

Echenique, Federico, Alistair J. Wilson, and Leeat Yariv. "Clearinghouses for two-sided matching: An experimental study." *Quantitative Economics* 7, no. 2 (2016): 449-482.

Enke, Benjamin, Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen van de Ven. "Cognitive Biases: Mistakes or Missing Stakes?." *The Review of Economics and Statistics* 105, no.4 (2023): 818–832.

Fischbacher, Urs, and Franziska Föllmi-Heusi. "Lies in disguise—an experimental study on cheating." *Journal of the European Economic Association* 11, no. 3 (2013): 525-547.

Gao, Yu, and Massimo Tavoni. "Forget-me-not: The persistent effect of information provision for adopting climate-friendly goods." *Management Science* 70, no. 7 (2024): 4480-4501.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. "When and why incentives (don't) work to modify behavior*." Journal of Economic Perspectives* 25, no. 4 (2011): 191-210.

Gneezy, Uri, and Jan Potters. "An experiment on risk taking and evaluation periods." *Quarterly Journal of Economics* 112, no. 2 (1997): 631-645.

Harrison, Glenn, W., and John A. List. "Field Experiments." *Journal of Economic Literature* 42, no 4 (2004): 1009–55.

Haushofer, Johannes, Sara Lowes, Abednego Musau, David Ndetei, Nathan Nunn, Moritz Poll, and Nancy Qian. "Stress, ethnicity, and prosocial behavior." *Journal of Political Economy Microeconomics* 1, no. 2 (2023): 225-69.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith. "Preferences, property rights, and anonymity in bargaining games." *Games and Economic Behavior* 7, no. 3 (1994): 346-380.

Niederle, Muriel, and Lise Vesterlund. "Gender and Competition." *Annual Review of Economics* 3, no. 1 (2011): 601-630.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. "Experimental tests of the endowment effect and the Coase theorem." *Journal of Political Economy* 98, no. 6 (1990): 1325-1348.

Karlan, Dean S., and Jonathan Zinman. "List randomization for sensitive behavior: An application for measuring use of loan proceeds." *Journal of Development Economics* 98, no. 1 (2012): 71-75.

Kay, Aaron C., and Lee Ross. "The perceptual push: The interplay of implicit cues and explicit situational construals on behavioral intentions in the Prisoner's Dilemma." *Journal of Experimental Social Psychology* 39, no. 6 (2003): 634-643.

Kessler, Judd, and Lise Vesterlund. "The external validity of laboratory experiments: The misleading emphasis on quantitative effects." *Handbook of Experimental Economic Methodology* 18 (2015): 392-405.

Lambdin, Charles, and Victoria A. Shaffer. "Are within-subjects designs transparent?." *Judgment and Decision Making* 4, no. 7 (2009): 554-566.

Levati, Maria Vittoria, Topi Miettinen, and Birendra Rai. "Context and interpretation in laboratory experiments: The case of reciprocity." *Journal of Economic Psychology* 32, no. 5 (2011): 846-856.

Liberman, Varda, Steven M. Samuels, and Lee Ross. "The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves." *Personality and Social Psychology Bulletin* 30, no. 9 (2004): 1175-1185.

List, John A., Robert P. Berrens, Alok K. Bohara, and Joe Kerkvliet. "Examining the Role of Social Isolation on Stated Preferences." *The American Economic Review* 94, no. 3 (2004): 741–52

Loewenstein, George. "Experimental economics from the vantage-point of behavioural economics." The *Economic Journal* 109, no. 453 (1999): 25-34.

Muller, Laurent, Martin Sefton, Richard Steinberg, and Lise Vesterlund. “Strategic Behavior and Learning in Repeated Voluntary Contribution Experiments.” *Journal of Economic Behavior & Organization* 67, no. 3 (2008): 782–93.

Mummolo, Jonathan, and Erik Peterson. "Demand effects in survey experiments: An empirical assessment." *American Political Science Review* 113, no. 2 (2019): 517-29.

Orne, Martin T. "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications." *American Psychologist* 17, no 11 (1982): 776-83.

Ortmann, Andreas., "Field Experiments in Economics: Some Methodological Caveats", Harrison, G.W, Carpenter, J, and List, J.A. (Ed.) *Field Experiments in Economics* (*Research in Experimental Economics, Vol. 10*), Emerald Group Publishing Limited, Leeds, (2005): 51-70

Pierce A. H. “The Subconscious Again.” *Journal of Philosophy, Psychology and Scientific Methods* 5, no. 10 (1908)

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. "Time discounting for primary and monetary rewards." *Economics Letters* 106, no. 2 (2010): 125-127.

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

Roth, Alvin E. "The early history of experimental economics." *Journal of the History of Economic Thought* 15, no. 2 (1993): 184-209.

Roux, Catherine, and Christian Thöni. "Do control questions influence behavior in experiments?" *Experimental Economics* 18 (2015): 185-194.

Shafir, Eldar. "Choosing versus rejecting: Why some options are both better and worse than others." *Memory & Cognition* 21, no. 4 (1993): 546-56.

Snyder, Mark. "Self-monitoring of expressive behavior." *Journal of Personality and Social Psychology* 30, no. 4 (1974): 526-37.

Tversky, Amos, and Daniel Kahneman. "The framing of decisions and the psychology of choice." *Science* 211, no. 4481 (1981): 453-458.

Tversky, Amos, and Daniel Kahneman. "The framing of decisions and the evaluation of prospects." In *Studies in Logic and the Foundations of Mathematics*, vol. 114, pp. 503-520. Elsevier, 1986.

Zizzo, Daniel John. "Experimenter demand effects in economic experiments." *Experimental Economics* 13 (2010): 75-98.