American Economic Review 2018, 108(11): 3266–3302
https://doi.org/10.1257/aer.20171330

Measuring and Bounding Experimenter Demand[†]

By Jonathan de Quidt, Johannes Haushofer, and Christopher Roth*

We propose a technique for assessing robustness to demand effects of findings from experiments and surveys. The core idea is that by deliberately inducing demand in a structured way we can bound its influence. We present a model in which participants respond to their beliefs about the researcher's objectives. Bounds are obtained by manipulating those beliefs with "demand treatments." We apply the method to 11 classic tasks, and estimate bounds averaging 0.13 standard deviations, suggesting that typical demand effects are probably modest. We also show how to compute demand-robust treatment effects and how to structurally estimate the model. (*JEL* C83, C90, D83, D91)

A basic concern in experimental work with human participants is that, knowing that they are being experimented on, the participants may change their behavior. Specifically, participants may try to infer the experimenter's objective from their treatment, and then act accordingly (Orne 1962; Rosenthal 1966; Zizzo 2010). For instance, participants who believe the researcher wants to show that people free-ride in public good games might play more selfishly than they otherwise would. Thus, instead of measuring the participant's "natural" choice, the data are biased by an unobservable *experimenter demand effect*. Demand effects pose a threat to external validity, because participants would make different choices if the experimenter were absent. They can affect estimates of average behavior and treatment effects, and have been raised as a concern in the context of lab experiments (List et al. 2004; List 2006; Levitt and List 2007), field experiments (Allcott and Taubinsky 2015; Dupas

*de Quidt: Institute for International Economic Studies, Stockholm University, 106 91 Stockholm, Sweden, and CESifo (email: jonathan.dequidt@iies.su.se); Haushofer: Department of Psychology and Woodrow Wilson School of Public and International Affairs, Princeton University, 427 Peretsman-Scully Hall, Princeton, NJ 08540, NBER, and Busara Center for Behavioral Economics (email: haushofer@princeton.edu); Roth: Department of Economics, University of Oxford, Keble College, Parks Road, Oxford, OX1 3PG, United Kingdom, and CSAE (email: christopher.roth@economics.ox.ac.uk). This paper was accepted to the *AER* under the guidance of Stefano DellaVigna, Coeditor. We thank three anonymous referees for useful comments. We are grateful to Johannes Abeler, Dan Benjamin, Stefano Caria, Rachel Cassidy, Tom Cunningham, Elwyn Davies, Armin Falk, Thomas Graeber, Don Green, Alexis Grigorieff, Johannes Hermle, George Loewenstein, Simon Quinn, Matthew Rabin, Gautam Rao, Bertil Tungodden, and Liad Weiss for comments. We thank Stefano DellaVigna, Lukas Kiessling, and Devin Pope for sharing code. Moreover, we thank seminar participants at Bergen, Berlin, Bonn, Busara Center for Behavioral Economics, CESifo, Cologne, IIES, LSE, Lund, Melbourne, Oxford, SITE, Stockholm, Sussex, Warwick, Wharton, and Wisconsin. We thank Justin Abraham for excellent research assistance. de Quidt acknowledges financial support from Handelsbanken's Research Foundations, grant no. B2014-0460:1. The experiments in this paper were funded by Princeton University. IRB approval was obtained at Princeton University and the University of Oxford. The experiments were pre-registered as trial 1248 on the American Economic Association RCT Registry, available at https://www.socialscienceregistry.org/trials/1248. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

[†]Go to https://doi.org/10.1257/aer.20171330 to visit the article page for additional materials and author disclosure statement(s).

and Miguel 2017; Al-Ubaydli et al. 2017), and survey responses (Clark and Schober 1992; Bertrand and Mullainathan 2001).[1]

The core idea of our paper is that one can construct plausible bounds on demand-free behavior and treatment effects by deliberately *inducing* experimenter demand and measuring its influence. For example, in a dictator game, we explicitly tell some participants that we expect they will give more than they normally would, while others are told we expect they will give less. Under the assumption that any underlying demand effect is less extreme than our manipulations (in a sense that we will formalize), choices under these instructions give upper and lower bounds on demand-free behavior, and by combining bounds from different experimental treatments we can estimate bounds on treatment effects.

We begin with a simple Bayesian model of decision-making that motivates our approach. In our model, an experiment defines a mapping from actions to utility. The experimenter is only interested in measuring the "natural" action (or changes in that action) that maximizes the participant's utility as derived from the experimental payoffs. However, the participant is also motivated to take actions that conform to the experimenter's research objectives. He infers those objectives from the design features, and distorts his action, biasing the results. Our demand treatments manipulate those beliefs to identify an interval containing the natural action. We remain agnostic about *why* the participant wishes to please the experimenter; motives could include altruism, a desire to conform, a misguided attempt to contribute to science, or an expectation of reciprocity from the experimenter.

We provide an extensive set of applications of the method. We conduct seven online experiments with approximately 19,000 participants in total, in which we construct bounds on demand-free behavior for eleven canonical tasks.[2] We employ two different types of demand treatments. "Weak" demand treatments signal an experimental hypothesis to our respondents: we tell them "We expect that participants who are shown these instructions will [work, invest, …] more/less than they normally would." We believe that these treatments are likely to be more informative than implicit signals about demand in typical studies, so in our view these bounds will be sufficient for most applications. Our "strong" demand treatments go further, telling participants "You will do us a favor if you [work, invest, …] more/less than you normally would."[3] These give rise to much more conservative bounds, which

[1] Zizzo (2010) discusses how demand effects can arise from different sources, such as perceived social pressure from the experimenter, or inferences about appropriate behavior. In psychology, experimenter demand effects are considered a specific case of "demand characteristics" (Orne 1962), which also include the simple effect of being observed ("Hawthorne" effects), or the effect of features of the environment on task construal. Researchers might also worry about "social desirability bias" (respondents taking actions they perceive to be moral or desirable, which may or may not relate to the researcher's objectives), or responses motivated by respondents' own preferences over the findings (e.g., respondents might misreport income in a survey to increase their eligibility for a program). In this paper we focus on inferences about the experimenter's objective, but the framework can easily be adapted to fit other inferences.

[2] Specifically, we study simple time, risk, and ambiguity preference elicitation tasks, a real effort task with and without performance incentives, a lying game, dictator game, ultimatum game (first and second mover), and trust game (first and second mover).

[3] We based this phrasing on Binmore, Shaked, and Sutton's (1985) experiment on the ultimatum game, in which the instructions included the line "You will be doing us a favour if you simply set out to maximize your winnings." These instructions were subsequently criticized precisely because they potentially induce experimenter demand (see, e.g., Zizzo 2010). In recent work, Ellingsen, Östling, and Wengström (2018) use similar language, deliberately using demand to try to shut down social preference motivations in games with communication.

may be useful for applications where concerns about demand are paramount. They also play an important role in our more structural applications, described below, and their strength makes them suited for studying demand effects in their own right.

We establish several novel facts about demand effects. Our first finding is that responses to the weak treatments are modest, averaging around 0.13 standard deviations, varying from close to 0 for unincentivized real effort to 0.29 standard deviations for trust game second movers. In most tasks, our estimates are not significantly different from zero. Overall, we interpret these results as suggesting that demand effects in typical experiments are likely to be small. Responses to our strong demand treatments are much larger, with bounds averaging 0.6 standard deviations and ranging from 0.23 to 1.06 standard deviations. While these bounds are likely more conservative than required in most applications, they illustrate that participants can respond substantially to strong signals about the researcher's objective, and thus researchers are right to pay close attention to potential demand effects in their studies.

The heterogeneity across tasks in responsiveness to our treatments reveals differing levels of *uncertainty* about the importance of experimenter demand in different tasks. For example, there is more uncertainty (i.e., wider bounds) about demand effects for trust game second movers than in the effort task. We provide an additional assumption, "monotone sensitivity," under which this heterogeneity can be interpreted as revealing variation in the *magnitude* of demand effects in different tasks, i.e., that demand effects are larger for trust game second movers.

Next, we apply the method to bounding treatment effect estimates, deriving bounds on the real effort response to performance pay. The bounds we obtain using our weak demand treatments are quite tight, corresponding to around 11 percent of the estimated treatment effect (or 0.07 standard deviations). The strong demand treatments generate wider bounds, but even these more conservative bounds exclude zero, supporting the qualitative finding that incentives increase effort. We apply standard methods to construct "demand-robust" confidence intervals on the bounds and on the underlying actions or treatment effects contained by those bounds. These intervals combine the standard parameter uncertainty due to sampling error with the additional uncertainty due to potential demand effects.

Third, we turn to point estimation of treatment effects. We ask whether applying same-signed demand treatments to both the control and treatment group (for example, demanding high effort from both groups) can reduce or eliminate bias due to experimenter demand. Intuitively, the goal is to "control for" demand by harmonizing beliefs across treatments. We show that this approach is valid under additional assumptions, and apply it to the effort experiment, obtaining a set of alternative estimates, all lying within 10 percent of the conventional treatment effect estimate.

Fourth, following the basic approach of DellaVigna and Pope (2018), we illustrate how sufficiently informative demand treatments can be used in conjunction with a structural model to obtain unconfounded estimates of structural parameters of interest and measure participants' value of conforming to the experimenter's wishes. We estimate that the value of pleasing the experimenter in our effort task is equivalent to increasing the monetary incentives by 20 percent.

Fifth, we explore some of the properties of demand effects. Our approach relies on a monotonicity assumption, essentially assuming that participants want

to comply with, rather than defy, the researchers' wishes. We find strong support for this assumption in average behavior, and at the individual level, using a within-participants design. We show using simple belief data that participants' beliefs about the experimental objective respond as expected to our demand treatments. We also compare our bounds to estimates of the effect of double anonymity in dictator games, a manipulation that has been interpreted as reducing demand. Finally, we examine four moderators of sensitivity to experimenter demand: incentivized versus hypothetical choice; gender; attention; and participant pool.

Finally, we provide an extended summary of recommendations for practitioners, covering how to apply the methods developed, and practical lessons learned from our own applications.

We contribute to the small literature discussing experimenter demand effects (Zizzo 2010; Fleming and Zizzo 2015; Shmaya and Yariv 2016), demand characteristics (Orne 1962), and obedience to the experimenter (Milgram 1963). We are aware of few attempts to directly assess the empirical importance of experimenter demand, and a key contribution of our paper is to provide a general framework for studying demand effects, and evidence from a wide range of standard tasks. In recent work, concurrent with our own, Mummolo and Peterson (2018) conduct two vignette studies on support for free speech and partisan news consumption, and a hypothetical audit study concerning racial bias in hiring, using treatments similar to our weak demand treatments.[4] While they do not construct bounds, they find modest responses to these treatments, in line with our findings.[5]

Relatedly, our paper contributes to the literature on social pressure (DellaVigna, List, and Malmendier 2012; DellaVigna et al. 2017) and moral suasion (Dal Bó and Dal Bó 2014).

We also relate to the literature which examines the effects of anonymity on behavior in the laboratory. Participants who believe their choices are being monitored might be more likely to try to please the experimenter. Hoffman et al. (1994) and List et al. (2004) find that varying anonymity can influence pro-social behavior, while Barmettler, Fehr, and Zehnder (2012) find little effect. Intriguingly, Loewenstein (1999) suggests that participants' responses to the anonymity treatments in Hoffman et al. (1994) could themselves be driven by demand. Our findings also complement work that explores the principal-agent relationship between experimenter and participant (Chassang, Padró i Miquel, and Snowberg 2012; Shmaya and Yariv 2016).

[4] For example, some participants in the audit study are told "We expect that job candidates with names indicating they are white will be more likely to receive an interview because of the historical advantages this group has had on the job market," while others are told "We expect that job candidates with names indicating they are African American will be more likely to receive an interview because corporations are increasingly looking to diversify their workforces."

[5] Other related papers include Cilliers, Dube, and Siddiqi (2015), who show that a white foreigner's presence in the lab in experiments in Sierra Leone distorted giving in dictator games; Lambdin and Shaffer (2009), who find that participants' ability to guess hypotheses varied (but was mostly low) across three different experimental tasks; Bischoff and Frank (2011), in which an actor (unsuccessfully) tried to induce demand effects by their delivery of instructions in a lab game; and Tsutsui and Zizzo (2014), who measure individual demand sensitivity by participants' propensity to select dominated lotteries from a list when told "it would be nice if some of you were to choose" them. List (2007) and Bardsley (2008) argue that behavior in the dictator game is to a large degree an artifact of the experimental situation. Small, Loewenstein, and Slovic (2007) assess the robustness of the "identifiable victim effect" to different question framings, and find that the effect disappears once the experimenter informs respondents about the effect.

Finally, our paper relates to the debate on how lab behavior generalizes to the field (Harrison and List 2004; List 2006; Levitt and List 2007; Falk and Heckman 2009; Camerer 2015; Kessler and Vesterlund 2015). There are multiple reasons why behavior might differ between lab and field, including demand effects. Our focus is on bounding the influence of demand while holding constant other design features. In some cases there may exist a "natural field experiment" counterpart to the design of interest, in which participants are unaware of the experiment, addressing demand alongside other external validity concerns. However, the set of studies that can be practically conducted as natural field experiments is limited. This literature often highlights a distinction between *qualitative* (directional) and *quantitative* effects. Either could be threatened by experimenter demand. Our approach can be used to put quantitative bounds on point estimates, but also to assess whether a qualitative finding could be explained by a demand effect, for instance by asking whether the bounds exclude zero or a sign reversal.

One indication of the level of concern about demand is the consideration given to it in study design. The experimental toolbox contains a number of techniques that are partly or wholly motivated by the goal of reducing the influence of experimenter demand. For example, researchers often work hard to conceal potential signals about the study objective (such as efforts to avoid making gender salient: Bordalo et al. forthcoming); favor between-participant designs despite the larger samples required (Charness, Gneezy, and Kuhn 2012);[6] or conduct costly natural field experiments (Harrison and List 2004).[7] These approaches plausibly make it more difficult for participants to infer the true experimental hypothesis, hopefully reducing the correlation between inference and treatment, or reduce participants' responsiveness to their inferences. But it is difficult to be sure that one has been successful, or that participants are not acting out some other conjecture that could be correlated in unpredictable ways with treatment. It is also difficult to know what is the set of studies that remain unpublished, or not even conducted, due to unresolved concerns about demand. Our bounding approach seeks to isolate the hidden demand effects by *amplifying* them with an explicit demand effect. It can be applied broadly without requiring major changes to experimental design, and we believe it will prove a useful addition to the toolbox.

The paper proceeds as follows. Section I presents a simple model of experimenter demand. Section II describes the experiments. Section III presents bounds on natural actions and treatment effects, demand-corrected point estimates, and structural estimates. Section IV examines properties of demand effects and the assumptions underlying our approach. Section V provides guidance for applying our approach in different settings. Section VI concludes. The online Appendix contains theoretical details and additional results.

[6] Charness, Gneezy, and Kuhn (2012, p. 2): "Within designs may lead to spurious effects, through respondents expecting to act in accord with some pattern, or attempting to provide answers to satisfy their perceptions of the experimenter's expectations… Demand effects are likely to be stronger in a within design."

[7] Other design features include abstract framing of choices, anonymized responses, homogenized delivery of instructions, and incentivized choice. Review articles by Zizzo (2010) and de Quidt, Vesterlund, and Wilson (forthcoming) provide a discussion; de Quidt, Vesterlund, and Wilson (forthcoming) also measure their adoption in published experimental papers.

I. Theory

We now derive a simple model of experimenter demand and demand treatments. We begin with the three central assumptions at the heart of our approach, and provide a Bayesian model that generates them. Next, we discuss demand treatment design. We conclude with a brief discussion of heterogeneity, and defiers, participants who do the opposite of the experimenter's wishes. Online Appendix Sections B.B5 and B.B6 extend the model to allow participants to infer the *importance* of the experimenter's objective, and to model demand treatments that ask participants to ignore the experimenter's objective.

We model a decision-maker (he) who has preferences over outcomes induced by his action $a \in \mathbb{R}$ in an experiment. Note that a could be continuous or discrete, but for simplicity we focus on the case of continuous actions with a natural ordering (more/less effort, investment, giving).

In the absence of demand effects, the optimal action is simply a function of the decision-making *environment*. We index environments by $\zeta \in Z$, where ζ captures aspects including participant characteristics (e.g., male/female, student/representative sample), setting (e.g., lab/field, online/in-person), experimental treatments, the content and framing of information provided to participants, and so on. A key component of ζ is information the participant has about other treatments (e.g., in a within-participant design), which might inform their beliefs about the experimental objective.

Given ζ, we define the "natural" action $a(\zeta)$ as that which would be taken absent any confounding motive for pleasing the experimenter.[8] The experimenter (she) is interested in measuring a specific action $a(\zeta)$ (e.g., the level of giving out of an endowment), or a treatment effect $a(\zeta_1) - a(\zeta_0)$ (e.g., the effect of incentives on effort provision). Unfortunately, her task is complicated by experimenter demand. After observing ζ, the decision-maker forms a conjecture about the experimenter's wishes or objectives, which may change his action. Instead of $a(\zeta)$, he chooses action $a^L(\zeta)$, where L signifies the presence of a "latent," unobserved experimenter demand influence. The influence could increase or decrease a: $a^L(\zeta) \gtreqless a(\zeta)$. We define the *latent demand effect* in environment ζ as the difference $a^L(\zeta) - a(\zeta)$.

While nonzero latent demand automatically biases estimates of mean actions, it does not necessarily bias estimates of treatment effects. To see this, note that the observed treatment effect can be decomposed as follows:

$$(1) \quad a^L(\zeta_1) - a^L(\zeta_0) = \underbrace{a(\zeta_1) - a(\zeta_0)}_{\text{Effect of interest}} + \underbrace{[a^L(\zeta_1) - a(\zeta_1)]}_{\text{Latent demand in } \zeta_1} - \underbrace{[a^L(\zeta_0) - a(\zeta_0)]}_{\text{Latent demand in } \zeta_0}.$$

The first term on the right-hand side is the treatment effect of interest. The second and third capture the potential bias due to experimenter demand. If both demand effects are equal they cancel and the treatment effect is identified, but they may not

[8] In some experiments, the experimenter essentially fills the role of a real-world authority figure. For example, part of the real-world response to incentives might include a response to perceived demand from an employer. For a researcher interested in the total effect of incentives, perceived demand may actually be part of the environment of interest, ζ, rather than a confound.

cancel, either because the participant's inference or his response to a given inference varies with ζ. The usual logic of a randomized experiment is to ensure that variation in treatment is orthogonal to potential confounds, but as demand effects may be driven by the treatment itself, randomization does not guard against bias.

Example 1: Consider two variants on the Dictator game, in which a participant is told to choose what fraction of \$10 to give to a recipient. In variant 0, he is told that the recipient is aware that the choice is taking place, while in variant 1 they are unaware (for instance, the money will just be added to a show-up fee). Absent any motive for pleasing the experimenter, the participant would prefer to give \$4, so the true treatment effect is $a(\zeta_1) - a(\zeta_0) = \$0$. However, in variant 0 he infers that the experimenter wants him to be generous, so he gives \$5, while in variant 1 he infers that the experimenter wants him to be selfish, so he gives \$0. The experimenter fails to measure true preferences in either case, and identifies a treatment effect that is in reality a demand effect.

A. *Demand Treatments*

We now assume that the experimenter has at her disposal a particular kind of treatment manipulation which we call a *demand treatment.* Negative demand treatments deliberately signal a demand that the decision-maker decrease his action, inducing $a^-(\zeta)$, while positive demand treatments demand an increase and induce $a^+(\zeta)$. Our first substantive assumption is a basic monotonicity condition.

ASSUMPTION 1 (Monotonicity): $a^-(\zeta) \leq a^L(\zeta) \leq a^+(\zeta)$.

Assumption 1 requires that demanding an increased action does not decrease it, and vice versa. It has a natural connection to the monotonicity condition in the estimation of local average treatment effects (Imbens and Angrist 1994): the assumption rules out "defier" behavior whereby participants do the opposite of what is demanded.

Our main assumption amounts to assuming that the demand treatments can bound the natural action of interest.

ASSUMPTION 2 (Bounding): $a^-(\zeta) \leq a(\zeta) \leq a^+(\zeta)$.

It implies bounds for natural actions (2) and treatment effects (3):

$$a(\zeta) \in [a^-(\zeta), a^+(\zeta)], \tag{2}$$

$$a(\zeta_1) - a(\zeta_0) \in [a^-(\zeta_1) - a^+(\zeta_0), a^+(\zeta_1) - a^-(\zeta_0)]. \tag{3}$$

For some purposes we may wish to be able to make comparative statements about demand in different environments. Although the latent demand effect is unobservable, the sensitivity of behavior to demand treatments may be informative about it. First, we define what we mean by "sensitivity."

DEFINITION 1 (Sensitivity): *Sensitivity is the difference in actions under positive and negative demand treatments:* $S(\zeta) = a^{+}(\zeta) - a^{-}(\zeta)$.

Remark 1: In addition to bounding the natural action, Assumptions 1 and 2 jointly imply that sensitivity $S(\zeta)$ provides an upper bound on the magnitude of the latent demand effect: $S(\zeta) \geq \left|a^{L}(\zeta) - a(\zeta)\right|$.

This fact enables us to use sensitivity $S(\zeta)$ to make statements of *comparative ignorance*, in the sense that if $S(\zeta_1) > S(\zeta_0)$ there is more scope for large latent demand effects under ζ_1 than ζ_0. But it could nevertheless be that the true latent demand effect is larger under ζ_0. Our third assumption, Monotone Sensitivity, allows us to make concrete claims about magnitudes.

DEFINITION 2 (Comparison Classes): *A comparison class* $Z^{C} \subseteq Z$ *is a set of environments such that Monotone Sensitivity holds for all* $z \in Z^{C}$.

ASSUMPTION 3 (Monotone Sensitivity): $S(z)$ *is strictly increasing in* $\left|a^{L}(z) - a(z)\right|$ *for all* $z \in Z^{C}$.

Monotone Sensitivity permits statements such as "latent demand is stronger for participant pool A than participant pool B" or "latent demand is stronger under incentive scheme A than incentive scheme B." We derive some comparison classes below using our Bayesian model.

B. *Bayesian Model*

We now provide a simple foundation for our main assumptions, and derive conditions under which they will or will not hold. The environment ζ determines the mapping from actions $a \in \mathbb{R}$ into outcomes or distributions over outcomes. The decision-maker's payoff is $v(a, \zeta)$, where v captures the payoff structure (mapping from actions to outcomes) and preferences (mapping from outcomes to utility). We assume v is strictly concave and differentiable, so the natural action $a(\zeta)$ solves $v_1(a(\zeta), \zeta) = 0$.

Latent Demand.—Demand enters preferences as follows. Upon observing ζ, the decision-maker makes an inference about the experimenter's objective, $h \in \{-1, 1\}$. If $h = -1$, he believes the experimenter benefits from him taking low actions, while if $h = 1$ he believes she benefits from high actions. He has a preference, ϕ, for pleasing the experimenter, which we allow to depend upon ζ.[9] We remain agnostic about *why* the participant wishes to please the experimenter; possible motives include altruism, a motive to conform, or a belief that he will ultimately be rewarded for doing so.

[9] We have in mind that ϕ might depend on the identity of the experimenter (e.g., a firm versus a researcher) or decision-maker (e.g., women might have different attitudes than men). ϕ might also vary with other features such as the salience of the benefit to the experimenter, or how important the participant believes his actions are for achieving the experimenter's objectives.

We assume utility takes the following separable form:

$$U(a,\zeta) = v(a,\zeta) + a\phi(\zeta)E[h\,|\,\zeta]. \tag{4}$$

The optimal action $a^L(\zeta)$ thus solves

$$v_1(a^L(\zeta),\zeta) + \phi E[h\,|\,\zeta] = 0, \tag{5}$$

so $a^L(\zeta) = a(\zeta) \Leftrightarrow \phi E[h\,|\,\zeta] = 0$. There is therefore no demand confound if either the decision-maker assigns equal likelihood to the preferred action being high or low ($E[h\,|\,\zeta] = 0$), or he does not care about the experimenter's objectives ($\phi = 0$) (these would be expected in a "natural field experiment," where the participant is unaware of the experiment). We assume the decision-maker's mean prior over h is $E[h] = 0$, so in the absence of any new information about h he chooses $a(\zeta)$. The relation between actions and beliefs is captured by $da^L(\zeta)/dE[h\,|\,\zeta] = -\phi/v_{11}(a,\zeta)$, which has the same sign as ϕ. Actions are monotone in beliefs.

We model learning as follows. The environment ζ includes a signal $h^L(\zeta) \in \{-1,1\}$ which the decision-maker believes is a sufficient statistic, i.e., $E[h\,|\,h^L(\zeta),\zeta] = E[h\,|\,h^L(\zeta)]$. He believes that with probability $p^L(\zeta)$, the signal is correct ($h^L = h$), and with probability $1-p^L(\zeta)$ it is pure noise ($h^L = \epsilon$, where ϵ equals -1 or 1 with equal probability). We impose that $p^L(\zeta) \in [0,1)$. It is straightforward to see that

$$E[h\,|\,h^L(\zeta)] = h^L(\zeta)p^L(\zeta). \tag{6}$$

The decision-maker's belief depends on ζ in two ways. First, via the sign of $h^L(\zeta)$, i.e., whether he believes that the experimenter wants a high or low action, which determines the *direction* of the latent demand effect. Second, via $p^L(\zeta)$, i.e., the perceived informativeness of the signal, which affects the *magnitude* of the latent demand effect.

Demand Treatments.—We assume that the experimenter can choose a "demand treatment" signal $h^T \in \{-1,1,\emptyset\}$; $h^T = \emptyset$ corresponds to the usual case in which no demand treatment is used, while $h^T = 1$ and $h^T = -1$ correspond to positive and negative demand treatments. These signals provide information about h so as to direct the decision-maker's beliefs. We assume that if $h^T = \emptyset$ the decision-maker does not update his belief about h (for example because their prior is that demand treatments are never used). This assumption is reasonable as (at present) demand treatments are rarely used in experiments. We maintain throughout that ζ (and hence $v(a,\zeta)$, $h^L(\zeta)$, $p^L(\zeta)$, and $\phi(\zeta)$) does not depend on the demand treatment, i.e., receiving a demand treatment does not change the decision-maker's interpretation of the maintained experimental environment or their motive for pleasing the experimenter. Instead the demand treatment is interpreted purely as informative about the direction of the experimenter's objective.[10]

[10] Formally, we assume that $\zeta(h^T) = \zeta, \forall\zeta$. This assumption will be stronger for some demand treatments and environments than others, and is an important consideration in the selection of appropriate demand treatments. If

The decision-maker believes that h^T is informative about h: with probability p^T, h^T equals h, and with probability $1 - p^T$ it equals η, which takes values -1 and 1 with equal probability. Here, η and ϵ are believed to be independent (we revisit this assumption in online Appendix Section B.B6). The Bayesian posterior is

$$E[h \mid h^T, h^L(\zeta)] = \frac{h^L(\zeta)p^L(\zeta) + h^T p^T}{1 + h^L(\zeta)p^L(\zeta)h^T p^T}. \tag{7}$$

Thus, if $h^L(\zeta) = h^T$, the demand treatment reinforces the participant's belief, while if the signals have opposite signs they offset one another.

Assumptions.—We now use the model to provide foundations for our main assumptions described in Section IA. Derivations can be found in online Appendix B.

First, Assumption 1 (Monotonicity) states that a positive demand treatment increases the action (relative to no demand treatment) and the negative demand treatment decreases it. It is straightforward to see that except for the trivial case $p^T = 0$, these conditions are satisfied if and only if $\phi \geq 0$, i.e., a weak preference for pleasing the experimenter.

PROPOSITION 1: *Monotonicity holds for all p^T if and only if $\phi \geq 0$.*

Second, Assumption 2 (Bounding) states that the demand treatments provide bounds on the true action. In the Bayesian model, given $\phi \geq 0$ (Monotonicity), the action is larger or smaller than $a(\zeta)$ when $\phi E[h \mid h^T, h^L] \geq 0$ or $\phi E[h \mid h^T, h^L] \leq 0$ respectively. Intuitively, whatever the latent demand effect, the demand treatment that opposes it must be informative enough to reverse the sign of beliefs. It is clear from inspection of (7) that this simply requires the demand treatments to be "more informative" than latent demand, $p^T \geq p^L(\zeta)$.

PROPOSITION 2: *Given $\phi \geq 0$, Bounding holds if and only if $p^T \geq p^L(\zeta)$.*

Finally, Assumption 3 (Monotone Sensitivity) states that within a comparison class Z^C of environments, differences in sensitivity are informative about differences in underlying latent demand. Latent demand and sensitivity can vary for multiple reasons, so there is no simple condition that guarantees when this assumption will and will not hold. In online Appendix Section B.B3, we work out some important cases. First, we show that Monotone Sensitivity holds when variation in demand effects is driven by differences in the strength of preference for pleasing the experimenter, ϕ. Second, we analyze Monotone Sensitivity when variation in demand effects is driven by differences in the payoff function, v, deriving specific conditions when v is additively or multiplicatively separable and providing examples such as variation in incentives. Third, we show that Monotone Sensitivity holds in a model

it does not hold then Bounding might fail because the demand treatments alter the natural action itself: $a(\zeta(\emptyset)) \notin [a(\zeta(-1)), a(\zeta(1))]$. In online Appendix Section B.B5, we extend the model to allow ϕ to depend on h^T and show that the Bounding condition remains unchanged.

of inattention to experimenter demand. Finally, we show that Monotone Sensitivity does *not* hold in general when environments differ in the beliefs they induce $(E[h \mid h^L(\zeta)])$. We use these findings when interpreting heterogeneous responses to demand treatments in Section IVD.

C. *"Weak" and "Strong" Demand Treatments*

There are many different ways to signal a desire for high or low actions. How should the experimenter choose? The model gives us a way to answer this question. The width of the bounds $[a^-(\zeta),\ a^+(\zeta)]$ is increasing in p^T. Therefore the tightest bounds, subject to satisfying Bounding $(p^T \geq p^L(\zeta))$, are obtained when $p^T = p^L(\zeta)$. In other words, we want the "least informative" demand treatment possible, subject to being "informative enough" for Bounding.[11] We want to choose demand treatments that are likely to be "stronger" or more informative than any latent demand in the study of interest, while avoiding excessively strong signals that lead to uninformative bounds.

In our empirical applications we employ two types of demand treatments, described in more detail below. Our "weak" manipulations explicitly signal what we expect participants to do; we believe these are already more informative than likely latent demand in typical experiments. Our "strong" manipulations go further, telling participants which action will "do us a favor." These lead to more conservative bounds, and may be useful for applications where researchers are especially concerned about demand effects. They also play a role in more structural applications, described in Sections IIID and IIIE.

D. *Heterogeneity and Defiers*

The approach naturally extends to the case where participants are heterogeneous and the experimenter is interested in average behavior or average treatment effects. If Monotonicity and Bounding hold for all agents individually, then they also hold for average actions, so we can simply reinterpret a, a^L, a^+, and a^- as representing average behaviors and our approach remains valid.

An important dimension of heterogeneity is in ϕ, the preference for pleasing the experimenter. Monotonicity requires a weak positive preference, $\phi \geq 0$. "Defiers" with $\phi < 0$ prefer to go against the experimenter's wishes. Bounding fails for these individuals, because $a^- > a^+$. We show in online Appendix Section B.B4 that the method is able to tolerate some defier behavior, but too much will lead to failure to bound the average natural action. We give an example where Bounding is satisfied provided the average participant is a complier. In general, for defier behavior to be "small enough," the joint distribution of preferences and beliefs must be such that the response by the compliers outweighs that of the defiers.

[11] This gives a novel reason why deception in experiments can be problematic. If the demand treatment is regarded as uninformative because participants are used to second-guessing what experimenters are really after, then the bounding exercise is invalidated. We thank an anonymous referee for this observation.

II. Sample and Experimental Design

We conducted seven experiments in total to demonstrate our approach and to provide estimates of demand sensitivity on a wide range of standard experimental tasks (to save space, we provide citations for the tasks in online Appendix E). Our respondents complete 1 of 11 tasks: a dictator game; a risky investment game, without or with ambiguity; a convex time budget task; a trust game (first or second mover); an ultimatum game (first or second mover); a lying game; and a real effort task with or without performance pay. We conduct all of our experiments online, primarily because the large number of treatments would be infeasible to implement in the laboratory. We designed the experiments to maximize comparability. For all experiments except the effort task, the action spaces are similar (they can be expressed as real numbers from 0 to 1); we pay the same show-up fee; recruit from the same participant pools; use the same mode of collection (online); the same response mode (sliders); and keep stakes as similar as possible.[12]

We employ two phrasings for our demand treatments. Our "weak" treatments explicitly tell participants that we expect high or low actions. For example, in the investment game, participants were told at the end of their instructions that "We expect that participants who are shown these instructions will invest more/less in the project than they normally would."[13] The strong treatments go further, telling participants that they will "do us a favor" by taking a higher or lower action. For example, in the dictator game, participants in the positive demand condition were told "You will do us a favor if you give more/less to the other participant than you normally would." We keep the phrasing of the demand treatments as homogeneous as possible across tasks. In the two-player games we do not provide information about demand treatments shown to the other player, but our approach could be extended to create common knowledge about demand.

Table 6 summarizes the design features of each experiment, and Table 7 provides design details, parameters, and the exact wording of the demand treatments for each task. Online Appendix Figure A1 gives an example of the experimental interface. Full experimental instructions can be found on the journal website.

A. *Participant Populations*

We conducted six experiments with approximately 16,000 participants (or "workers") on Amazon Mechanical Turk (MTurk) (Experiments 1–3 and 5–7), and one experiment with around 3,000 participants using an online panel sample representative of the US population in terms of region, age, income, and gender (Experiment 4). MTurk is an online labor marketplace that is frequently used by researchers for surveys and experiments. It is attractive because it offers a large

[12] For the effort task, we replicated the design employed in DellaVigna and Pope (2018, forthcoming). The primary differences with our other tasks are a higher show-up fee and a different response mode (effort).

[13] It is not completely straightforward to design demand treatments that report the experimental hypothesis, because if the experimenter truly hypothesizes that the action will be high in one treatment, telling participants she expects it to be low could be considered deceptive. By referring to "participants who are shown these instructions" (which include the demand treatment) we avoid this issue, because it is indeed true that we expect high actions from participants in the positive demand treatment group and low actions in the negative demand treatment group.

and diverse pool of workers. There is some evidence that MTurk workers are more attentive to instructions than college students (Hauser and Schwarz 2016). To participate in our MTurk experiments, workers had to live in the United States, have an overall approval rating of more than 95 percent, and have completed more than 500 tasks on MTurk, fairly standard parameters in research on MTurk.[14]

Most workers on MTurk are experienced in taking surveys, which might affect the external validity of our results. We used the representative sample, whose participants are less experienced with social science experiments, to replicate a subset of our findings. The sample is maintained by a market research company, *Research Now*.

B. *Pre-Analysis Plans*

Our experiments were conducted in a sequence, between May 2016 and May 2017. Each is described in a pre-analysis plan (PAP) posted online prior to launch.[15] The sequence is laid out in Table 6. For each experiment, the PAP details the data to be collected, treatment variables, experimental instructions, and how we planned to analyze that experiment's data.

However, presenting the data experiment-by-experiment is repetitious. Therefore, for brevity and clarity of exposition, in the paper we pool the data and analyze all tasks side-by-side for our weak and strong demand treatments separately (this structure was described in pre-analysis plan 5). Our main analysis uses data from MTurk respondents with real stakes, which we have for all 11 tasks studied. In the analysis of heterogeneity we introduce hypothetical choice data from MTurk and the representative panel, which were collected for a subset of tasks. When averaging across tasks we weight observations to give equal weight to each task.

Other than this pooling across experiments, our analysis closely follows what was pre-specified.[16] For completeness, online Appendix C presents all pre-specified analyses, experiment-by-experiment. We refer to findings in the text if relevant.

[14] We excluded prior participants when recruiting for experiments 2 and 3. Technically this is achieved by applying a "qualification" flag to the MTurk accounts of prior participants, which can then be used to prevent them from seeing or accepting new MTurk tasks posted by us. At the time of running experiments 5 and 6, we had essentially exhausted the active participant pool, and to avoid undue delays in recruitment we therefore allowed prior participants to take part. Around 36 percent of the respondents in these experiments had not participated before. In experiment 7, which was conducted some time later, we did exclude prior participants, but a server communication error meant that not all accounts received the qualification flag, and as a result some prior participants did take part. Seventy percent of the respondents in this experiment had not participated before. Our results are virtually unchanged by the dropping of participants who completed more than one of our experiments; results are available upon request.

[15] The pre-analysis plans were posted on the Social Science Registry and can be found here: https://www.socialscienceregistry.org/trials/1248.

[16] In some experiments we proposed to standardize responses based on average choices in the no-demand condition. Because we did not collect no-demand data for all tasks, for consistency we always standardize based on the negative demand treatment group (a simple and inconsequential linear transformation). For our real-effort tasks, which were based on DellaVigna and Pope (2018), we pre-specified that we would apply their exclusion criteria to the analysis dataset (excluding participants who take more than 30 minutes, take the task more than once, score 0 or more than 4,000 points, or have invalid MTurk IDs). In our other experiments we did not pre-specify exclusions, but for consistency we also drop participants who submitted multiple responses (less than 0.5 percent). This is inconsequential for the results.

Table 1—Response to Weak Demand Treatments, All Incentivized Tasks

	Time	Risk	Ambiguity Aversion	Effort 0 cent bonus	Effort 1 cent bonus	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust Game 1	Trust Game 2
Panel A. Unconditional means											
Positive demand	0.770	0.524	0.557	0.331	0.484	0.537	0.382	0.470	0.413	0.455	0.398
	(0.027)	(0.023)	(0.024)	(0.014)	(0.012)	(0.012)	(0.014)	(0.014)	(0.014)	(0.023)	(0.017)
No demand		0.541					0.313				
		(0.021)					(0.015)				
Negative demand	0.766	0.472	0.499	0.343	0.469	0.530	0.318	0.443	0.362	0.430	0.348
	(0.027)	(0.021)	(0.024)	(0.014)	(0.013)	(0.011)	(0.014)	(0.013)	(0.013)	(0.025)	(0.012)
Panel B. Sensitivity (positive − negative)											
Raw data	0.005	0.052	0.058	−0.012	0.015	0.007	0.063	0.027	0.051	0.025	0.050
	(0.038)	(0.031)	(0.034)	(0.019)	(0.018)	(0.016)	(0.020)	(0.019)	(0.019)	(0.034)	(0.021)
z-score	0.012	0.156	0.174	−0.063	0.078	0.042	0.240	0.158	0.281	0.076	0.289
	(0.096)	(0.091)	(0.102)	(0.101)	(0.094)	(0.102)	(0.075)	(0.112)	(0.102)	(0.104)	(0.125)
		[0.096]					[0.002]				
Panel C. Monotonicity											
Positive − neutral		−0.051					0.261				
(z-score)		(0.092)					(0.078)				
		[0.237]					[0.002]				
Negative − neutral		−0.207					0.021				
(z-score)		(0.087)					(0.078)				
		[0.056]					[0.357]				
Observations	422	739	390	388	381	412	758	360	411	352	346

Notes: This table uses data from incentivized MTurk respondents with weak demand treatments. Panel A displays mean actions with standard errors in the positive, negative, and no-demand conditions, respectively. Panel B presents the raw and z-scored sensitivity of behavior to our demand treatments. Panel C displays the response to our positive and negative demand treatments separately, when "no demand" choices were also collected. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets, adjusting across tests within each task when testing the Monotonicity assumption.

C. *Summary Statistics*

Online Appendix Tables D1 to D7 present the pre-specified balance tables for all of the experiments. Tables D8 to D15 provide summary statistics on our respondents. Table D12 highlights that respondents from the online panel are representative of the US population by gender, income, age, and region, and other observables. Attrition was low, below 2 percent on average, and did not differ across demand treatment arms (Tables D16 and D17).

III. Applying the Method

A. *Bounding Natural Actions*

In this section we provide bounds on natural actions estimated using our weak and strong demand treatments. For a subset of tasks we also measured behavior with no demand treatment, and describe these results in Section IVA where we discuss Monotonicity. Our objects of interest here are mean behavior in the positive $(a^{+}(\zeta))$ and negative $(a^{-}(\zeta))$ demand conditions.

Panel A of Table 1 and Figure 2 show mean actions by task and demand treatment for incentivized MTurk respondents with weak treatments. Panel B of Table 1

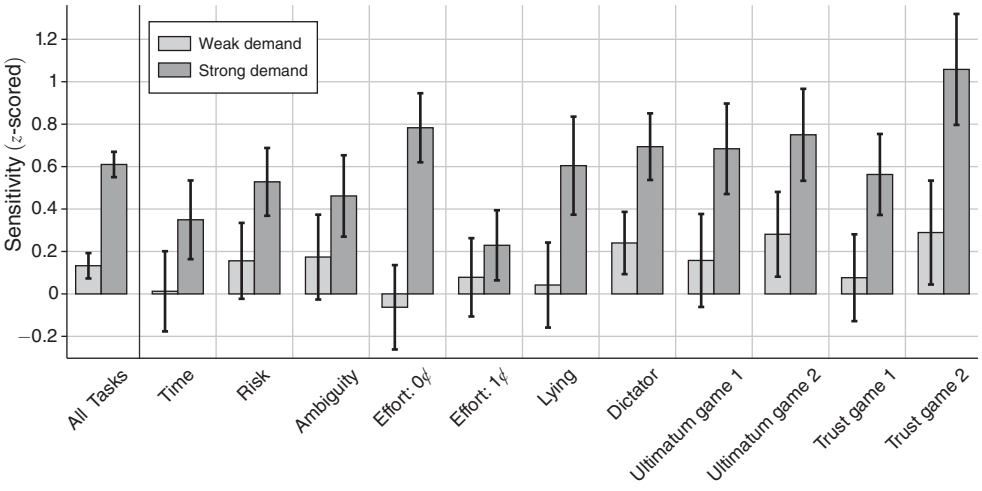

FIGURE 1. SENSITIVITY TO DEMAND TREATMENTS, z-SCORED

Notes: This figure uses data from incentivized MTurk respondents with weak and strong demand treatments. It presents the z-scored sensitivity of behavior to our demand treatments, i.e., the normalized difference in behavior between the positive and negative demand conditions. Error bars indicate 95 percent confidence intervals.

and Figure 1 display sensitivities $(a^{+}(\zeta) - a^{-}(\zeta))$, in both raw and z-scored units. Sensitivity is modest, averaging around 0.13 standard deviations, and frequently not significantly different from zero. The strongest responses (between 0.2–0.3 standard deviations) were observed for the dictator game, the ultimatum game second mover, and the trust game second mover. As we have argued, the weak manipulations seem likely to satisfy bounding for typical applications, so these results give cause for optimism.

Panel A of Table 2 and Figure 2 show mean actions in the different demand treatment arms employing strong treatments. Panel B of Table 2 and Figure 1 display sensitivities. Behavior is responsive to our strong demand treatments, and sensitivity is significantly different from zero in all tasks, averaging around 0.6 standard deviations. Sensitivity is particularly high in the dictator game, for second movers in the trust and ultimatum games, and for unincentivized effort. These manipulations are significantly stronger than likely implicit signals in most experiments or surveys, so provide quite conservative upper bounds on typical demand biases. However, they do demonstrate that participants are motivated to respond to signals about the researcher's goals, and that responses can be significant when those signals are strong. Thus, the attention researchers pay to potential demand effects at the study design stage is well justified.

B. *Bounding Treatment Effects*

Our real effort experiments replicate treatments from DellaVigna and Pope (2018). Participants alternately pressed the "a" and "b" keyboard buttons for 10 minutes, earning one point per pair. One group was told that their score "will not affect [their] payment," while a second group received 1 cent per 100 points. By combining the

TABLE 2—RESPONSE TO STRONG DEMAND TREATMENTS, ALL INCENTIVIZED TASKS

	Time	Risk	Ambiguity Aversion	Effort 0 cent bonus	Effort 1 cent bonus	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust Game 1	Trust Game 2
Panel A. Unconditional means											
Positive demand	0.795	0.550	0.583	0.405	0.492	0.606	0.434	0.520	0.474	0.535	0.469
	(0.024)	(0.020)	(0.024)	(0.011)	(0.011)	(0.013)	(0.015)	(0.013)	(0.014)	(0.024)	(0.017)
No demand	0.786	0.466		0.341	0.476		0.282				
	(0.025)	(0.022)		(0.012)	(0.012)		(0.015)				
Negative demand	0.659	0.373	0.428	0.255	0.449	0.510	0.251	0.404	0.337	0.350	0.288
	(0.028)	(0.019)	(0.023)	(0.011)	(0.011)	(0.014)	(0.014)	(0.014)	(0.015)	(0.022)	(0.015)
Panel B. Sensitivity (*positive* − *negative*)											
Raw data	0.137	0.177	0.155	0.150	0.043	0.096	0.183	0.116	0.136	0.185	0.181
	(0.037)	(0.027)	(0.033)	(0.016)	(0.016)	(0.019)	(0.021)	(0.018)	(0.020)	(0.032)	(0.023)
z-score	0.349	0.528	0.462	0.783	0.229	0.604	0.694	0.684	0.750	0.563	1.058
	(0.095)	(0.082)	(0.098)	(0.083)	(0.084)	(0.118)	(0.080)	(0.109)	(0.111)	(0.097)	(0.133)
	[0.001]	[0.001]		[0.001]	[0.020]		[0.001]				
Panel C. Monotonicity											
Positive − neutral	0.022	0.252		0.333	0.084		0.574				
(z-score)	(0.088)	(0.088)		(0.085)	(0.088)		(0.082)				
	[0.363]	[0.001]		[0.001]	[0.159]		[0.001]				
Negative − neutral	−0.327	−0.276		−0.450	−0.145		−0.120				
(z-score)	(0.097)	(0.086)		(0.085)	(0.086)		(0.080)				
	[0.001]	[0.001]		[0.001]	[0.101]		[0.046]				
Observations	727	728	404	731	714	365	770	409	421	382	371

Notes: This table uses data from incentivized MTurk respondents with strong demand treatments. Panel A displays mean actions with standard errors in the positive, negative, and no-demand conditions, respectively. Panel B presents the raw and z-scored sensitivity of behavior to our demand treatments. Panel C displays the response to our positive and negative demand treatments separately, when "no demand" choices were also collected. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets, adjusting across tests within each task when testing the Monotonicity assumption.

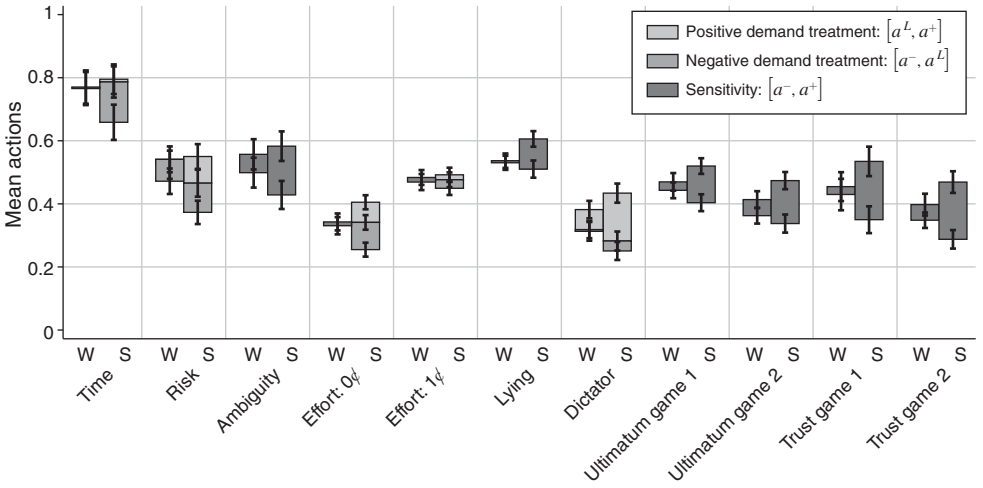

FIGURE 2. BOUNDING NATURAL ACTIONS

Notes: This figure uses data from incentivized MTurk respondents with weak (W) and strong (S) demand treatments. It displays mean responses by task and demand treatment. Upper (lower) points correspond to positive (negative) demand treatments (a^+ and a^-), intermediate points to "no demand" treatments (a^L, not collected for all tasks). Lighter shaded sections indicate the response to positive and negative demand treatments separately, dark shaded sections indicate sensitivity when a^L was not measured. Error bars indicate 95 percent confidence intervals.

Table 3—Bounding Treatment Effects

	Conventional	Weak bounds		Strong bounds	
	Treatment effect	Lower	Upper	Lower	Upper
Count	540.720	530.001	588.270	177.421	948.978
	(66.763)	(64.532)	(61.499)	(62.379)	(64.148)
Count (z-scored)	0.686	0.673	0.747	0.225	1.205
	(0.085)	(0.082)	(0.078)	(0.079)	(0.081)

Notes: This table uses data from the real effort experiments with weak and strong demand treatments (experiments 3 and 6). Column 1 shows conventional treatment effect estimates. Columns 2 to 5 show lower and upper bounds estimated using weak and strong treatments. We apply the "ironing" procedure described in Section IIIB when constructing the weak estimates. Robust standard errors in parentheses. "Count" is the raw score from the experiment, Count (z-scored) is standardized using the negative demand condition, pooled across incentive treatment arms.

bounds estimated for each incentive treatment we can construct bounds on the treatment effect of performance pay on effort provision.[17]

Table 3 displays the conventional treatment effect $(a^L(1) - a^L(0)$, where "1" and "0" correspond to the reward per 100 points), the upper bound of the treatment effect $(a^+(1) - a^-(0))$, and the lower bound $(a^-(1) - a^+(0))$. In words, the lower bound on the treatment effect is given by comparing participants who received performance pay, coupled with a negative demand treatment, to participants who received no performance pay, coupled with a positive demand treatment. We first show the bounds generated using our weak treatments, which are quite tight, ranging from 0.67 to 0.75 standardized units.[18] The width of these bounds corresponds to only 11 percent of the estimated treatment effect (or 0.07 standard deviations), suggesting a limited role for experimenter demand in explaining the effort response to incentives. Naturally, the bounds created using the more conservative strong treatments are much wider, ranging from 0.23 to 1.21 standard deviations. Even these conservative bounds support the qualitative finding that effort responds to incentives.

C. *Confidence Intervals*

It is possible to compute confidence intervals for (i) the bounds themselves, and (ii) the parameters contained by those bounds (a natural action or treatment effect), following Imbens and Manski (2004) (see online Appendix Section B.B7 for details). The latter can be thought of as "demand-robust" confidence intervals, combining conventional parameter uncertainty due to sampling error with the additional

[17] Our pre-analysis plans did not explicitly describe the bounding of treatment effects, but it is an immediate extension of the approach to bounding actions.

[18] In constructing the bounds using our weak treatments, we note that the average effort in the no-incentive condition was actually slightly higher for those receiving negative demand than those receiving positive demand, i.e., we observe a small monotonicity failure $(a^+(0) < a^-(0))$. When sensitivity is low, such outcomes can easily arise due to sampling variation; both values here are statistically indistinguishable. In such cases, the procedure we propose in this section could lead to bounds on the treatment effect with negative width. A conservative approach, which we follow, is to first "iron" the bounds on the no-incentive condition, by averaging them. Formally, one can compute $a^+_{iron}(\zeta) = \max\{a^+(\zeta), 0.5[a^+(\zeta) + a^-(\zeta)]\}$ and $a^-_{iron}(\zeta) = \min\{a^-(\zeta), 0.5[a^+(\zeta) + a^-(\zeta)]\}$, and then use these values when computing the bounds on the treatment effect, which become $a^+(1) - a^-_{iron}(0)$, and $a^-(1) - a^+_{iron}(0)$. Because in this case $a^+_{iron}(0) = a^-_{iron}(0)$, the width of the weak bounds on the treatment effect is simply equal to $a^+(1) - a^-(1)$.

uncertainty about possible demand effects. Uncertainty due to sampling error can be reduced in the usual way by increasing sample size (specifically, in the demand treatment arms), while uncertainty due to demand is reduced by selecting minimally informative demand treatments, subject to Bounding (see Section IC). Online Appendix Table A3 presents confidence intervals computed from individual tasks using both the weak and strong demand treatments. Table A4 presents confidence intervals on the bounds and treatment effect of the effect of incentive pay in the effort experiment. Zero lies outside these confidence intervals, providing statistical support for the finding that incentives increased effort.

D. *Controlling for Demand*

The nonparametric bounding approach described above yields bounds on treatment effects, but researchers may be interested in point estimates that "control for" demand effects. Intuitively, one might apply same-signed demand treatments (positive-positive or negative-negative) to the treatment group and the control group, with the goal of harmonizing demand between treatments. In this section we describe how using this approach can eliminate bias if demand treatments are assumed to be fully informative ($p^T = 1$), and can reduce bias in other cases. Derivations are given in online Appendix Section B.B8.[19]

We will assume throughout that Monotonicity holds strictly, i.e., $\phi > 0$ ($\phi = 0$ would imply no demand bias). The participant's usual first-order condition, with demand treatment h^T and optimal action $a^*(\zeta, h^T)$, is $v_1(a^*(\zeta, h^T), \zeta) + \phi(\zeta) E[h \mid h^T, h^L(\zeta)] = 0$. A first-order Taylor approximation around the natural action $a(\zeta)$ yields

$$a^*(\zeta, h^T) \approx a(\zeta) + \Phi(\zeta) E[h \mid h^T, h^L(\zeta)], \tag{8}$$

where $\Phi(\zeta) \equiv -\phi(\zeta)/v_{11}(a(\zeta), \zeta)$ is a slope term capturing the effect of beliefs on actions, which we term "responsiveness." Note that Φ is positive as $v_{11} < 0$.

Assume two treatment groups, $\zeta \in \{0, 1\}$, with identical demand treatments $h^T \in \{-1, 1, \emptyset\}$, from which we estimate a treatment effect $a^*(1, h^T) - a^*(0, h^T)$. Its bias relative to the true effect can be decomposed as follows:

$$\begin{aligned} \textit{Bias} &= [a^*(1, h^T) - a^*(0, h^T)] - [a(1) - a(0)] \\ &\approx \underbrace{\Phi(1)\big(E[h \mid h^T, h^L(1)] - E[h \mid h^T, h^L(0)]\big)}_{\text{Bias due to beliefs}} + \underbrace{\big(\Phi(1) - \Phi(0)\big) E[h \mid h^T, h^L(0)]}_{\text{Bias due to "responsiveness"}}. \end{aligned}$$

The first term captures differences in beliefs between the treatment and control environments, for example because they induce differences in latent demand. The second captures differences in behavioral responsiveness, given beliefs, for example

[19] We thank the editor, Stefano DellaVigna, as well as an anonymous referee for suggesting this line of inquiry.

because the treatment and control groups are at different locations on the cost of effort function.[20]

Fully Informative Demand Treatments.—Importantly, in the special case where researchers are willing to assume that demand treatments are fully informative ($p^T = 1$), we can eliminate the bias due to beliefs: if h^T is fully informative, $E[h|h^T, h^L(1)] = E[h|h^T, h^L(0)] = 1$ or -1. We are left with the bias due to differences in responsiveness. We can then ask whether this bias is important, by testing for differences in sensitivity between treatment and control (an interaction effect):[21]

$$\underbrace{\big[a^*(1,1) - a^*(1,-1)\big]}_{\text{Sensitivity } (\zeta=1)} - \underbrace{\big[a^*(0,1) - a^*(0,-1)\big]}_{\text{Sensitivity } (\zeta=0)} \approx 2\big(\Phi(1) - \Phi(0)\big).$$

If this term is small, we can obtain a point estimate of the demand-free treatment effect by comparing behavior on two same-signed demand treatment, essentially we are "controlling for" the influence of demand.

If sensitivity differs significantly between treatment and control, we can still approximate the treatment effect by averaging the estimates obtained with two positive and two negative demand treatments:

$$0.5\Big(\big[a^*(1,1) - a^*(0,1)\big] + \big[a^*(1,-1) - a^*(0,-1)\big]\Big) \approx a(1) - a(0).$$

This approach is equivalent to estimating the treatment effect from the midpoints of the bounds for the treatment and control groups. It relies on the symmetry of the first-order Taylor approximation.

Less Informative Treatments.—Alternatively, researchers might wish to use same-signed weaker demand treatments to align beliefs among participants, without requiring $p^T = 1$. In general this will not eliminate bias entirely, but we can derive conditions under which the bias will be reduced. Since differences in responsiveness will no longer be testable, we focus on the prospect of reducing the bias due to beliefs, which will be sufficient if variation in responsiveness between treatments is small.[22] We find that when the latent demand biases have opposite signs ($h^L(1) = -h^L(0)$, which is the typical scenario that concerns researchers) our Bounding assumption is sufficient for two same-signed demand treatments to reduce the bias due to beliefs. When the latent demand biases have the same sign ($h^L(1) = h^L(0)$), same-signed demand treatments that *reinforce* latent demand (i.e., $h^T = h^L(1)$)

[20] In some settings it may be possible to sign the bias due to responsiveness. If demand treatments are applied, and bounding holds, the sign of $E[h|h^T, h^L(0)]$ is known and equal to the sign of h^T. Knowledge of the shape of v can then help us to sign $\Phi(1) - \Phi(0)$. For example, in the real effort case, we expect responsiveness to decrease as effort increases, due to the curvature of the cost of effort function. That implies $\Phi(1) - \Phi(0) < 0$, in which case the bias due to responsiveness is negative when positive demand treatments are used.

[21] Or, equivalently, testing whether the treatment effect estimate differs when two positive versus two negative demand treatments are used.

[22] In other words, we ask when $\big|E[h|h^T, h^L(1)] - E[h|h^T, h^L(0)]\big| < \big|E[h|h^L(1)] - E[h|h^L(0)]\big|$, for $h^T \in \{-1, 1\}$.

always reduce bias. Sufficiently strong opposite-signed treatments reduce bias, but Bounding is not enough to guarantee this.

In summary, the Bounding assumption covers all cases except where the demand effects in treatment and control agree with one another and disagree with the demand treatments used. To apply this approach, therefore, researchers may need to use judgment about the likely sign of demand effects in their experiment, or report a range of estimates.

Applications.—We apply the approaches developed above to our effort experiment in online Appendix Table A1. For the strong demand treatments, where we have argued $p^T = 1$ is not an unreasonable assumption, we see large and statistically significant differences in sensitivity between the 0 and 1 treatment groups, so we instead apply the "midpoint" technique. For the weak demand treatments, we report treatment effect estimates for both positive-positive and negative-negative demand treatment applications. Encouragingly, the estimates are all quite similar to one another, lying within 10 percent of the conventional treatment effect estimate.

E. *Structural Estimates*

Under further assumptions, strong demand treatments permit structural estimation of demand-free model parameters (v), as well as ϕ and $E\left[h \mid h^L\right]$. Knowing v allows the researcher to make predictions about behavior absent experimenter demand. Knowing ϕ allows them to quantify the importance of experimenter demand. Measuring beliefs can enable them to diagnose and eliminate the sources of latent demand effects. We illustrate how structural estimation can be performed using the real effort experiment. Because our model simply nests that of DellaVigna and Pope (2018)—henceforth, DP—we follow their approach to structural estimation.[23]

DP estimate the following utility function (expressed in our notation):

$$v(a) = (s + \zeta)\, a - c(a). \tag{9}$$

The action a is effort, measured in points on the task, s is an intrinsic motivation parameter (workers may exert effort because they enjoy the task), and $c(a)$ is a cost of effort function. We assume the environment enters v only via the piece rate, so let $\zeta \in \{0, 1, 4\}$ be a real number. DP solve the first-order condition and estimate the model parameters using nonlinear least squares (NLLS).[24]

Adding demand to this utility function gives

$$U(a, \zeta) = (s + \zeta + \phi(\zeta)\, E[\, h \mid h^T, h^L(\zeta)\,]\,)\, a - c(a) \tag{10}$$

with corresponding first-order condition

$$s + \zeta + \phi(\zeta)\, E[\, h \mid h^T, h^L(\zeta)\,] - c'(a^*(\zeta)) = 0. \tag{11}$$

[23] We note that the structural analysis was not included in our pre-analysis plan.
[24] They also employ a minimum distance procedure. We stick to NLLS for brevity.

DP consider two alternative forms for c: first, a power function $c(a) = ka^{1+\gamma}/(1+\gamma)$, yielding optimal effort equal to

$$a^*(\zeta) = \left(\frac{s+\zeta+\phi(\zeta)E[h|\,h^T,h^L(\zeta)]}{k}\right)^{\frac{1}{\gamma}}. \tag{12}$$

Second, an exponential form $c(a) = k\exp(\gamma a)/\gamma$, with effort level

$$a^*(\zeta) = \frac{1}{\gamma}\log\left(\frac{s+\zeta+\phi(\zeta)E[h|\,h^T,h^L(\zeta)]}{k}\right). \tag{13}$$

We have seven treatment groups in total: neutral treatments with piece rates equal to 0 cents, 1 cent, and 4 cents per 100 points on the task; and positive and negative strong demand treatments in the 0 and 1 cent groups.[25] Noting that $E[h|h^L(\zeta)] = p^L(\zeta)h^L(\zeta) \in (-1,1)$, we can treat it as a single parameter whose sign identifies h^L and whose magnitude identifies $p^L(\zeta)$. This leaves us with 10 parameters: s, k, γ, $\phi(0)$, $\phi(1)$, $\phi(4)$, $p^L(0)h^L(0)$, $p^L(1)h^L(1)$, $p^L(4)h^L(4)$, and p^T, so we need to impose some further restrictions.

First we assume that ϕ is fixed: $\phi(0) = \phi(1) = \phi(4) = \phi$, eliminating two parameters. In other words, varying incentives do not change the participants' desire to please the experimenter. Second, as in the previous section, we assume $p^T = 1$, which implies that $E[h|h^T,h^L] = h^T$. By assumption this is not justified for our weak demand treatments, so we focus on the strong treatments. We are left with seven parameters, s, k, γ, ϕ, $p^L(0)h^L(0)$, $p^L(1)h^L(1)$, and $p^L(4)h^L(4)$, and are therefore exactly identified. We additionally estimate a specification in which we restrict latent demand to depend only on whether monetary incentives are present, i.e., $p^L(1)h^L(1) = p^L(4)h^L(4)$.

While we use the same model as DP, identification comes from a different source. Under the assumption of no latent demand (as in DP), s, γ, and k are identified from the three *neutral treatment* groups. When latent demand is present, the model parameters (s, γ, k, ϕ) are identified from the *demand treatment* groups; with these in hand the neutral treatments allow us to back out the beliefs $p^L(\zeta)h^L(\zeta)$.

Full details of the estimation procedure, which follows DP, are provided in online Appendix Section B.B9. We estimate equation (12) in logs, and equation (13) in levels. Estimation results are presented in Table 4. Columns 1–3 correspond to the power cost function, and columns 4–6 to the exponential cost function. In each case we first mirror DP by estimating s, γ, and k using only the neutral treatments, assuming that there is no latent demand.[26] Second, we include all treatment groups and impose that latent demand depends only on whether monetary incentives are present ($p^L(1)h^L(1) = p^L(4)h^L(4)$). Third, we allow latent demand to differ across all

[25] We also collected data using weak demand treatments, but we do not use it in this analysis a) because it was collected in a separate experiment and b) because for estimation we need to impose the parameter restriction $p^T = 1$, which we do not believe is satisfied in the weak treatments.

[26] There are some differences between our parameter estimates and DP's earlier work, which may reflect changes in the participant pool over time.

TABLE 4—STRUCTURAL ESTIMATES

	Power cost of effort			Exponential cost of effort		
	log count (1)	log count (2)	log count (3)	Count (4)	Count (5)	Count (6)
ϕ		0.175	0.249		0.205	0.300
		(0.092)	(0.095)		(0.079)	(0.066)
$h^L(0)p^L(0)$		−0.735	−0.516		−0.525	−0.187
		(0.172)	(0.303)		(0.191)	(0.249)
$h^L(>0)p^L(>0)$		−0.609			0.849	
		(2.194)			(1.799)	
$h^L(1)p^L(1)$			−0.473			0.155
			(1.110)			(0.694)
$h^L(4)p^L(4)$			−6.508			−6.600
			(3.360)			(1.963)
s	0.034	0.179	0.273	0.031	0.229	0.493
	(0.051)	(0.095)	(0.126)	(0.046)	(0.096)	(0.208)
k	4.7e-26	7.5e-24	6.5e-17	4.2e-08	2.1e-06	1.8e-04
	(3.1e-25)	(2.9e-23)	(3.1e-16)	(1.8e-07)	(3.7e-06)	(2.9e-04)
γ	7.260	6.583	4.433	6.5e-03	4.6e-03	2.3e-03
	(2.216)	(1.303)	(1.707)	(2.1e-03)	(8.7e-04)	(8.2e-04)
Observations	727	1,691	1,691	727	1,691	1,691
R^2	0.122	0.166	0.166	0.167	0.204	0.206

Notes: This table uses data from the the real effort experiment on MTurk with strong demand treatments. Coefficients s and ϕ are measured in cents. s measures the respondents' intrinsic motivation. ϕ measures the monetary value of acting according to the experimental objective. γ is the effort cost curvature and k is the scaling parameter. $h^L(\zeta)p^L(\zeta)$ latent demand in incentive condition ζ. $h^L(>0)p^L(>0)$ in the combined 1-cent and 4-cent incentive conditions. Robust standard errors in parentheses.

three incentive levels. Coefficients s and ϕ are measured in cents per 100 points. Therefore, $s = 1$ is interpreted as intrinsic motivation playing an equivalent role to an incentive of 1 cent per 100 points.

Our main finding is a nontrivial preference for pleasing the experimenter. Our estimates of ϕ take values in the range 0.2–0.3 and are similar across specifications. A value of 0.2 implies that moving from complete uncertainty ($E[h|h^L] = 0$) to complete certainty that high effort is desired ($E[h|h^L] = 1$) increases effort as much as increasing the incentive by 0.2 cents per 100 points.

Our estimates of $E[h|h^L]$ are mostly negative, consistent with latent demand decreasing effort. However, the estimates are noisy and typically not significantly different from zero. We estimate that in the 4 cent treatment, $E[h|h^L(4)] \approx -6.5$, while the theory requires $E[h|h^L(4)] \in (-1, 1)$ (we note that the estimate is noisy and -1 lies well within the 95 percent confidence interval). This most likely reflects the fact that our demand treatments were only applied to the 0 and 1 cent treatment groups, so the effort cost function must be extrapolated far out of sample to estimate beliefs for the 4 cent group. We provide further discussion on this point, and an illustrative figure, in online Appendix Section B.B9.

IV. Properties of Demand Effects

In this section, we examine some of the properties of demand effects and the assumptions underlying our approach. We begin with a discussion of Monotonicity, examining whether it holds first on average and then at the individual level. Second we turn to the central mechanism that drives behavior in the model: changes in beliefs due signals about demand. Third, we consider the Bounding assumption. Although we cannot test it directly (since natural actions are not observed), we show that our bounds seem reasonable given existing evidence on responsiveness to a particular design feature—anonymity in the dictator game—that has been argued to potentially induce variation in demand. Fourth, we study heterogeneity in sensitivity to our demand treatments, focusing on four dimensions: incentives, gender, attention, and participant pool. These are cases where we might expect our Monotone Sensitivity assumption to hold, such that variation in sensitivity is informative about underlying variation in latent demand. Fifth, we examine the effect of our demand treatments on the variance and full distribution of actions.

A. *Monotonicity*

Monotonicity on Average.—Our first theoretical assumption is Monotonicity: $a^{+}(\zeta) \geq a^{L}(\zeta) \geq a^{-}(\zeta)$. Panel C of Table 1 and panel C of Table 2 examine this assumption for the subset of tasks in which we collected data without applying demand treatments.[27] We estimate the following equation using the incentivized MTurk respondents, in which POS_i and NEG_i are dummy variables for the positive and negative demand treatments, and the no-demand condition is the reference group:

$$ZY_i = \pi_0 + \pi_1 POS_i + \pi_2 NEG_i + \varepsilon_i. \tag{14}$$

We find strong support for Monotonicity in average actions. The strong demand treatments always moved average actions in the intended direction, and in most cases the differences are statistically significant. We find a significant negative response to negative weak demand in the investment game, and a significant positive response to weak positive demand in the dictator game. Responses to the positive demand treatment in the investment game and the negative demand treatment in the dictator game have the wrong signs, but are close to zero and not statistically significant. Finally, our data from the representative sample are fully consistent with Monotonicity for both the weak and strong treatments (see online Appendix Table C18).

Testing Monotonicity Within-Person.—Our seventh experiment uses a within-participant design, collecting data on behavior first without, and then with a demand treatment. This allows us to examine Monotonicity directly at the individual level, and identify defiers, who try to do the opposite of the experimenter's wishes.

[27] We have data for the dictator and investment games with weak and strong treatments, plus convex time budgets and real effort with only the strong treatments. Because the weak and strong treatments were applied in separate experiments, we analyze the data separately.

Intuitively, by observing who increases and who decreases their action in response to a positive demand treatment, we can identify who is a complier and who is a defier. As discussed in Section ID, "too much" defiance can invalidate our bounds.

The design is as follows. MTurk participants were told that they would complete two tasks, and be paid according to one of them, selected by chance. One-half played the dictator game twice, and one-half the investment game twice. They first completed the task without any demand treatment, then again with the addition of a strong positive or negative demand treatment. We thus have four groups, split by dictator/investment game and positive/negative demand.

The model implies a simple interpretation of the data. Participants observe the first task, form a belief about h, and make a choice. They then observe the second task with the demand treatment, update their belief, and make a new choice. Strict compliers, with $\phi > 0$, will increase their action relative to task 1, strict defiers with $\phi < 0$ will decrease it, and those with $\phi = 0$ should take the same action in both tasks.[28]

Our main findings are captured by Figure 3, which plots actions from tasks 1 and 2. In the positive demand treatments, strict compliers lie above the 45-degree line, strict defiers lie below, and those who did not change their action lie on the line. Only about 5 percent of our respondents are strict defiers. About 30 percent do not change their behavior in response to our demand treatments, while the remaining 65 percent strictly comply with our demand treatments (proportions are similar across tasks). Thus, we find very little evidence of defiance.

Online Appendix Table A2 presents mean actions and sensitivities estimated from the within design and the equivalent objects from the earlier between-participants experiments. For the within experiment, "no demand" cells are computed from task 1, while demand treatment cells and sensitivities from task 2. The sensitivities are quantitatively very similar in the between and within designs. This is encouraging, as it suggests researchers can simply and cheaply obtain bounds using within-participant demand treatments, avoiding the need to recruit additional participants to apply our method.

Within-participant data can be used to construct "defier-corrected" bounds.[29] These, with confidence intervals, are displayed in online Appendix Table A6. They are almost identical to the conventional bounds, reflecting the low rate of defiance, and giving further comfort that defiance is quantitatively unimportant. Table A5 reports raw actions separately for compliers and defiers.

[28] The within design might fail to perfectly classify respondents, for two reasons. First, the theory assumes that ζ, and therefore the natural action, $a(\zeta)$, is independent of the demand treatment, h^T. This is a strong assumption in our within design, because it is clear that the response to h^T is part of the analysis, which could change participants' interpretation of ζ. However, if participants infer that our interest is in showing people respond to our demand treatments, compliers would increase and defiers to decrease their actions, in which case we would still arrive at the correct classification. Second, it might matter that participants have made a prior choice, either out of a concern for consistency (reducing responsiveness to our demand treatments) or a motive to conceal their defier/complier identity.

[29] For defiers, $a(\zeta) \in [a^+(\zeta), a^-(\zeta)]$, so if the proportion of compliers is c, the natural action lies in the interval $[cE[a^-(\zeta) \mid \phi \geq 0] + (1-c)E[a^+(\zeta) \mid \phi < 0], cE[a^+(\zeta) \mid \phi \geq 0] + (1-c)E[a^-(\zeta) \mid \phi < 0]]$. In practice, one simply inverts the demand treatment variable for participants identified as defiers and computes bounds as before. The construction of defier-corrected bounds was not included in our pre-analysis plan.

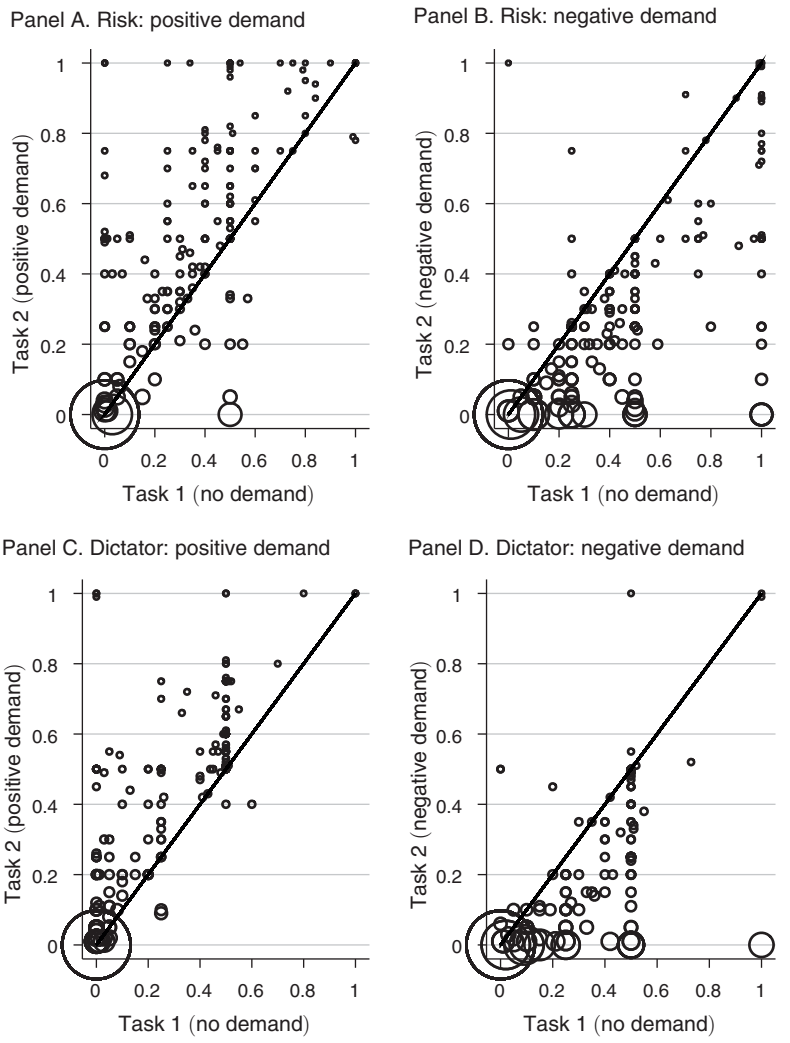

Figure 3. Measuring Defiance through a Within-Participant Design

Notes: This figure uses MTurk data from experiment 7 and displays the scatterplot of responses in task 1 ("no demand" condition) and task 2 (demand condition). Points above the 45-degree line indicate an increase in the action, and points below the 45-degree line a decrease. The size of the rings is proportional to the number of observations.

B. *Beliefs*

The core mechanism in our model is that participants form beliefs about the experimenter's objective in response to implicit or explicit signals. We examine this assumption with simple, unincentivized belief data collected after participants had completed their experimental task. The purpose of the measures was a manipulation check, to ascertain that participants' beliefs responded as expected to the demand treatments. We asked two questions: "What do you think is the result that the researchers of this study want to find?"; and "What do you think was the hypothesis of this research study?" Responses were binary: participants could respond that

they thought the objective/hypothesis was either a high or low action.[30] We assume that participants report a high belief if their posterior ($E[h|h^L]$ or $E[h|h^T, h^L]$) is positive, and a low belief if negative, so the average response tells us the fraction of participants with high beliefs.

Results for incentivized MTurk respondents are presented in online Appendix Tables A8 and A9. They confirm that our treatments moved average responses in the anticipated direction. Overall, the levels of beliefs and magnitudes of shifts in beliefs are similar for the strong and weak treatments, i.e., both were equally successful in fixing the sign of beliefs. In the theory, strong and weak treatments are equally effective at fixing the sign of beliefs if $p^T \geq p^L$, but stronger treatments lead to more extreme posteriors.[31]

C. *Comparison of Effect Sizes*

Is the bounding assumption reasonable? Although it is not directly testable, we compare our bounds to previous manipulations that have been hypothesized to induce demand effects. Our examples all come from the dictator game and include four studies that varied participants' degree of anonymity, and a study in Sierra Leone that varied the presence of a white foreigner.[32] We present effect sizes from these experiments and our own in online Appendix Table A11.

Sensitivity to our weak treatments (a 17 percent reduction in giving under negative versus positive demand) is very close to the average effect size across these 5 studies (around a 21 percent reduction in giving in response to treatment), and our strong treatments comfortably bound this average (a 42 percent reduction). Considering individual studies, our weak bounds are close in magnitude to those from Bolton, Katok, and Zwick (1998); Barmettler, Fehr, and Zehnder (2012); and Cilliers, Dube, and Siddiqi (2015); but smaller than those from Hoffman et al. (1994) and Hoffman, McCabe, and Smith (1996). These two studies in particular, however, have been criticized for inducing potentially strong experimenter demand (Loewenstein 1999), so may represent a scenario where the more conservative strong bounds are preferable. Their effect sizes are close to or a bit larger than (and not significantly different from) our strong bounds.

This exercise is of course only suggestive, since responses in these studies include direct effects of anonymity on behavior, as well as potential experimenter demand. Additionally, the studies we consider were conducted in the laboratory and differ in various other ways from our online setting. The results are nevertheless encouraging, in particular that our weak bounds seem to perform quite well.

[30] One could collect richer belief measures and incentivize responses, but asking for fine-grained beliefs about *our own* motivations seemed quite unnatural, particularly as there was no objective truth against which to score. Our measures may of course be subject to their own demand bias.

[31] $p^T \geq p^L$ also implies that *all* participants' beliefs should have the "correct" sign following a demand treatment. Not all of our participants reported correct beliefs following a demand treatment. This could be due to measurement error in our belief data, or, as we discuss in online Appendix Section B.B3, participants might be inattentive to our demand treatments. If they are also inattentive to latent demand signals such participants do not threaten Bounding.

[32] Hoffman et al. (1994); Hoffman, McCabe, and Smith (1996); Bolton, Katok, and Zwick (1998); and Barmettler, Fehr, and Zehnder (2012) study the effect of "double blind" versus "single blind" anonymity in dictator games. To our knowledge, this is the complete set. Cilliers, Dube, and Siddiqi (2015) study the effect of white foreigner presence.

D. *Heterogeneity*

Does sensitivity to demand treatments vary by design and participant characteristics? Here, we examine heterogeneous responses to our strong and weak demand treatments on four pre-specified dimensions: by whether choices are incentivized or hypothetical; gender; attentiveness; and participant pool (MTurk versus representative online panel). Whether or not this heterogeneity can be interpreted as informative about differences in underlying latent demand depends upon whether Monotone Sensitivity holds for the environments under consideration, i.e., whether they belong to the same comparison class. We show in online Appendix Section B.B3 that variation in incentives, attention, and the preference for pleasing the experimenter, ϕ (which may differ by gender or participant pool), form valid bases for comparison classes.

Incentivized versus Hypothetical Choices.—In MTurk experiments 1 and 2 we randomly assigned participants to make either hypothetical or incentivized choices. In theory, we would expect higher sensitivity in hypothetical choice, as the cost of deviating from the natural action is lower. To test this prediction, we regress standardized actions on a dummy, POS_i, taking value 1 for the positive demand treatment and 0 for the negative treatment, a dummy indicating incentivized choice M_i, and their interaction:

$$ZY_i = \beta_0 + \beta_1 POS_i + \beta_2 M_i \times POS_i + \beta_3 M_i + \varepsilon_i. \tag{15}$$

Results for the weak and strong demand treatments are presented in Table 5. Interestingly, participants making hypothetical or incentivized choices responded very similarly to experimenter demand, in each task and on average, and if anything sensitivity is slightly higher when incentivized.

Relatedly, we ask how sensitivity differs when we increase performance pay in the effort task. Reasonable assumptions would imply sensitivity is decreasing in performance pay (see online Appendix B.B3). Table 2 shows that sensitivity to our strong treatments was around 3.5 times higher when effort was unincentivized, as predicted. We do not see the same pattern under the weak treatments, though this may simply reflect the fact that sensitivity to these treatments was low.

The mixed evidence on responsiveness to incentives is somewhat surprising. One possibility is that our incentivized choices still involve relatively low stakes, and that we would see a difference at higher stakes. Additionally, the theory allows ϕ to depend upon ζ, and another possibility is that raising the stakes also raises participants' desire to please the experimenter (e.g., due to reciprocity). We see this as an interesting avenue for future work. Our results relate to previous work examining the effects of incentives on behavior in the lab (Camerer et al. 1999).

Gender and Attention.—We measure self-reported gender in all tasks on MTurk and in the representative panel, and attentiveness in all tasks except the effort task (since DP did not measure this variable). We define a participant as attentive if they

Table 5—Heterogeneity in Response to Weak and Strong Demand Treatments (*z-Scored*)

	All games	Time	Risk	Ambiguity aversion	Effort 0 cent bonus	Effort 1 cent bonus
Panel A. Weak: design characteristics						
Sensitivity × incentive	0.073		0.149			
	(0.085)		(0.125)			
Observations	1,963		970			
Panel B. Weak: respondent characteristics						
Sensitivity × male	0.038	−0.028	0.057	0.069	0.305	0.033
	(0.061)	(0.174)	(0.179)	(0.203)	(0.239)	(0.236)
Observations	4,450	422	473	390	388	381
Sensitivity × attention	0.119	−0.402	−0.077	0.434		
	(0.116)	(0.395)	(0.307)	(0.504)		
Observations	3,681	422	473	390		
Sensitivity × representative sample	0.032		0.032			
	(0.084)		(0.127)			
Observations	2,125		1,041			
Panel C. Strong: design characteristics						
Sensitivity × incentive	−0.007	−0.063	0.196			
	(0.080)	(0.132)	(0.116)			
Observations	2,989	994	996			
Panel D. Strong: respondent characteristics						
Sensitivity × male	−0.152	−0.212	−0.090	−0.382	0.075	0.005
	(0.064)	(0.168)	(0.160)	(0.192)	(0.197)	(0.214)
Observations	4,800	491	482	404	492	472
Sensitivity × attention	0.117	0.319	0.471	−0.276		
	(0.140)	(0.393)	(0.401)	(0.414)		
Observations	3,836	491	482	404		
Sensitivity × representative sample	0.027		−0.121			
	(0.081)		(0.118)			
Observations	2,184		1,070			

	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust game 1	Trust game 2
Panel A. Weak: design characteristics						
Sensitivity × incentive		−0.002				
		(0.115)				
Observations		993				
Panel B. Weak: respondent characteristics						
Sensitivity × male	0.060	0.029	−0.089	0.003	0.257	−0.239
	(0.189)	(0.146)	(0.192)	(0.185)	(0.230)	(0.229)
Observations	412	515	360	411	352	346
Sensitivity × attention	0.226	0.585	0.094	0.368	0.116	−0.398
	(0.305)	(0.328)	(0.296)	(0.230)	(0.362)	(0.301)
Observations	412	515	360	411	352	346
Sensitivity × representative sample		0.032				
		(0.110)				
Observations		1,084				
Panel C. Strong: design characteristics						
Sensitivity × incentive		0.072				
		(0.121)				
Observations		999				
Panel D. Strong: respondent characteristics						
Sensitivity × male	−0.223	−0.201	−0.137	−0.144	0.098	−0.361
	(0.217)	(0.153)	(0.187)	(0.201)	(0.216)	(0.240)
Observations	365	511	409	421	382	371
Sensitivity × attention	0.255	−0.024	−0.272	0.229	0.918	−0.091
	(0.358)	(0.530)	(0.394)	(0.538)	(0.409)	(0.311)
Observations	365	511	409	421	382	371
Sensitivity × representative sample		0.176				
		(0.112)				
Observations		1,114				

Notes: Outcome variables are *z*-scored at the task level. Panels A and C display heterogeneous treatment effects by design characteristics, i.e., whether choices are incentivized or hypothetical. Panels B and D display heterogeneous treatment effects by respondent characteristics: gender, attention, and population. "Male" equals 1 for males, "attention" equals 1 if the respondent passed the attention screener, "representative sample" equals 1 for representative sample respondents.

passed an attention screener at the beginning of the task.[33] We estimate the following equation:

$$ZY_i = \beta_0 + \beta_1 POS_i + \beta_2 H_i + \beta_3 H_i \times POS_i + \varepsilon_i, \tag{16}$$

where H_i is the dimension of heterogeneity of interest.

As can be seen in Table 5, we find that women respond more strongly to the strong demand treatments than men, with sensitivity around 0.15 standard deviations higher, but no significant difference for the weak treatments (where overall sensitivity and thus statistical power is lower). We interpret the evidence as suggestive of greater desire to please the experimenter among women, which relates to the literature on gender differences in preferences (Croson and Gneezy 2009).

Turning to attention, only 10 percent of MTurk respondents failed our screener, so we have little power to detect differences in sensitivity. Table 5 shows higher sensitivity (around 0.12 standard deviations) to our weak and strong manipulations among attentive participants, but these effects are not significant.[34]

In the representative online panel we find significantly higher sensitivity among women, and among attentive participants (see online Appendix Section C.C4). Approximately 65 percent failed the screener, increasing our power here.

MTurk versus Representative Online Panel.—Some researchers are concerned that MTurk workers are experienced research participants and may behave differently than a more representative participant pool. In addition, MTurkers need to maintain a high work "acceptance" rating and may therefore be especially motivated to please the researcher (Berinsky, Huber, and Lenz 2012). To address such concerns, and to test an additional dimension of heterogeneity, we replicated the MTurk dictator game and investment game experiments with respondents from a representative online panel, whose participants are less experienced in the types of tasks we consider. We used both weak and strong demand treatments, or no demand treatment. All choices were incentivized at the same stakes as in the MTurk experiments.[35]

[33] We use the screener developed by Berinsky, Margolis, and Sances (2014). It presents participants with a paragraph of text that appears to direct them to select their preferred online news sources from a list, but concealed in the text is an instruction to instead choose two specific options. The assumption is that attentive respondents read the question and follow the concealed instruction, while inattentive respondents do not. Passing the attention check is weakly positively correlated with previous completion of MTurk tasks, so we also consider heterogeneity using a representative online panel whose respondents are generally less experienced and are unlikely to have seen the screener before. Moreover, there is little variation in sensitivity by experience; results are available on request.

[34] Our pre-analysis plans specified that these heterogeneity tests would be conducted at the experiment level, rather than averaged across all tasks within demand treatments. We perform these tests in online Appendix C. Experiment 1 (strong treatments) finds higher sensitivity for women ($p = 0.10$) and attentive participants ($p = 0.10$). Experiment 2 (weak treatments) finds slightly higher sensitivity for men ($p = 0.25$) and attentive participants ($p = 0.53$). Experiment 3 (effort, strong treatments) finds almost identical sensitivity for men and women ($p = 0.95$).

[35] Respondents in the online panel were incentivized with \$1 stakes in the panel currency, which they can use to buy products in the survey provider's online store. We discovered after the study that, while some of the products in the store have a value equivalent to \$1, others have lower value. This means that the effective stake size in the representative online panel may have been lower than on MTurk. Since we find no differences in response to demand treatments depending on whether choices are incentivized or hypothetical on MTurk, we do not expect this to be an important concern.

Table 5 tests for differences in sensitivity between MTurk and representative survey participants, pooling tasks and for each task separately.[36]

Representative panel participants responded very similarly to MTurk participants, with sensitivity on average 0.03 standard deviations higher (not significant) under both weak and strong treatments. There are some small differences in sensitivity to the strong treatments at the game level (significant at 10 percent for the dictator game), but little evidence of systematic differences between participant pools.

E. *Demand and the Distribution of Actions*

We have focused on analysis of mean behavior, but other moments may respond to our demand treatments. For example, by aligning beliefs, they might reduce the variance of observed actions. Online Appendix Table A12 shows that variance is very similar and in most cases slightly lower under the demand treatments relative to no demand. Online Appendix Figures A2 and A3 plot the cumulative distribution of actions for each task and demand treatment, showing that the demand treatments shift the full distribution of behavior. Encouragingly, these shifts seem to almost always satisfy first-order stochastic dominance, consistent with monotonicity.

V. Using the Method in Practice

We now provide some practical guidance on using the methods developed in this paper. First, we discuss settings in which demand treatments can be employed. Second, many of the applications in this paper have been to "levels" of behavior, so we list a few examples of other cases where one might be specifically interested in bounding levels. Third, we summarize the set of techniques and recommendations we have developed. Online Appendix B.B10 uses a diagram to work through an example of each technique.

We have two main settings in mind for applications. First, demand treatments can be applied in *experiments* in the laboratory, online, or in the field. We expect their primary use will be for the various robustness checks and estimation procedures we have outlined, but they can also be used for studying demand effects themselves. A natural next step in this agenda would be to compare demand sensitivity in the lab and online, which may differ due to differences in attentiveness or social interaction with the experimenter. Second, they can readily be applied in *surveys*. Our estimates from hypothetical dictator games, convex time budgets, and investment games, which are commonly used as survey questions, show that reasonable bounds are obtained even when choices are not incentivized. Applications include standalone surveys (e.g., on political views, inflation expectations, labor market outcomes) or field experiments, which often rely on survey data. For instance, participants might be told: "The researchers expect respondents who received the intervention (e.g., cash, bednets, education) to report more favorable outcomes."

While the majority of experiments are aimed at estimating treatment effects, researchers are often interested in mean responses in both surveys and experiments,

[36] Our pre-analysis plan specified the test pooled across the strong and weak demand treatments: we perform this test in online Appendix C.C4 and find no significant difference.

Table 6—Overview of Experiments

Experiment	Sample	Tasks	Demand treatments	Real or Hypothetical
Experiment 1 (May 18, 2016–May 30, 2016)	MTurk ($N = 4{,}479$)	Dictator game, investment game, and convex time budgets	Strong positive demand, strong negative demand and no-demand treatment	Both real stakes and hypothetical choices
Experiment 2 (July 5, 2016–July 25, 2016)	MTurk ($N = 2{,}950$)	Dictator game and investment game	Weak positive demand, weak negative demand and no-demand treatment	Both real stakes and hypothetical choices
Experiment 3 (Aug. 26, 2016–Aug. 27, 2016)	MTurk ($N = 1{,}691$)	Effort experiment with 1 cent bonus and Effort experiment with no bonus. Also effort experiment with 4 cent bonus (no demand treatments were applied to this group)	Strong positive, strong negative and no-demand treatment	Real stakes (real effort experiment)
Experiment 4 (Aug. 18, 2016–Sep. 1, 2016)	Research now representative online panel ($N = 2{,}933$)	Dictator game and investment game	Strong positive demand, strong negative demand, weak positive demand, and weak negative demand, and no-demand treatment	Real stakes
Experiment 5 (Sep. 12, 2016–Sep. 20, 2016)	MTurk ($N = 5{,}045$)	Trust game (first and second mover), ultimatum game (first and second mover), lying game, ambiguous investment game, and convex time budgets	Strong positive demand, strong negative demand, weak positive demand, and weak negative demand	Real stakes
Experiment 6 (Sep. 19, 2016–Sep. 20, 2016)	MTurk ($N = 769$)	Effort experiment with 1 cent bonus, Effort experiment with no bonus	Weak positive demand and weak negative demand	Real stakes (real effort experiment)
Experiment 7 (May 18, 2017–May 20, 2017)	MTurk ($N = 999$)	Dictator game and investment game	Within design: Task 1: no demand treatment; Task 2: strong positive demand or strong negative demand	Real stakes

Notes: This table summarizes the key design features of each of the experiments. In experiment 5 for convex time budgets we only employ the weak demand treatments, while for all other tasks in experiment 5 we employ both strong and weak demand treatments.

and might be concerned about robustness. We provide a few examples. Policymakers might be intrinsically interested in levels of policy views about taxation or immigration; beliefs about these objects; willingness to contribute to public goods; inflation or growth expectations; consumption plans; or time use. In the lab, we are often interested in the level of giving in dictator games; offers and frequency of rejections in ultimatum games; competitiveness of specific subpopulations (e.g., men versus women); the amount of lying in coin-flip games; or the degree of risk or ambiguity aversion (e.g., for calibrating models).

TABLE 7—DETAILS OF EXPERIMENTAL TASKS

Task	Dictator game	Investment game	Ambiguous investment game
Description	Choose to split money between yourself and another participant	Choose how much to invest in a risky project	Choose to how much to invest in an uncertain project
Sample	Strong: MTurk between ($N = 1{,}503$); MTurk within (998); Rep. Sample ($N = 899$); Weak: MTurk between ($N = 1{,}477$); Rep. Sample ($N = 865$)	Strong: MTurk between ($N = 1{,}492$); MTurk within (1,000); Rep Sample ($N = 902$); Weak: MTurk between ($N = 1{,}473$); Rep. Sample ($N = 883$)	Strong: MTurk ($N = 404$); Weak: MTurk ($N = 390$)
Show up fee	\$0.25	\$0.25	\$0.25
Action space	$a \in [0, 1]$ is fraction of \$1 sent to the other participant	$a \in [0, 1]$ is fraction of \$1 invested	$a \in [0, 1]$ is fraction of \$1 invested
Payoffs	Amount not sent $(1 - a)$	"give more (less) to the other participant than you normally would"	\$ $(1 - a) + 3a$ with probability 0.4, \$ $(1 - a)$ otherwise
Strong demand treatment: "You will do us a favor if you"	"give more (less) to the other participant than they normally would"	"invest more (less) than they normally would"	"invest more (less) than they normally would"
Weak demand treatment: "We expect that participants who are shown these instructions will"	"give more (less) to the other participant than they normally would"	"invest more (less) than you normally would"	"invest more (less) than they normally would"

Task	Convex time budgets	Effort: no bonus	Effort: 1-cent bonus
Description	Choose between receiving money today versus money in seven days	Alternately press the a and b button without receiving any bonus	Alternately press the a and b button, receiving 1 cent per 100 points
Sample	Strong: MTurk ($N = 1{,}484$); Rep. Sample ($N = 899$); Weak: MTurk ($N = 422$)	Strong: MTurk ($N = 731$); Weak: MTurk ($N = 388$)	Strong: MTurk ($N = 714$) Weak: MTurk ($N = 381$)
Show up fee	\$0.25	\$1	\$1
Action space	$a \in [0, 1.2]$ is the amount to be received in 7 days	$a \in [0, 4{,}000]$ is number of a-b button presses	$a \in [0, 4{,}000]$ is number of a-b button presses
Payoffs	\$ $(1 - a)/1.2$ is received within 24 hours, and \$ a is received in 7 days	No payoffs beyond show-up fee	1 cent per 100 button presses
Strong demand treatment: "You will do us a favor if you"	"choose to receive more (less) in seven days than you normally would"	"work harder (less hard) than you normally would"	"work harder (less hard) than you normally would"
Weak demand treatment: "We expect that participants who are shown these instructions will"	"choose to receive more (less) in seven days than they normally would"	"work harder (less hard) than they normally would"	"work harder (less hard) than they normally would"

(*Continued*)

TABLE 7—DETAILS OF EXPERIMENTAL TASKS (*Continued*)

Task	Trust game first mover	Trust game second mover	Ultimatum game first mover
Description	Choose to send an amount of money to the other player	Choose to send back some money to the other player; (strategy method)	Offer a split to the other player
Sample	Strong: MTurk ($N = 382$); Weak: MTurk ($N = 352$)	Strong: MTurk ($N = 371$); Weak: MTurk ($N = 346$)	Strong: MTurk ($N = 409$); Weak: MTurk ($N = 360$)
Show up fee	\$0.25	\$0.25	\$0.25
Action space	$a \in [0, 0.2, 0.4, 0.6, 0.8, 1]$ is fraction of \$1 sent	$a \in [0, 1.2]$ is amount returned, averaged over each possible nonzero amount received	$a \in [0, 1]$ is offer to the other player
Payoffs	\$2a is sent to second mover, who decides how much to send back; \$$(1 - a)$ not sent is kept with certainty	Amount not sent back	$1 - a$ if the offer is accepted, 0 if it is rejected
Strong demand treatment: "You will do us a favor if you"	"send more (less) to the other participant than you normally would"	"send back more (less) to the other participant than you normally would"	"offer more (less) to the other participant than you normally would"
Weak demand treatment: "We expect that participants who are shown these instructions will"	"send more (less) to the other participant than they normally would"	"send back more (less) to the other participant than they normally would"	"offer more (less) to the other participant than they normally would"

Task	Ultimatum game second mover	Lying
Description	Specify the smallest offer you would accept	Report the number of "Heads" in 10 coinflips
Sample	Strong: MTurk ($N = 421$); Weak: MTurk ($N = 411$)	Strong: MTurk ($N = 365$); Weak: MTurk ($N = 412$)
Show up fee	\$0.25	\$0.25
Action space	$a \in [0, 1]$ is min. acceptable offer: reject all offers below this amount	$a \in [0, 1, \ldots, 10]$ is number of heads
Payoffs	Amount received if it exceeds a, otherwise 0	10 cents per "Heads" reported: \$0.1a
Strong demand treatment: "You will do us a favor if you"	"require a higher (lower) minimum amount than you normally would"	"report more (fewer) heads than you normally would"
Weak demand treatment: "We expect that participants who are shown these instructions will"	"require a higher (lower) minimum amount than they normally would"	"report more (fewer) heads than they normally would"

A further use of levels estimated in the lab or surveys is to predict behavior in other contexts (e.g., using risk, time or social preference measures to predict real-world behaviors). The extent to which these measures are predictive may be sensitive to demand effects, which can be thought of as a form of measurement error. Our approach can be used to shed light on how important such errors might be. Within-subject applications even allow the researcher to measure and control for participant-level estimates of demand sensitivity.

We make the following recommendations on how to use demand treatments. First, in most studies, we believe "weak" manipulations will give sufficiently conservative bounds, because explicit signals about the study hypothesis are likely to be more informative than implicit messages from the design. If potential demand confounds are a first-order concern, researchers may find stronger language, similar to our "strong" manipulations, helpful for further robustness. Our phrasings were chosen for broad applicability, but researchers with a specific application in mind may prefer to design their own demand treatments to best suit their setting.[37] With bounds in hand, researchers can compute demand-robust confidence intervals following Imbens and Manski (2004).

Second, demand treatments can be applied within-participant by adding a small number of questions or tasks to the end of a study. These are repetitions of questions or tasks presented earlier in the study, now including a demand treatment. Our estimates suggest that this approach yields similar bounds to a between-participant design, but is much less demanding of sample size. It also allows researchers to identify which participants are most sensitive to demand, and compute "defier-corrected" bounds.

Third, we have shown how demand treatments can be used for point identification of treatment effects, applying same-signed demand treatments to the treatment and control group. If demand treatments are "sufficiently informative," this approach can eliminate biases due to differences in beliefs, and any remaining bias due to differences in behavioral responsiveness can be tested for. We have also shown how sufficiently informative demand treatments can be used for structural identification of models, by plausibly eliminating nuisance parameters due to unobservable beliefs.

Fourth, in a study with many treatment arms, adding all of the possible demand manipulations may become impractical. In such settings, researchers could add demand manipulations to a subset of groups, and then compare treatment effect magnitudes to demand sensitivity measured in those groups. When an experiment features many different and complicated choices, researchers may find it worthwhile to consider what overarching beliefs could affect their estimates (for example, participants might believe that they should misreport their valuations in willingness-to-pay elicitation), and target those with demand treatments, rather than manipulating individual actions.

Finally, researchers conducting similar experiments to those in this paper may find our estimates useful for benchmarking purposes.

[37] When bounding treatment effects, one could refer to the effect of interest in the demand treatment. For example, one could tell participants: "You are in the high incentive treatment and will be compared with a group that has low incentives. We expect that incentives will increase effort."

VI. Conclusion

We propose a technique for assessing the robustness of experimental results to demand effects. We deliberately induce demand in a structured way to measure its influence and to construct bounds on demand-free behavior and treatment effects. We formalize the intuition behind the procedure with a simple model in which participants form beliefs about the experimental objective and gain utility from conforming to it. Bounds are obtained by intentionally manipulating those beliefs.

Across 11 canonical experimental tasks we find modest responses to demand manipulations that explicitly signal the researcher's hypothesis, with bounds averaging around 0.13 standard deviations in width. We argue that these treatments reasonably bound the magnitude of demand effects in typical experiments, so our findings give cause for optimism.

Using stronger manipulations we show how to obtain demand-robust point estimates of treatment effects, and analyze demand effects structurally. In a real effort task with incentives of 1 cent per 100 points, we estimate a utility of pleasing the experimenter of around 0.2 cents per 100 points. Combining demand treatments with structural estimation can enable identification of preference parameters free of demand confounds.

Future work might employ similar treatments to study how to mitigate demand in experiments, for example by examining how demand sensitivity varies with features of the environment. One avenue for further exploration is the effect of incentives, given the central role they play in experiments.

REFERENCES

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105 (8): 2501–38.

Al-Ubaydli, Omar, John A. List, Danielle LoRe, and Dana Suskind. 2017. "Scaling for Economists: Lessons from the Non-Adherence Problem in the Medical Literature." *Journal of Economic Perspectives* 31 (4): 125–44.

Bardsley, Nicholas. 2008. "Dictator Game Giving: Altruism or Artefact?" *Experimental Economics* 11 (2): 122–33.

Barmettler, Franziska, Ernst Fehr, and Christian Zehnder. 2012. "Big Experimenter Is Watching You! Anonymity and Prosocial Behavior in the Laboratory." *Games and Economic Behavior* 75 (1): 17–34.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (3): 351–68.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58 (3): 739–53.

Bertrand, Marianne, and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *American Economic Review* 91 (2): 67–72.

Binmore, K., A. Shaked, and J. Sutton. 1985. "Testing Noncooperative Bargaining Theory: A Preliminary Study." *American Economic Review* 75 (5): 1178–80.

Bischoff, Ivo, and Björn Frank. 2011. "Good News for Experimenters: Subjects Are Hard to Influence by Instructors' Cues." *Economics Bulletin* 31 (4): 3221–25.

Bolton, Gary E., Elena Katok, and Rami Zwick. 1998. "Dictator Game Giving: Rules of Fairness versus Acts of Kindness." *International Journal of Game Theory* 27 (2): 269–99.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Forthcoming. "Beliefs about Gender." *American Economic Review*.

Camerer, Colin F. 2015. "The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List." In *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter, 249–95. Cambridge, UK: Oxford University Press.

Camerer, Colin F., Robin M. Hogarth, David V. Budescu, and Catherine Eckel. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework." *Journal of Risk and Uncertainty* 19 (1–3): 7–42.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn. 2012. "Experimental Methods: Between-Subject and Within-Subject Design." *Journal of Economic Behavior and Organization* 81 (1): 1–8.

Chassang, Sylvain, Gerard Padró I Miquel, and Erik Snowberg. 2012. "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments." *American Economic Review* 102 (4): 1279–309.

Cilliers, Jacobus, Oeindrila Dube, and Bilal Siddiqi. 2015. "The White-Man Effect: How Foreigner Presence Affects Behavior in Experiments." *Journal of Economic Behavior and Organization* 118: 397–414.

Clark, Herbert H., and Michael F. Schober. 1992. "Asking Questions and Influencing Answers." In *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*, edited by Judith M. Tanur, 15–28. New York: Russell Sage Foundation.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2): 448–74.

Dal Bó, Ernesto, and Pedro Dal Bó. 2014. "'Do the Right Thing': The Effects of Moral Suasion on Cooperation." *Journal of Public Economics* 117: 28–38.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127 (1): 1–56.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao. 2017. "Voting to Tell Others." *Review of Economic Studies* 84 (1): 143–81.

DellaVigna, Stefano, and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies* 85 (2): 1029–69.

DellaVigna, Stefano, and Devin Pope. Forthcoming. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy.*

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and Bounding Experimenter Demand: Dataset." *American Economic Review.* https://doi.org/10.1257/aer.20171330.

de Quidt, Jonathan, Lise Vesterlund, and Alistair Wilson. Forthcoming. "Experimenter Demand Effects." In *Handbook of Research Methods and Applications in Experimental Economics*, edited by Aljaž Ule and Arthur Schram. Cheltenham, UK: Edward Elgar Publishing.

Dupas, Pascaline, and Edward Miguel. 2017. "Impacts and Determinants of Health Levels in Low-Income Countries." In *Handbook of Economic Field Experiments,* Vol. 2, edited by Esther Duflo and Abhijit Banerjee, 3–93. Amsterdam: Elsevier.

Ellingsen, Tore, Robert Östling, and Erik Wengström. 2018. "How Does Communication Affect Beliefs in One-Shot Games with Complete Information?" *Games and Economic Behavior* 107: 153–81.

Falk, Armin, and James J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science* 326 (5952): 535–38.

Fleming, Piers, and Daniel John Zizzo. 2015. "A Simple Stress Test of Experimenter Demand Effects." *Theory and Decision* 78 (2): 219–31.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009–55.

Hauser, David J., and Norbert Schwarz. 2016. "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than Do Subject Pool Participants." *Behavior Research Methods* 48 (1): 400–407.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior* 7 (3): 346–80.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith. 1996. "On Expectations and the Monetary Stakes in Ultimatum Games." *International Journal of Game Theory* 25 (3): 289–301.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Imbens, Guido W., and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72 (6): 1845–57.

Kessler, Judd B., and Lise Vesterlund. 2015. "The External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effects." In *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter. Cambridge, UK: Oxford University Press.

Lambdin, Charles, and Victoria A. Shaffer. 2009. "Are Within-Subjects Designs Transparent?" *Judgment and Decision Making* 4 (7): 554–66.

Levitt, Steven D., and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21 (2): 153–74.

List, John A. 2006. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions." *Journal of Political Economy* 114 (1): 1–37.

List, John A. 2007. "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy* 115 (3): 482–93.

List, John A., Robert P. Berrens, Alok K. Bohara, and Joe Kerkvliet. 2004. "Examining the Role of Social Isolation on Stated Preferences." *American Economic Review* 94 (3): 741–52.

Loewenstein, George. 1999. "Experimental Economics from the Vantage-Point of Behavioural Economics." *Economic Journal* 109 (453): F23–34.

Milgram, Stanley. 1963. "Behavioral Study of Obedience." *The Journal of Abnormal and Social Psychology* 67 (4): 371–78.

Mummolo, Jonathan, and Erik Peterson. 2018. "Demand Effects in Survey Experiments: An Empirical Assessment." Unpublished.

Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *American Psychologist* 17 (11): 776–83.

Rosenthal, Robert. 1966. *Experimenter Effects in Behavioral Research.* New York: Appleton-Century-Crofts.

Shmaya, Eran, and Leeat Yariv. 2016. "Experiments on Decisions under Uncertainty: A Theoretical Framework." *American Economic Review* 106 (7): 1775–1801.

Small, Deborah A., George Loewenstein, and Paul Slovic. 2007. "Sympathy and Callousness: The Impact of Deliberative Thought on Donations to Identifiable and Statistical Victims." *Organizational Behavior and Human Decision Processes* 102 (2): 143–53.

Tsutsui, Kei, and Daniel John Zizzo. 2014. "Group Status, Minorities and Trust." *Experimental Economics* 17 (2): 215–44.

Zizzo, Daniel John. 2010. "Experimenter Demand Effects in Economic Experiments." *Experimental Economics* 13 (1): 75–98.