Measuring and Bounding Experimenter Demand

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

This version: April 4, 2018

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 eleven 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: B41, C91, C92
Keywords: Experimenter Demand, Beliefs, Bounding

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).

* de Quidt: Institute for International Economic Studies and CESifo, address: Institute for International Economic Studies, Stockholm University, 106 91 Stockholm, Sweden, email: jonathan.dequidt@iies.su.se. Haushofer: Department of Psychology, Princeton University, NBER, and Busara Center for Behavioral Economics, address: 427 Peretsman-Scully Hall, Princeton University, Princeton, NJ 08540, United States, email: haushofer@princeton.edu. Roth: Department of Economics, University of Oxford, and CSAE, address: Keble College, Parks Road, Oxford, OX1 3PG, United Kingdom, email: christopher.roth@economics.ox.ac.uk. We thank the editors, Stefano DellaVigna and Esther Duflo, and 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 in 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.

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 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

[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.

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 11 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 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 zero 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, thus researchers are right to pay close attention to potential demand effects in their studies.

[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.

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's 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 treatment. We also compare our bounds to estimates of the effect of double anonymity in dictator games, one 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, 2014; 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 (2017) 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

[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.

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).

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, 2012; 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., 2016); favor between-participant designs despite the larger samples required (Charness, Gneezy and Kuhn, 2012);[6] or conduct costly natural field ex-

[6]Charness, Gneezy and Kuhn (2012) write (p2), "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."

periments (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 remains 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. A set of web appendices contains theoretical details and additional results.

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. Web appendices 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. 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).

[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 (2018) provide a discussion, de Quidt, Vesterlund and Wilson (2018) also measure their adoption in published experimental papers.

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:

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

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 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

[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.

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 zero. 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 (Angrist and Imbens, 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* $S(\tau)$ *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 $v_1(a(\zeta), \zeta) = 0$.

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.

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

[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.

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]

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. η and ϵ are believed to be independent (we revisit this assumption in web Appendix B.B6). The Bayesian posterior is:

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

[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 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 web Appendix B.B5 we extend the model to allow ϕ to depend on h^T and show that the Bounding condition remains unchanged.

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 I.A. Derivations can be found in web 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 $\phi E[h|h^T, h^L] \geq 0$ or $\phi E[h|h^T, h^L] \leq 0$ 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 web Appendix 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 of inattention to experimenter demand. Finally, we show that Monotone Sensitivity does *not* hold in general when environments differ in

the beliefs they induce ($E[h|h^L(\zeta)]$). We use these findings when interpreting heterogeneous responses to demand treatments in section IV.D.

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 III.D and III.E.

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

[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.

individuals, because $a^- > a^+$. We show in web Appendix B.B4 that the method is able to tolerate some defier behavior, but too much will lead to failures of Bounding. 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 that the response by the compliers outweighs that of the defiers.

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 web Appendix E). Our respondents complete one of eleven 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

[12] For the effort task, we replicated the design employed in DellaVigna and Pope (2017) and DellaVigna and Pope (2018). 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.

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. Figure A1 gives an example from the experimental interface. Full experimental instructions can be found on the journal webpage.

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 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 U.S, 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*.

[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 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. 70 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.

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 eleven 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, web Appendix C presents all pre-specified analyses, experiment-by-experiment. We refer to findings in the text if relevant.

C. *Summary statistics*

Tables D1 to D7 in the web Appendix 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).

[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 that take more than 30 minutes, take the task more than once, score zero 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.

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 IV.A 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 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 providing 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.

[Insert Table 1, Table 2, and Figures 1 and 2]

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 were told that their score "will not affect [their] payment," while a second group received one cent per 100 points. By combining the 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.

[Insert Table 3]

C. Confidence intervals

It is possible to compute confidence intervals for (a) the bounds themselves, and (b) the parameters contained by those bounds (a natural action or treatment effect), following Imbens and Manski (2004) (see Appendix 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 $a^+(1) - a^-(0)$.

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 I.C). 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 web Appendix 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 $v_1(a^*(\zeta, h^T), \zeta) + \phi(\zeta)E[h|h^T, h^L(\zeta)] = 0$. A first-order Taylor approximation around the natural action $a(\zeta)$ yields:

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

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." 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 $a^*(1, h^T) - a^*(0, h^T)$.

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

Its bias relative to the true effect can be decomposed as follows:

$$\begin{aligned} Bias &= [a^*(1, h^T) - a^*(0, h^T)] - [a(1) - a(0)] \\ &\approx \underbrace{\Phi(1)\left(E[h|h^T, h^L(1)] - E[h|h^T, h^L(0)]\right)}_{\text{Bias due to beliefs}} + \underbrace{(\Phi(1) - \Phi(0))\, E[h|h^T, h^L(0)]}_{\text{Bias due to ``responsiveness''}} \end{aligned}$$

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 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]

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

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

[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.

positive and two negative demand treatments:

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

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. $h^T = h^L(1)$) 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 above-developed approaches to our effort experiment in web 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

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

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 $E[h|h^L]$. 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) (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:

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

with corresponding first-order condition

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

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

[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.

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

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 web Appendix B.B9. We estimate equation (12) in logs and equation (13) in levels. Es-

[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.

timation 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 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 web Appendix B.B9.

[Insert Table 4]

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

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

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 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 is fully consistent with Monotonicity for both the weak and strong treatments (see Table C18).

[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.

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 experimenters wishes. 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 I.D "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 one of them, selected by chance. Half played the dictator game twice, and 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.

Table A2 presents mean actions and sensitivities estimated from the within design and the equivalent objects from the earlier between-participants experi-

[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.

ments. 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 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.

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 Tables A8–A9 in the web Appendix. 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

[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 $[cE[a^-(\zeta)|\phi \geq 0] + (1-c)E[a^+(\zeta)|\phi < 0], cE[a^+(\zeta)|\phi \geq 0] + (1-c)E[a^-(\zeta)|\phi < 0]]$. 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.

[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.

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 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 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.

The 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.

[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 Appendix 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 vs. 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 Appendix 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 vs. 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 one for the positive demand treatment and zero 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 web 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 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]

[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 web Appendix C. Experiment 1 (strong treatments) finds higher sensitivity for women (p=0.10) and attentive

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

[Insert Table 5]

MTurk vs. 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] 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.

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.

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

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. Table A12 shows that variance is very similar and in most cases slightly lower under the demand treatments relative to no demand. 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. Web 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, 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 coinflip games; or the degree of risk or ambiguity aversion (e.g. for calibrating models).

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 in most cases. 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 com-

[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."

pute "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.

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 eleven 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 es-

timates 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." *The 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.

Angrist, Joshua, and Guido Imbens. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–476.

Bardsley, Nicholas. 2008. "Dictator Game Giving: Altruism or Artefact?" *Experimental Economics*, 11(2): 122–133.

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): 351–368.

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): 739–753.

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

Binmore, Ken, Avner Shaked, and John Sutton. 1985. "Testing Noncooperative Bargaining Theory: A Preliminary Study." *The American Economic Review*, 75(5): 1178–1180.

Bischoff, Ivo, and Björn Frank. 2011. "Good News for Experimenters: Subjects are Hard to Influence by Instructors' Cues." *Economics Bulletin*, 31(4): 3221–3225.

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): 269–299.

Bordalo, Pedro, Katie Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Beliefs about Gender." *Mimeo, Harvard University*.

Camerer, Colin F. 2012. "The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List." In *The Methods of Modern Experimental Economics*. 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*, 7–48.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn. 2012. "Experimental Methods: Between-subject and Within-subject Design." *Journal of Economic Behavior & 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." *The American Economic Review*, 102(4): 1279–1309.

Cilliers, Jacobus, Oeindrila Dube, and Bilal Siddiqi. 2015. "The Whiteman Effect: How Foreigner Presence Affects Behavior in Experiments." *Journal of Economic Behavior & 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.* , ed. Judith M Tanur. Russell Sage Foundation.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature*, 47(2): 448–474.

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, and Devin Pope. 2017. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy, forthcoming.*

DellaVigna, Stefano, and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies*, 85(2): 1029–1069.

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

DellaVigna, Stefano, John A List, Ulrike Malmendier, and Gautam Rao. 2017. "Voting to Tell Others." *Review of Economic Studies*, 84(1): 143–181.

de Quidt, Jonathan, Lise Vesterlund, and Alistair Wilson. 2018. "Experimenter Demand Effects." In *Handbook of Research Methods and Applications in Experimental Economics, forthcoming.* , ed. Aljaž Ule and Arthur Schram. 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, 3–93. 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: 153–181.

Falk, Armin, and James J Heckman. 2009. "Lab Experiments are a Major Source of Knowledge in the Social Sciences." *Science*, 326(5952): 535–538.

Fleming, Piers, and Daniel John Zizzo. 2014. "A Simple Stress Test of Experimenter Demand Effects." *Theory and Decision*, 78(2): 219–231.

Harrison, Glenn W, and John A List. 2004. "Field Experiments." *Journal of Economic Literature*, 42(4): 1009–1055.

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): 400–7.

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.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 7(3): 346–380.

Imbens, Guido W, and Charles F Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica*, 72(6): 1845–1857.

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

Lambdin, Charles, and Victoria A. Shaffer. 2009. "Are Within-subjects Designs Transparent?" *Judgment and Decision Making*, 4(7): 554–566.

Levitt, Steven D, and John A List. 2007. "What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *The Journal of Economic Perspectives*, 21(2): 153–174.

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): 482–493.

List, John A, Robert P Berrens, Alok K Bohara, and Joe Kerkvliet. 2004. "Examining the Role of Social Isolation on Stated Preferences." *The American Economic Review*, 94(3): 741–752.

Loewenstein, George. 1999. "Experimental Economics From the Vantage-point of Behavioural Economics." *The Economic Journal*, 109(453): 25–34.

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

Mummolo, Jonathan, and Erik Peterson. 2017. "Demand Effects in Survey Experiments: An Empirical Assessment." *available at SSRN: https://ssrn.com/abstract=2956147*.

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.

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

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

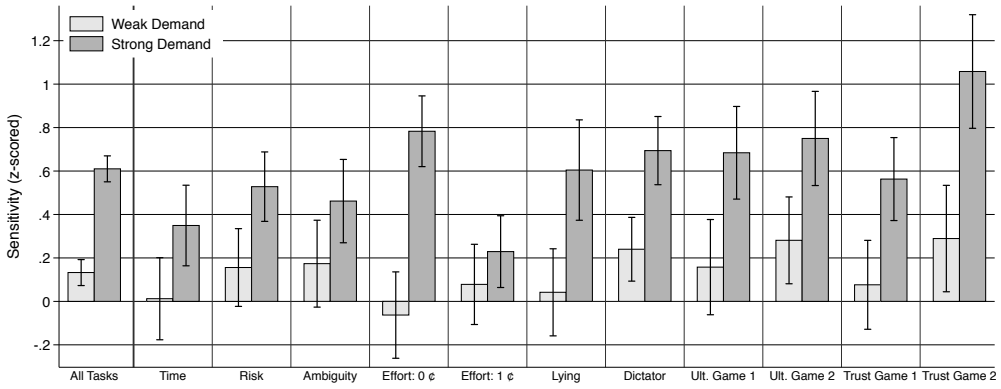
Small, Deborah A, George Loewenstein, and Paul Slovic. 2007. "Sympathy and Callousness: The Impact of Deliberative Thought on Donations to Identifiable and Statistical Victims." *Organizational Behavior and Human Decision Processes*, 102(2): 143–153.

Tsutsui, Kei, and Daniel John Zizzo. 2014. "Group Status, Minorities and Trust." *Experimental Economics*, 17: 215–244.

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

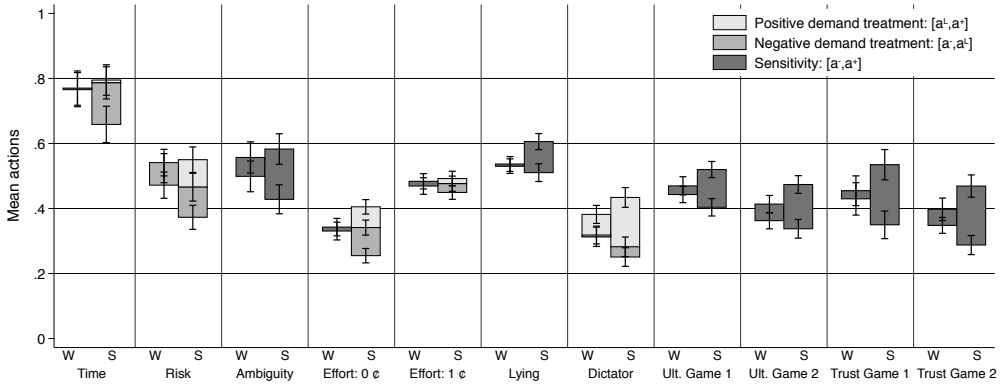
VII. Main Figures and Tables

Figure 1. Sensitivity to demand treatments, z-scored



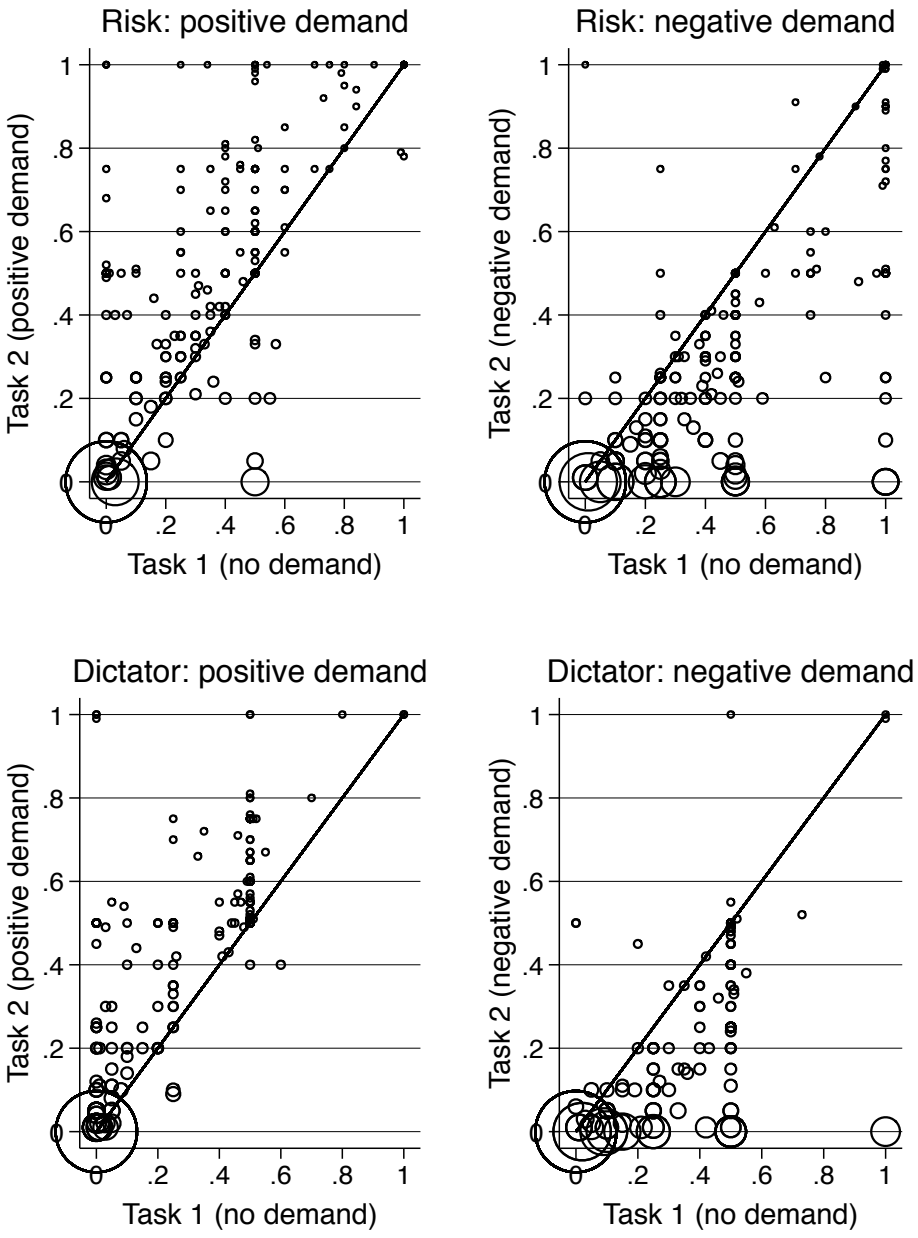
Note: This figure uses data from incentivized MTurk respondents with weak and strong demand treatments. It presents the z-scored sensitivity of behavior to our demand treatments, i.e. the normalized difference in behavior between the positive and negative demand conditions. Error bars indicate 95 percent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

Figure 2. Bounding natural actions



Note: 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.

Figure 3. Measuring defiance through a within-participant design



Note: 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.

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 (z-score)		-0.051					0.261				
		(0.092)					(0.078)				
		[0.237]					[0.002]				
Negative - Neutral (z-score)		-0.207					0.021				
		(0.087)					(0.078)				
		[0.056]					[0.357]				
Observations	422	739	390	388	381	412	758	360	411	352	346

Note: 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.

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 (z-score)	0.022	0.252		0.333	0.084		0.574				
	(0.088)	(0.088)		(0.085)	(0.088)		(0.082)				
	[0.363]	[0.001]		[0.001]	[0.159]		[0.001]				
Negative - Neutral (z-score)	-0.327	-0.276		-0.450	-0.145		-0.120				
	(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

Note: 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.

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)

Note: 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 6 show lower and upper bounds estimated using weak and strong treatments. We apply the "ironing" procedure described in section III.B 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.

Table 4—Structural Estimates

	Power cost of effort			Exponential cost of effort		
	(1) Log Count	(2) Log Count	(3) Log Count	(4) Count	(5) Count	(6) Count
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	1691	1691	727	1691	1691
R-squared	0.122	0.166	0.166	0.167	0.204	0.206

Note: 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. $h^L(\zeta)p^L(\zeta)$ is latent demand in incentive condition ζ. $h^L(>0)p^L(>0)$ is latent demand in the combined 1-cent and 4-cent incentive conditions. Robust standard errors in parentheses.

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	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust Game 1	Trust Game 2
Panel A: Weak - Design Characteristics												
Sensitivity × Incentive	0.073		0.149					-0.002				
	(0.085)		(0.125)					(0.115)				
Observations	1963		970					993				
Panel B: Weak - Respondent Characteristics												
Sensitivity × Male	0.038	-0.028	0.057	0.069	0.305	0.033	0.060	0.029	-0.089	0.003	0.257	-0.239
	(0.061)	(0.174)	(0.179)	(0.203)	(0.239)	(0.236)	(0.189)	(0.146)	(0.192)	(0.185)	(0.230)	(0.229)
Observations	4450	422	473	390	388	381	412	515	360	411	352	346
Sensitivity × Attention	0.119	-0.402	-0.077	0.434			0.226	0.585	0.094	0.368	0.116	-0.398
	(0.116)	(0.395)	(0.307)	(0.504)			(0.305)	(0.328)	(0.296)	(0.230)	(0.362)	(0.301)
Observations	3681	422	473	390			412	515	360	411	352	346
Sensitivity × Representative sample	0.032		0.032					0.032				
	(0.084)		(0.127)					(0.110)				
Observations	2125		1041					1084				
Panel C: Strong - Design Characteristics												
Sensitivity × Incentive	-0.007	-0.063	0.196					0.072				
	(0.080)	(0.132)	(0.116)					(0.121)				
Observations	2989	994	996					999				
Panel D: Strong - Respondent Characteristics												
Sensitivity × Male	-0.152	-0.212	-0.090	-0.382	0.075	0.005	-0.223	-0.201	-0.137	-0.144	0.098	-0.361
	(0.064)	(0.168)	(0.160)	(0.192)	(0.197)	(0.214)	(0.217)	(0.153)	(0.187)	(0.201)	(0.216)	(0.240)
Observations	4800	491	482	404	492	472	365	511	409	421	382	371
Sensitivity × Attention	0.117	0.319	0.471	-0.276			0.255	-0.024	-0.272	0.229	0.918	-0.091
	(0.140)	(0.393)	(0.401)	(0.414)			(0.358)	(0.530)	(0.394)	(0.538)	(0.409)	(0.311)
Observations	3836	491	482	404			365	511	409	421	382	371
Sensitivity × Representative sample	0.027		-0.121					0.176				
	(0.081)		(0.118)					(0.112)				
Observations	2184		1070					1114				

Note: 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 one for males, "attention" equals one if the respondent passed the attention screener, "representative sample" equals one for representative sample respondents.

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 7)	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

Note: 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.

TABLE 7—DETAILS OF EXPERIMENTAL TASKS

Task	**Description**	**Sample**	**Show up fee**	**Action space**	**Payoffs**	**Strong Demand Treatment:** "You will do us a favor if you..."	**Weak Demand Treatment:** "We expect that participants who are shown these instructions will..."
Dictator Game	Choose to split money between yourself and another participant.	Strong: MTurk between (N=1,503); MTurk within (998); Rep. Sample (N=899). Weak: MTurk between (N=1,477); Rep. Sample (N=865).	$.25	$a \in [0,1]$ is fraction of $1 sent to the other participant.	Amount not sent $(1-a)$.	"...give more (less) to the other participant than you normally would."	"...give more (less) to the other participant than they normally would."
Investment Game	Choose how much to invest in a risky project.	Strong: MTurk between (N=1,492); MTurk within (1000); Rep Sample (N=902). Weak: MTurk between (N=1,473); Rep. Sample (N=883).	$.25	$a \in [0,1]$ is fraction of $1 invested	$\$(1-a)+3a$ with probability 0.4, $\$(1-a)$ otherwise.	"...invest more (less) than you normally would."	"...invest more (less) than they normally would."
Ambiguous Investment Game	Choose to how much to invest in an uncertain project.	Strong: MTurk (N= 404). Weak: MTurk (N= 390).	$.25	$a \in [0,1]$ is fraction of $1 invested	$\$(1-a)+3a$ with unstated probability, $\$(1-a)$ otherwise.	"...invest more (less) than you normally would."	"...invest more (less) than they normally would."
Convex Time Budgets	Choose between receiving money today vs. money in seven days.	Strong: MTurk (N=1,484); Rep. Sample (N=899). Weak: MTurk (N=422).	$.25	$a \in [0,1.2]$ is the amount to be received in 7 days	$\$(1-a)/1.2$ is received within 24 hours, and $\$a$ is received in 7 days.	"...choose to receive more (less) in seven days than you normally would."	"...choose to receive more (less) in seven days than they normally would."
Effort: No bonus	Alternately press the a and b button without receiving any bonus.	Strong: MTurk (N=731). Weak: MTurk (N=388).	$1	$a \in [0,4000]$ is number of a-b button presses	No payoffs beyond show-up fee	"...work harder (less hard) than you normally would."	"...work harder (less hard) than they normally would."
Effort: 1-cent bonus	Alternately press the a and b button, receiving 1 cent per 100 points.	Strong: MTurk (N=714). Weak: MTurk (N=381).	$1	$a \in [0,4000]$ is number of a-b button presses	1 cent per 100 button presses.	"...work harder (less hard) than you normally would."	"...work harder (less hard) than they normally would."
Trust Game 1st mover	Choose to send an amount of money to the other player.	Strong: MTurk (N=382). Weak: MTurk (N=352).	$.25	$a \in [0,.2,.4,.6,.8,1]$ is fraction of $1 sent	$\$2a$ is sent to second mover, who decides how much to send back. $\$(1-a)$ not sent is kept with certainty.	"...send more (less) to the other participant than you normally would."	"...send more (less) to the other participant than they normally would."
Trust Game 2nd mover	Choose to send back some money to the other player. (Strategy method)	Strong: MTurk (N=371). Weak: MTurk (N=346).	$.25	$a \in [0,1.2]$ is amount returned, averaged over each possible nonzero amount received.	Amount not sent back.	"...send back more (less) to the other participant than you normally would."	"...send back more (less) to the other participant than they normally would."
Ultimatum Game 1st mover	Offer a split to the other player.	Strong: MTurk (N=409). Weak: MTurk (N=360).	$.25	$a \in [0,1]$ is offer to the other player	$1-a$ if the offer is accepted, 0 if it is rejected.	"...offer more (less) to the other participant than you normally would."	"...offer more (less) to the other participant than they normally would."
Ultimatum Game 2nd mover	Specify the smallest offer you would accept.	Strong: MTurk (N=421). Weak: MTurk (N=411).	$.25	$a \in [0,1]$ is min. acceptable offer: reject all offers below this amount.	Amount received if it exceeds a, otherwise zero.	"...require a higher (lower) minimum amount than you normally would."	"...require a higher (lower) minimum amount than they normally would."
Lying	Report the number of "Heads" in 10 coinflips.	Strong: MTurk (N=365). Weak: MTurk (N=412).	$.25	$a \in [0,1,\ldots,10]$ is number of heads.	10 cents per "Heads" reported: $\$0.1a$.	"...report more (fewer) heads than you normally would."	"...report more (fewer) heads than they normally would."

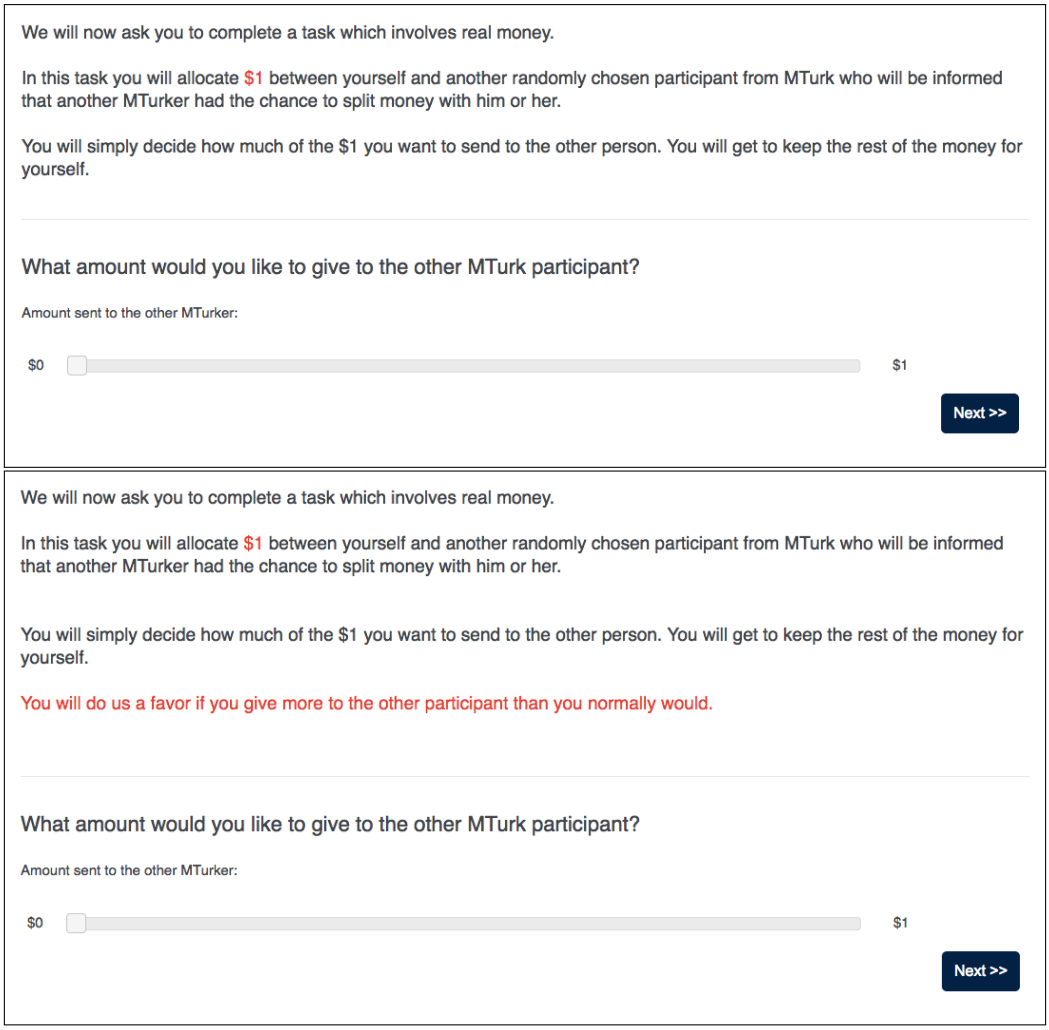
Measuring and Bounding Experimenter Demand

Jonathan de Quidt Johannes Haushofer Christopher Roth

Online Appendix

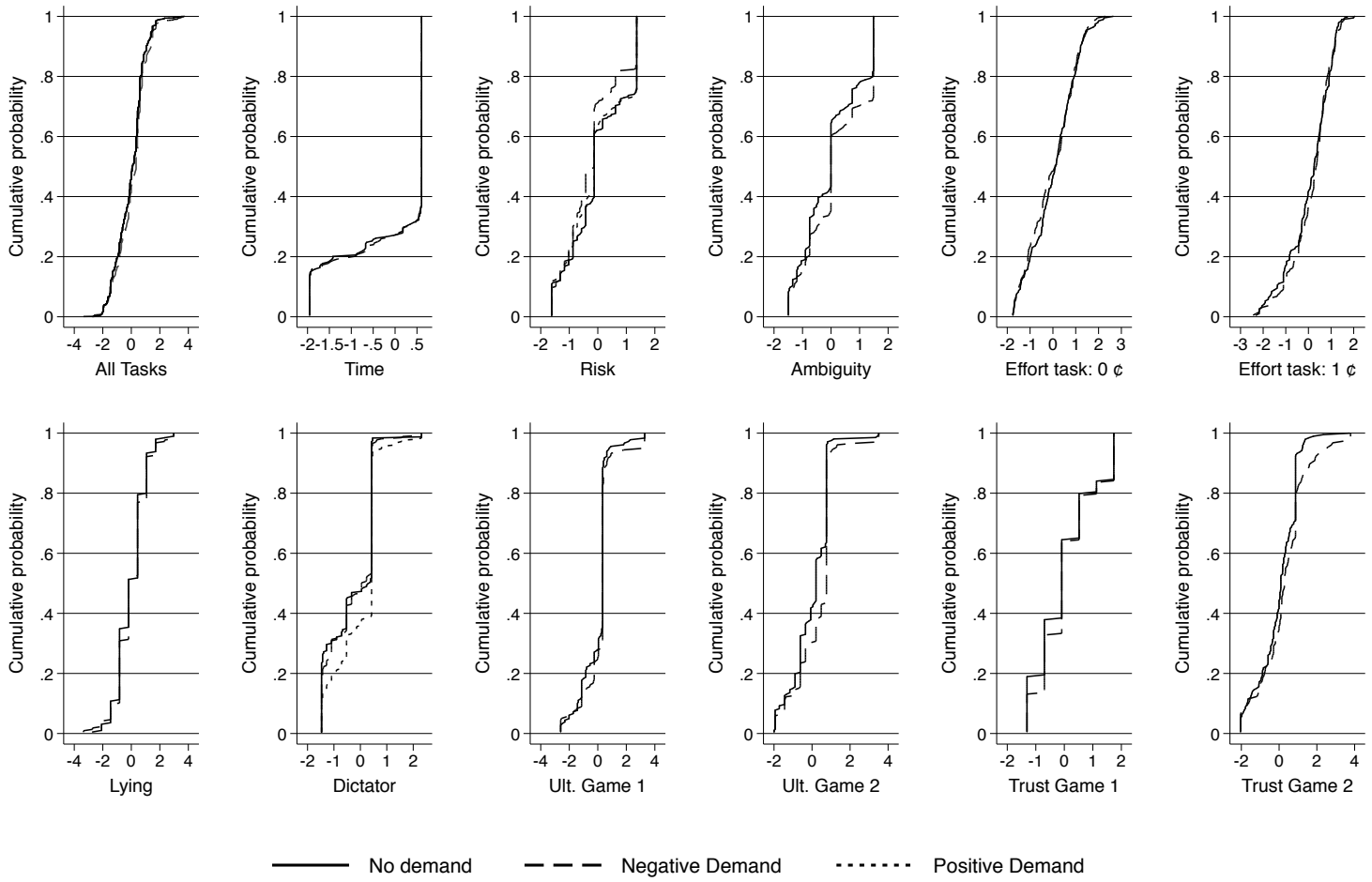
A. Additional Figures and Tables

FIGURE A1. INTERFACE EXAMPLE



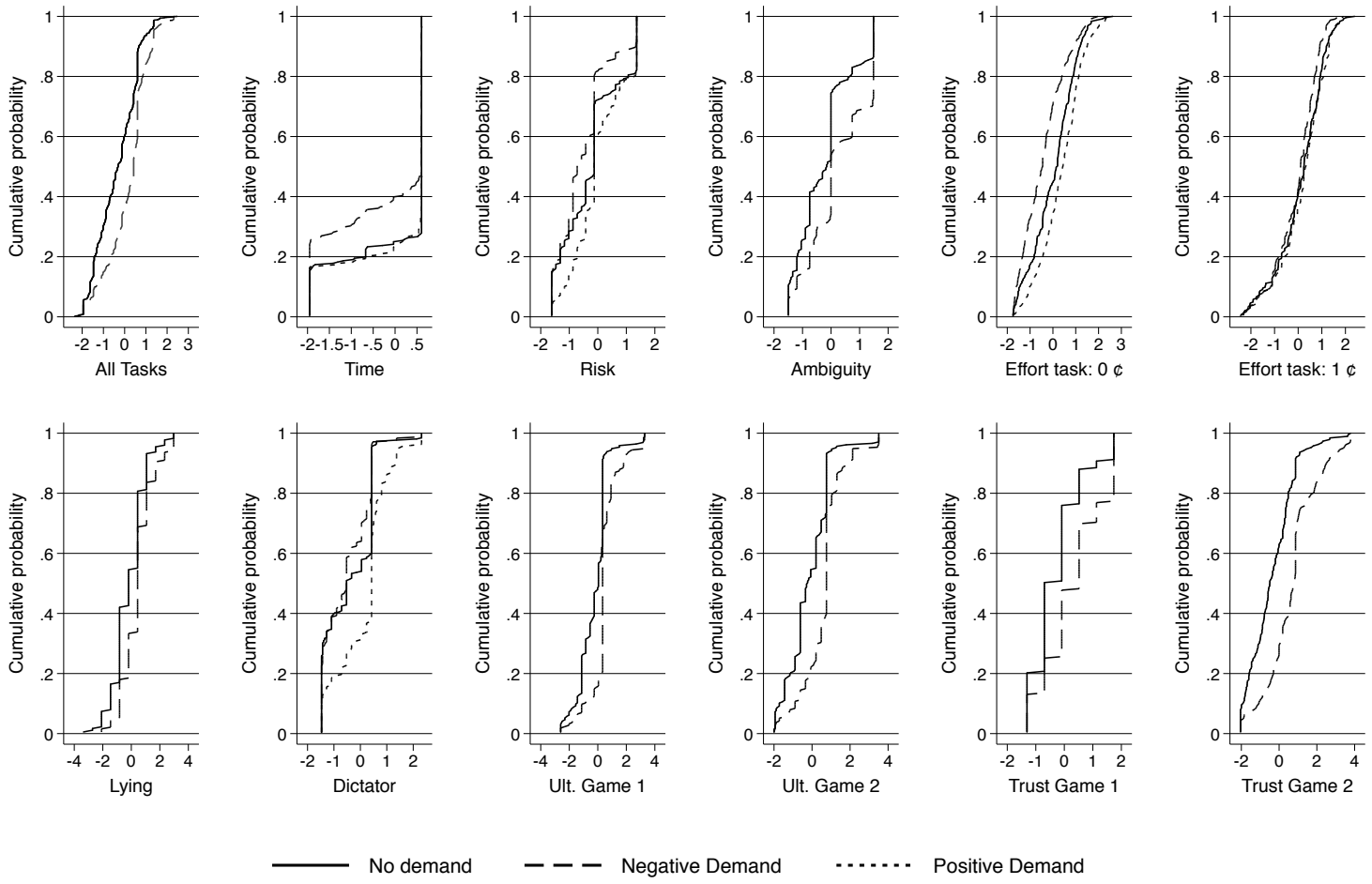
Note: We present two examples of the experimental interface, taken from the dictator game. The first frame corresponds to the real stakes, "no demand" condition, and the second frame to the real stakes, positive demand condition.

Figure A2. Distribution of z-scored actions by task and demand treatment, weak treatments



Note: This figure uses data from incentivized MTurk respondents with weak demand treatments, and displays the cumulative distribution function of z-scored actions by task and demand treatment arm.

Figure A3. Distribution of z-scored actions by task and demand treatment, strong treatments



Note: This figure uses data from incentivized MTurk respondents with strong demand treatments, and displays the cumulative distribution function of z-scored actions by task and demand treatment arm.

Table A1—Controlling for demand

	Conventional	Sensitivity difference	Strong	Weak	
	Treatment Effect	(Strong 1-cent - Strong 0-cent)	Midpoint	positive-positive	negative-negative
Count	540.720	430.009	494.774	588.270	530.001
	(66.763)	(89.477)	(49.321)	(61.499)	(64.532)
Count (z-scored)	0.686	0.546	0.628	0.747	0.673
	(0.085)	(0.114)	(0.063)	(0.078)	(0.082)

Note: This table uses data from the real effort experiments (experiment 3 and experiment 6). We follow the "controlling for demand" procedure outlined in Section III.D to estimate the treatment effect of incentives on effort provision. Column (1) shows the conventional treatment effect estimate (data from experiment 3). Column (2) tests for differences in sensitivity to our strong demand treatments between the 0-cent and 1-cent groups, and finds a significant difference. Therefore in column (3) we apply the "midpoint" technique with strong demand treatments to estimate the treatment effect. Columns (4) and (5) approximate the treatment effect using same-signed weak demand treatments. We apply the "ironing" procedure described in section III.B when constructing these estimates. Count is the raw-score of points scored in the real effort task. Count (z-scored) uses the mean and standard deviation from the negative demand condition. Robust standard errors in parentheses. Note that strong and weak treatment data were collected in separate experiments.

Table A2—Results from the Within Design

	Dictator			Risk		
	Within	Between	Difference	Within	Between	Difference
Panel A: Unconditional Means						
Positive demand	0.384	0.434	-0.050	0.560	0.550	0.010
	(0.017)	(0.015)	(0.023)	(0.021)	(0.020)	(0.029)
No demand	0.273	0.282	-0.010	0.448	0.466	-0.018
	(0.011)	(0.015)	(0.019)	(0.015)	(0.022)	(0.027)
Negative demand	0.195	0.251	-0.056	0.318	0.373	-0.055
	(0.014)	(0.014)	(0.020)	(0.019)	(0.019)	(0.027)
Panel B: Sensitivity (positive - negative)						
Raw data	0.189	0.183	0.006	0.242	0.177	0.065
	(0.022)	(0.021)	(0.031)	(0.029)	(0.027)	(0.040)
Z-score	0.794	0.745	0.048	0.709	0.520	0.188
	(0.093)	(0.086)	(0.127)	(0.084)	(0.080)	(0.116)
Panel C: Monotonicity						
Positive - Neutral (z-score)	0.514	0.617	-0.103	0.377	0.248	0.129
	(0.044)	(0.088)	(0.129)	(0.041)	(0.087)	(0.124)
Negative - Neutral (z-score)	-0.380	-0.128	-0.251	-0.427	-0.272	-0.155
	(0.045)	(0.086)	(0.123)	(0.042)	(0.084)	(0.119)
Observations	499	770	1269	500	728	1228

Note: This table uses data from the within design (experiment 7) and incentivized choices from the dictator game and the investment game in experiment 1. These experiments employ strong demand treatments. Panel A displays the unconditional means by task and demand treatment arm. Panel B displays the estimates of sensitivity. Panel C tests Monotonicity. Note that estimates from Panel C do not add up to the sensitivity estimates from Panel B as sensitivity is estimated between participants while monotonicity tests are within-participant.

Table A3—Confidence intervals for bounds on natural actions

	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: Weak Demand											
Interval	[0.766, 0.770]	[0.472, 0.524]	[0.499, 0.557]	[0.343, 0.331]	[0.469, 0.484]	[0.469, 0.484]	[0.318, 0.382]	[0.443, 0.470]	[0.362, 0.413]	[0.430, 0.455]	[0.348, 0.398]
95% CI on interval	[0.716, 0.821]	[0.438, 0.561]	[0.459, 0.597]	[0.316, 0.358]	[0.448, 0.504]	[0.510, 0.557]	[0.296, 0.405]	[0.422, 0.493]	[0.342, 0.436]	[0.387, 0.494]	[0.328, 0.426]
95% CI on parameter	[0.724, 0.812]	[0.445, 0.553]	[0.468, 0.588]	[0.320, 0.353]	[0.452, 0.499]	[0.514, 0.553]	[0.301, 0.400]	[0.427, 0.488]	[0.346, 0.431]	[0.396, 0.486]	[0.332, 0.420]
Observations	422	739	390	388	381	412	758	360	411	352	346
Panel B: Strong Demand											
Interval	[0.659, 0.795]	[0.373, 0.550]	[0.428, 0.583]	[0.255, 0.405]	[0.449, 0.492]	[0.449, 0.492]	[0.251, 0.434]	[0.404, 0.520]	[0.337, 0.474]	[0.350, 0.535]	[0.288, 0.469]
95% CI on interval	[0.612, 0.834]	[0.342, 0.583]	[0.391, 0.622]	[0.236, 0.424]	[0.432, 0.511]	[0.487, 0.626]	[0.227, 0.459]	[0.381, 0.541]	[0.314, 0.496]	[0.314, 0.574]	[0.263, 0.498]
95% CI on parameter	[0.622, 0.826]	[0.349, 0.576]	[0.399, 0.613]	[0.240, 0.420]	[0.436, 0.507]	[0.493, 0.622]	[0.232, 0.454]	[0.386, 0.536]	[0.319, 0.491]	[0.322, 0.565]	[0.269, 0.491]
Observations	727	728	404	731	714	365	770	409	421	382	371

Note: This table uses data from incentivized MTurk respondents with strong and weak demand treatments. It first presents estimated bounds on the natural action, then 95 percent confidence intervals on those bounds, then 95 percent confidence intervals on the parameter (natural action) contained in the bounds.

Table A4—Confidence intervals for bounds on treatment effects

	Treatment Effect: Score in Effort Task
Weak treatments	
Interval	[530.001, 588.270]
95% CI on interval	[410.310, 701.645]
95% CI on parameter	[434.271, 678.736]
Observations	769
Strong treatments	
Interval	[177.421, 948.978]
95% CI on interval	[74.817, 1054.492]
95% CI on parameter	[97.479, 1031.187]
Observations	1445

Note: This table uses data from incentivized MTurk respondents with weak and strong demand treatments (experiments 3 and 6). It first presents estimated bounds on the treatment effect of incentives on effort, then 95 percent confidence intervals on those bounds, then 95 percent confidence intervals on the parameter (treatment effect) contained in the bounds.

Table A5—Results from the Within Design: Compliers and Defiers

	Dictator			Risk		
	All	Compliers	Defiers	All	Compliers	Defiers
Positive - Neutral (z-score)	0.514	0.777	-0.402	0.377	0.704	-0.601
	(0.044)	(0.055)	(0.122)	(0.041)	(0.052)	(0.100)
Observations	265	179	7	247	146	16
Negative - Neutral (z-score)	-0.380	-0.796	1.028	-0.427	-0.721	0.529
	(0.045)	(0.059)	(0.329)	(0.042)	(0.049)	(0.199)
Observations	234	122	8	253	161	16

Note: This table uses data from the within design (experiment 7). The outcome variable is the change in standardized action between task 1 and task 2. We separately present the results for the whole sample, compliers, and defiers.

Table A6—Within Design: Defier-corrected bounds and confidence intervals

	Risk	Dictator
Panel A: Standard Bounds		
Interval	[0.318, 0.560]	[0.195, 0.384]
95% CI on interval	[0.286, 0.595]	[0.172, 0.412]
95% CI on parameter	[0.293, 0.587]	[0.177, 0.406]
Observations	500	499
Panel B: Adjusted Bounds		
Interval	[0.308, 0.571]	[0.185, 0.392]
95% CI on interval	[0.277, 0.606]	[0.163, 0.420]
95% CI on parameter	[0.284, 0.598]	[0.168, 0.414]
Observations	500	499

Note: This table uses data from the within design (experiment 7). In Panel A we compute our standard bounds and confidence intervals. In Panel B we compute the adjusted bounds which take into account defier behavior.

TABLE A7—BELIEF ABOUT THE EXPERIMENTAL OBJECTIVE IN RESPONSE TO THE WEAK DEMAND TREATMENTS

	Belief: Time	Belief: Risk	Belief: Ambiguity Aversion	Belief: Effort 0 cent bonus	Belief: Effort 1 cent bonus	Belief: Lying	Belief: Dictator Game	Belief: Ult. Game 1	Belief: Ult. Game 2	Belief: Trust Game 1	Belief: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.832	0.757	0.760	0.788	0.979	0.779	0.540	0.698	0.688	0.612	0.669
	(0.026)	(0.028)	(0.031)	(0.030)	(0.010)	(0.028)	(0.032)	(0.034)	(0.032)	(0.036)	(0.038)
No demand		0.620					0.321				
		(0.030)					(0.030)				
Negative demand	0.603	0.370	0.330	0.277	0.358	0.467	0.234	0.238	0.383	0.112	0.083
	(0.034)	(0.031)	(0.034)	(0.032)	(0.035)	(0.036)	(0.026)	(0.032)	(0.034)	(0.024)	(0.020)
Panel B: Sensitivity (Positive - Negative)											
Raw data	0.229	0.388	0.430	0.511	0.621	0.312	0.306	0.461	0.304	0.500	0.585
	(0.042)	(0.042)	(0.046)	(0.044)	(0.036)	(0.046)	(0.041)	(0.047)	(0.047)	(0.044)	(0.043)
Z-score	0.471	0.776	0.909	1.117	1.240	0.627	0.678	0.994	0.633	1.092	1.417
	(0.087)	(0.084)	(0.096)	(0.095)	(0.073)	(0.092)	(0.091)	(0.101)	(0.098)	(0.095)	(0.104)
		[0.001]					[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)		0.274					0.485				
		(0.082)					(0.096)				
		[0.001]					[0.001]				
Negative - Neutral (z-score)		-0.501					-0.193				
		(0.087)					(0.088)				
		[0.001]					[0.009]				
Observations	422	739	390	388	381	412	758	360	411	352	346

Note: This table uses data from incentivized MTurk respondents with weak demand treatments. The outcome variables take value one if the respondents believed that the experimenter wanted a high action. Panel A displays mean beliefs with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of beliefs 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.

TABLE A8—BELIEF ABOUT THE EXPERIMENTAL OBJECTIVE IN RESPONSE TO THE STRONG DEMAND TREATMENTS

	Belief: Time	Belief: Risk	Belief: Ambiguity Aversion	Belief: Effort 0 cent bonus	Belief: Effort 1 cent bonus	Belief: Lying	Belief: Dictator Game	Belief: Ult. Game 1	Belief: Ult. Game 2	Belief: Trust Game 1	Belief: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.802	0.705	0.701	0.776	0.942	0.815	0.651	0.572	0.664	0.407	0.385
	(0.025)	(0.030)	(0.032)	(0.027)	(0.015)	(0.028)	(0.029)	(0.034)	(0.032)	(0.035)	(0.036)
No demand	0.720	0.537		0.485	0.888		0.355				
	(0.029)	(0.032)		(0.032)	(0.020)		(0.030)				
Negative demand	0.622	0.424	0.335	0.296	0.511	0.562	0.244	0.309	0.357	0.295	0.217
	(0.031)	(0.031)	(0.033)	(0.029)	(0.033)	(0.037)	(0.028)	(0.033)	(0.033)	(0.034)	(0.030)
Panel B: Sensitivity (Positive - Negative)											
Raw data	0.181	0.281	0.366	0.480	0.431	0.252	0.407	0.263	0.306	0.112	0.168
	(0.040)	(0.043)	(0.046)	(0.039)	(0.036)	(0.047)	(0.040)	(0.047)	(0.047)	(0.049)	(0.047)
Z-score	0.372	0.563	0.773	1.050	0.861	0.507	0.901	0.567	0.637	0.245	0.406
	(0.083)	(0.087)	(0.098)	(0.086)	(0.073)	(0.094)	(0.089)	(0.102)	(0.097)	(0.106)	(0.114)
	[0.001]	[0.001]		[0.001]	[0.001]		[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)	0.169	0.337		0.635	0.108		0.654				
	(0.079)	(0.088)		(0.092)	(0.050)		(0.092)				
	[0.023]	[0.001]		[0.001]	[0.011]		[0.001]				
Negative - Neutral (z-score)	-0.203	-0.226		-0.415	-0.754		-0.247				
	(0.089)	(0.089)		(0.095)	(0.077)		(0.090)				
	[0.022]	[0.003]		[0.001]	[0.001]		[0.002]				
Observations	727	728	404	731	714	365	770	409	421	382	371

Note: This table uses data from incentivized MTurk respondents with strong demand treatments. The outcome variables take value one if the respondents believed that the experimenter wanted a high action. Panel A displays mean beliefs with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of beliefs 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.

Table A9—Belief about the experimental hypothesis in response to the weak demand treatments

	Belief: Time	Belief: Risk	Belief: Ambiguity Aversion	Belief: Effort 0 cent bonus	Belief: Effort 1 cent bonus	Belief: Lying	Belief: Dictator Game	Belief: Ult. Game 1	Belief: Ult. Game 2	Belief: Trust Game 1	Belief: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.788	0.749	0.704	0.782	0.963	0.871	0.464	0.726	0.732	0.628	0.682
	(0.028)	(0.028)	(0.033)	(0.030)	(0.014)	(0.023)	(0.032)	(0.033)	(0.031)	(0.036)	(0.038)
No demand		0.534					0.160				
		(0.031)					(0.024)				
Negative demand	0.458	0.261	0.222	0.231	0.326	0.528	0.106	0.354	0.359	0.296	0.182
	(0.034)	(0.029)	(0.030)	(0.030)	(0.034)	(0.036)	(0.019)	(0.036)	(0.034)	(0.035)	(0.028)
Panel B: Sensitivity (Positive - Negative)											
Raw data	0.331	0.488	0.482	0.552	0.637	0.343	0.358	0.373	0.372	0.333	0.500
	(0.044)	(0.040)	(0.044)	(0.042)	(0.037)	(0.042)	(0.037)	(0.049)	(0.046)	(0.050)	(0.047)
Z-score	0.681	0.978	0.982	1.244	1.286	0.706	0.836	0.825	0.750	0.731	1.161
	(0.092)	(0.080)	(0.090)	(0.096)	(0.074)	(0.087)	(0.086)	(0.108)	(0.092)	(0.110)	(0.109)
		[0.001]					[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)		0.431					0.708				
		(0.084)					(0.092)				
		[0.001]					[0.001]				
Negative - Neutral (z-score)		-0.547					-0.128				
		(0.084)					(0.071)				
		[0.001]					[0.024]				
Observations	422	739	390	388	381	412	758	360	411	352	346

Note: This table uses data from incentivized MTurk respondents with weak demand treatments. The outcome variables take value one if the respondents believed that the experimenter expected a high action. Panel A displays mean beliefs with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of beliefs 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.

TABLE A10—BELIEF ABOUT THE EXPERIMENTAL HYPOTHESIS IN RESPONSE TO THE STRONG DEMAND TREATMENTS

	Belief: Time	Belief: Risk	Belief: Ambiguity Aversion	Belief: Effort 0 cent bonus	Belief: Effort 1 cent bonus	Belief: Lying	Belief: Dictator Game	Belief: Ult. Game 1	Belief: Ult. Game 2	Belief: Trust Game 1	Belief: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.727	0.595	0.593	0.727	0.934	0.788	0.454	0.670	0.659	0.578	0.588
	(0.028)	(0.033)	(0.034)	(0.029)	(0.016)	(0.030)	(0.030)	(0.032)	(0.032)	(0.035)	(0.037)
No demand	0.682	0.484		0.423	0.855		0.143				
	(0.030)	(0.032)		(0.032)	(0.023)		(0.022)				
Negative demand	0.639	0.420	0.400	0.267	0.576	0.625	0.186	0.284	0.435	0.290	0.243
	(0.031)	(0.031)	(0.035)	(0.028)	(0.033)	(0.037)	(0.025)	(0.032)	(0.035)	(0.034)	(0.031)
Panel B: Sensitivity (Positive - Negative)											
Raw data	0.089	0.175	0.193	0.459	0.358	0.163	0.268	0.386	0.224	0.288	0.345
	(0.042)	(0.045)	(0.049)	(0.040)	(0.036)	(0.047)	(0.039)	(0.046)	(0.047)	(0.049)	(0.048)
Z-score	0.183	0.351	0.393	1.036	0.722	0.336	0.624	0.855	0.451	0.634	0.801
	(0.087)	(0.090)	(0.100)	(0.091)	(0.074)	(0.097)	(0.092)	(0.101)	(0.095)	(0.107)	(0.112)
	[0.118]	[0.001]		[0.001]	[0.001]		[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)	0.093	0.222		0.685	0.159		0.725				
	(0.085)	(0.091)		(0.097)	(0.056)		(0.087)				
	[0.268]	[0.015]		[0.001]	[0.001]		[0.001]				
Negative - Neutral (z-score)	-0.090	-0.128		-0.350	-0.563		0.101				
	(0.090)	(0.089)		(0.096)	(0.080)		(0.077)				
	[0.268]	[0.052]		[0.001]	[0.001]		[0.069]				
Observations	727	728	404	731	714	365	770	409	421	382	371

Note: This table uses data from incentivized MTurk respondents with strong demand treatments. The outcome variables take value one if the respondents believed that the experimenter expected a high action. Panel A displays mean beliefs with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of beliefs 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.

Table A11—Overview of studies varying anonymity in dictator games

Study	Description of the treatment	Sample	Sample Size	Stake Size	Effect Size	Statistical significance
Hoffman et al. (1994)	Double blind compared to single blind	Student sample from the University of Arizona	101	$10	61 percent reduction in giving	p<0.01
Hoffman, McCabe and Smith (1996)	Double blind compared to single blind	Student sample from the University of Arizona	114	$10	37 percent reduction in giving	p<0.01
Bolton, Katok and Zwick (1998)	Double blind compared to single blind	Student sample at Penn State University	60	$5	22 percent increase in giving	p>0.1
Barmettler, Fehr and Zehnder (2012)	Double blind compared to single blind	Student samples from the University of Zurich (UZH)	103	20 Swiss Frank ($22)	16.8 percent reduction in giving	p>0.1
Cilliers, Dube and Siddiqi (2015)	Presence of non-foreign experimenter vs.presence of a white foreign experimenter	Poor households from Sierra Leone	708	4000 Leones (approximately $1)	16 percent reduction in giving	p<0.01
Our estimates based on demand treatments						
de Quidt et al. (2018)	Weak negative demand treatment compared to weak positive demand treatment	MTurk respondents	515	$1	17 percent reduction in giving	p<0.01
de Quidt et al. (2018)	Strong negative demand treatment compared to strong positive demand treatment	MTurk respondents	511	$1	42 percent reduction in giving	p<0.01

Note: This table provides an overview of dictator game studies which vary the anonymity of experimenter-subject interactions and the presence of a foreign (white) experimenter. Our estimates of treatment effects for the studies by Hoffman et al. (1994) and Hoffman, McCabe and Smith (1996) are based on inspection of the cumulative distribution functions and and probability distribution functions reported in the paper (details of our calculations are available upon request). These papers did not report mean behavior across treatment arms. In Hoffman et al. (1994) we compare behavior in "Double Blind treatment 1" and "Double Blind treatment 2" to behavior in the "Dictator random entitlement, exchange". In Hoffman, McCabe and Smith (1996) we compare behavior in "Double Blind treatment 1" and "Double Blind treatment 2" to behavior in the "Single Blind 1" condition. In Bolton, Katok and Zwick (1998) we compare behavior in the "Anonymity" condition to behavior in the "6card1game" condition. In Barmettler, Fehr and Zehnder (2012) we compare behavior in the "Double Anonymity" condition to behavior in the "Single Anonymity" condition. In Cilliers, Dube and Siddiqi (2015) we compare behavior when a white foreigner was or was not present in the session. The average reduction in giving across the studies using equal weights is a 21.76 percent, or 20.37 percent when weighted by sample size.

Table A12—Overview of standard deviations across 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: Weak Demand											
Positive demand	0.385	0.348	0.341	0.193	0.165	0.170	0.222	0.189	0.194	0.314	0.217
No demand	.	0.339	.	.	.	.	0.234	.	.	.	.
Negative demand	0.389	0.317	0.334	0.190	0.178	0.158	0.226	0.170	0.182	0.329	0.172
Panel B: Strong Demand											
Positive demand	0.379	0.331	0.340	0.177	0.179	0.172	0.267	0.184	0.202	0.334	0.234
No demand	0.386	0.340	.	0.182	0.184	.	0.246	.	.	.	.
Negative demand	0.437	0.322	0.319	0.176	0.162	0.183	0.229	0.189	0.209	0.291	0.205

Note: This table uses data from incentivized MTurk respondents with weak and strong demand treatments and displays the standard deviations across the different demand treatment arms.

B. Theoretical Appendix

B1. Proof of Proposition 1 (Monotonicity)

We require that $a^+(\zeta) \geq a^L(\zeta) \geq a^-(\zeta)$. We are therefore interested in the sign of $\phi(E[h|h^T, h^L(\zeta)] - E[h|h^L(\zeta)])$. We have:

$$\begin{aligned}\phi(E[h|h^T, h^L(\zeta)] - E[h|h^L(\zeta)]) &= \phi\left(\frac{h^L(\zeta)p^L(\zeta) + h^T p^T}{1 + h^L(\zeta)p^L(\zeta)h^T p^T} - h^L(\zeta)p^L(\zeta)\right) \\ &= \phi h^T p^T \frac{(1 - h^L(\zeta)^2 p^L(\zeta)^2)}{1 + h^L(\zeta)p^L(\zeta)h^T p^T}\end{aligned}$$

Because we assumed that $p^L(\zeta) < 1$, this expression has the same sign as $\phi h^T p^T$. We want to show that $\phi(E[h|h^T = 1, h^L(\zeta)] - E[h|h^L(\zeta)]) \geq 0$ and $\phi(E[h|h^T = -1, h^L(\zeta)] - E[h|h^L(\zeta)]) \leq 0$. This follows trivially when $p^T = 0$. When $p^T > 0$ if follows if and only if $\phi \geq 0$.

B2. Proof of Proposition 2 (Bounding)

In the Bayesian model, given $\phi \geq 0$ (Monotonicity), the action is larger or smaller than $a(\zeta)$ when $\phi E[h|h^T, h^L] \geq 0$ or $\phi E[h|h^T, h^L] \leq 0$ respectively. Given that $\phi \geq 0$, we need $E[h|h^T = 1, h^L] \geq 0$ and $E[h|h^T = -1, h^L] \leq 0$. This is guaranteed if h^T and h^L have the same sign, so we simply need to check whether it holds when the demand treatment and latent demand are in opposite directions, i.e. $E[h|h^T = 1, h^L = -1] \geq 0$ and $E[h|h^T = -1, h^L = 1] \leq 0$. Given our restriction $p^L(\zeta) < 1$, inspection of (7) reveals that these conditions hold if and only if $p^T \geq p^L(\zeta)$, i.e. the decision-maker perceives the demand treatment as at least as informative about h as the latent demand signal.

B3. Conditions for Monotone Sensitivity

Assumption 3 (Monotone Sensitivity) assumes that sensitivity $S(\zeta) = a^+(\zeta) - a^-(\zeta)$ is (strictly) monotone in the size of the latent demand effect $\left|a^L(\zeta) - a(\zeta)\right|$. Here we examine cases under which that is and is not the case. We assume throughout that Assumptions 1 and 2 hold.

Variation driven by ϕ.

We are interested in how ϕ affects latent demand ($d\left|a^L(\zeta) - a(\zeta)\right|/d\phi$) and sensitivity ($dS(\zeta)/d\phi$). From (5) we obtain:

$$\frac{d(a^L(\zeta) - a(\zeta))}{d\phi} = -\frac{h^L(\zeta)p^L(\zeta)}{v_{11}(a^L(\zeta), \zeta)}$$

which has the same sign as $h^L(\zeta)$, allowing us to write $\frac{d\left|a^L(\zeta)-a(\zeta)\right|}{d\phi} = -\frac{p^L(\zeta)}{v_{11}(a^L(\zeta),\zeta)} \geq 0$.

Turning to sensitivity, we have:

$$\begin{aligned}\frac{dS(\zeta)}{d\phi} &= \frac{da^+(\zeta)}{d\phi} - \frac{da^-(\zeta)}{d\phi} \\ &= -\frac{1}{v_{11}(a^+(\zeta),\zeta)}\frac{h^L(\zeta)p^L(\zeta) + p^T}{1 + h^L(\zeta)p^L(\zeta)p^T} + \frac{1}{v_{11}(a^-(\zeta),\zeta)}\frac{h^L(\zeta)p^L(\zeta) - p^T}{1 - h^L(\zeta)p^L(\zeta)p^T}\end{aligned}$$

By Assumption 2, $h^L(\zeta)p^L(\zeta)+p^T \geq 0$ and $h^L(\zeta)p^L(\zeta)+p^T \leq 0$, so both terms are positive, i.e. $\frac{dS(\zeta)}{d\phi} \geq 0$. Therefore Monotone Sensitivity holds and any set of environments that differ only in ϕ constitutes a comparison class, i.e. for such environments, sensitivity is informative about the magnitude of latent demand effects.

EXAMPLE 2: *Suppose participant pool A is more concerned for pleasing the experimenter than participant pool B. Then latent demand effects and sensitivity will be larger in magnitude in participant pool A.*

Variation driven by v.

Suppose that ζ can be separated into a parameter, z, and a remainder term, ζ', that v is differentiable in z and that ϕ, h^L and p^L do not depend on z. z could be a preference parameter (e.g. risk aversion) or a design parameter (e.g. the scale of incentives). We write $U(a, \zeta', z) = v(a, \zeta', z) + a\phi(\zeta')E[h|\zeta']$ and modify the

first-order conditions accordingly.

$$\begin{aligned}\frac{d(a^L(\zeta',z)-a(\zeta',z))}{dz} &= \frac{da^L(\zeta',z)}{dz} - \frac{da(\zeta',z)}{dz} \\ &= -\left[\frac{v_{13}(a^L(\zeta',z),\zeta',z)}{v_{11}(a^L(\zeta',z),\zeta',z)} - \frac{v_{13}(a(\zeta',z),\zeta',z)}{v_{11}(a(\zeta',z),\zeta',z)}\right] \\ \frac{dS(\zeta',z)}{dz} &= -\left[\frac{v_{13}(a^+(\zeta',z),\zeta',z)}{v_{11}(a^+(\zeta',z),\zeta',z)} - \frac{v_{13}(a^-(\zeta',z),\zeta',z)}{v_{11}(a^-(\zeta',z),\zeta',z)}\right]\end{aligned}$$

It is clear from inspecting these conditions that we need to know how v_{13}/v_{11} varies with a, i.e.:

$$\frac{d\frac{v_{13}(a,\zeta',z)}{v_{11}(a,\zeta',z)}}{da} = \frac{v_{11}(a,\zeta',z)v_{113}(a,\zeta',z) - v_{111}(a,\zeta',z)v_{13}(a,\zeta',z)}{v_{11}(a,\zeta',tz)}$$

It is difficult to make general statements about these objects for general utility functions, so we focus attention on two special cases of interest.

Multiplicative separability.

Suppose that $v(a,\zeta',z) = \nu(a,\zeta')f(z)$ and define z such that $f'(z) > 0$. Then

$$\begin{aligned}\frac{d\left(a^L(\zeta',z)-a(\zeta',z)\right)}{dz} &= -f'(z)\left[\frac{\nu_1(a^L(\zeta',z),\zeta')}{\nu_{11}(a^L(\zeta',z),\zeta')} - \frac{\nu_1(a(\zeta',z),\zeta')}{\nu_{11}(a(\zeta',z),\zeta')}\right] \\ &= -f'(z)\frac{\nu_1(a^L(\zeta',z),\zeta')}{\nu_{11}(a^L(\zeta',z),\zeta')}\end{aligned}$$

Since by concavity $\nu_1(a,\zeta') > 0$ for $a < a(\zeta',z)$ and $\nu_1(a,\zeta') < 0$ for $a > a(\zeta',z)$, we have $\frac{d\left|a^L(\zeta',z)-a(\zeta',z)\right|}{dz} \leq 0$. Similarly

$$\frac{dS(\zeta)}{dz} = -f'(z)\left[\frac{\nu_1(a^+(\zeta',z),\zeta')}{\nu_{11}(a^+(\zeta',z),\zeta')} - \frac{\nu_1(a^-(\zeta',z),\zeta')}{\nu_{11}(a^-(\zeta',z),\zeta')}\right]$$

Since $\nu_1(a^+(\zeta',z),\zeta') \leq 0$ and $\nu_1(a^-(\zeta',z),\zeta') \geq 0$, we have $\frac{dS(\zeta)}{dz} \leq 0$. Therefore Monotone Sensitivity holds and any set of environments that varies only in z is a valid comparison set.

Intuitively, this case captures changes in the slope of payoffs that leave the optimal natural action unchanged. For example, an increase in the scale of incentives that makes the payoff function "more concave" around the natural action makes

deviating from the natural action more costly and so decreases the magnitude of latent demand and sensitivity.

EXAMPLE 3 (Belief scoring): *Consider a belief-reporting task rewarded by a quadratic scoring rule. A risk-neutral participant reports a belief, a, which is the probability of an event A. He is paid $\frac{z}{2}\left[1-(\mathbb{I}[A]-a)^2\right]$ where $\mathbb{I}[A]=1$ if A is true and 0 otherwise. The utility function is $U(a,\zeta',z)=\frac{z}{2}\left[1-\mu(1-a)^2-(1-\mu)(-a)^2\right]+a\phi(\zeta')E[h|\zeta']$, so $f(z)=z$. The optimal action solves $z\left[\mu(1-a^*)-(1-\mu)a^*\right]+\phi(\zeta')E[h|\zeta']=0$ or $a^*=\mu+\frac{\phi(\zeta')E[h|\zeta']}{z}$. Increases in z are equivalent to decreases in ϕ and decrease both the magnitude of latent demand effects, and sensitivity.*

Additive separability.

Suppose that $v(a,\zeta',z)=v(a,\zeta')+af(z)$ and define z such that $f'(z)>0$. Then:

$$\frac{d(a^L(\zeta',z)-a(\zeta',z))}{dz}=-f'(z)\left[\frac{1}{\nu_{11}(a^L(\zeta',z),\zeta')}-\frac{1}{\nu_{11}(a(\zeta',z),\zeta')}\right]$$

and

$$\frac{dS(\zeta)}{dz}=-f'(z)\left[\frac{1}{\nu_{11}(a^+(\zeta',z),\zeta')}-\frac{1}{\nu_{11}(a^-(\zeta',z),\zeta')}\right]$$

What matters in this case is how the concavity of v (and therefore ν) with respect to a varies with a. Suppose $\nu_{111}<0$, so ν_{11} is decreasing in a, i.e. concavity is increasing. Then $\frac{dS(\zeta)}{dz}<0$, i.e. increases in z decrease sensitivity. If $a^L(\zeta',z)>a(\zeta',z)$ then $\frac{d(a^L(\zeta',z)-a(\zeta',z))}{dz}<0$ and if $a^L(\zeta',z)<a(\zeta',z)$ then $\frac{d(a^L(\zeta',z)-a(\zeta',z))}{dz}>0$, so $\frac{d\left|a^L(\zeta',z)-a(\zeta',z)\right|}{dz}<0$ and Monotone Sensitivity holds. Monotone Sensitivity also holds (with the inequalities reversed) for $\nu_{111}>0$.

EXAMPLE 4 (Effort provision): *A participant performs a real-effort task for piece rate z with cost of effort $C(a)$, $C'>0$,$C''>0$, $C'''>0$. $U(a,\zeta',z)=za-C(a)+a\phi(\zeta')E[h|\zeta']$. The optimal action a^* solves $z-C'(a^*)+\phi(\zeta')E[h|\zeta']=0$. As z increases, a^* increases and responsiveness to latent demand or demand treatments decreases.*

Variation driven by inattention.

Suppose that with some probability ξ the participant is an attentive type who pays careful attention to the decision-making environment, and with probability

$1-\xi$, he is inattentive. When inattentive, he takes some action $a^I(\zeta)$. $a^I(\zeta)$ might be equal to $a(\zeta)$, in which case the participant is only inattentive to experimenter demand, but it might differ if the participant is also inattentive to other design features.

While until now we have treated the actions as those of a representative agent, for this analysis it is more natural to work with expected or average actions over a sample. Denote by $\bar{a}(\zeta) = \xi a(\zeta) + (1-\xi)a^I(\zeta)$ the expected natural action, define $\bar{a}^L(\zeta), \bar{a}^+(\zeta), \bar{a}^-(\zeta)$ equivalently and let $\bar{S}(\zeta) = \bar{a}^+(\zeta) - \bar{a}^-(\zeta)$. The latent demand effect is now equal to $\left|\bar{a}^L(\zeta) - \bar{a}(\zeta)\right| = \xi \left|a^L(\zeta) - a(\zeta)\right|$, while $\bar{S}(\zeta) = \xi S(\zeta)$. Hence, if the variation in latent demand effects is driven by variation in attention, ξ, Monotone Sensitivity will hold, and any set of environments that varies only in participant attentiveness is a valid comparison set. Note that since we have assumed the participant is inattentive to both latent demand and the demand treatment, Bounding will hold if $p^T \geq p^L$ as before.

Variation driven by beliefs.

Consider changes to the environment that influence behavior only by altering participants' beliefs about the experimenter's objective, i.e. we consider variation in $h^L(\zeta)p^L(\zeta)$. Call this term H. $a(\zeta)$ is unaffected, so:

$$\frac{d(a^L(\zeta) - a(\zeta))}{dH} = -\frac{\phi(\zeta)}{v_{11}(a^L(\zeta), \zeta)} \geq 0$$

and therefore $\frac{d\left|a^L(\zeta)-a(\zeta)\right|}{dH} = -\frac{\phi(\zeta)}{v_{11}(a^L(\zeta),\zeta)} \times \text{sign}(a^L(\zeta) - a(\zeta)) = -\frac{\phi(\zeta)h^L(\zeta)}{v_{11}(a^L(\zeta),\zeta)}$ which is positive when $h^L(\zeta) = 1$ (because an increase in H means the participant's beliefs are shifting toward certainty that the experimenter wants a high action) and negative when $h^L(\zeta) = -1$ (because the participant is becoming more uncertain about the experimenter's wishes).

Next we turn to demand treatment effects. First we derive the response of the participant's posterior:

$$\begin{aligned}\frac{d\frac{H+h^Tp^T}{1+Hh^Tp^T}}{dH} &= \frac{\left(1 + Hh^Tp^T\right) - \left(H + h^Tp^T\right)h^Tp^T}{\left(1 + Hh^Tp^T\right)^2} \\ &= \frac{1 - \left(h^Tp^T\right)^2}{\left(1 + Hh^Tp^T\right)^2} = \frac{1 - p^{T2}}{\left(1 + Hh^Tp^T\right)^2}\end{aligned}$$

So:

$$\frac{dS(\zeta)}{dH} = -\phi(\zeta)(1-p^{T2})\left[\frac{1}{(1+Hp^T)^2 v_{11}(a^+(\zeta),\zeta)} - \frac{1}{(1-Hp^T)^2 v_{11}(a^-(\zeta),\zeta)}\right]$$

The sign of this expression depends on the sign of H and how v_{11} changes with a. However, it is straightforward to see that Monotone Sensitivity *will not* hold in general, and in fact sensitivity will tend to be higher when latent demand is weaker. To see this, consider the simple case where v_{11} is constant. Then we have:

$$\begin{aligned}\frac{dS(\zeta)}{dH} &= -\frac{\phi(\zeta)(1-p^{T2})}{v_{11}}\left[\frac{\left(1-Hp^T\right)^2 - \left(1+Hp^T\right)^2}{(1+Hp^T)^2(1-Hp^T)^2}\right] \\ &= -\frac{\phi(\zeta)(1-p^{T2})}{v_{11}}\left[\frac{-4Hp^T}{(1+Hp^T)^2(1-Hp^T)^2}\right]\end{aligned}$$

which is positive when $h^L = -1$ and negative when $h^L = 1$, i.e. it has the opposite sign to $\frac{d\left|a^L(\zeta)-a(\zeta)\right|}{dH}$. The reason is that as H approaches zero, the participant becomes more uncertain about the experimenter's wishes and is therefore very responsive to the new information in the demand treatments. Meanwhile as H approaches 1 or -1, the participant is very confident about the value of h. Although his confidence can be undermined by a demand treatment in the opposite direction, he responds little to a demand treatment that confirms his beliefs, so sensitivity is low.

B4. Defiers

In this section we discuss defiance. We first derive a special case that illustrates how valid bounds can be obtained even when some participants defy the experimenter. Then we present three examples where defier behavior causes our key assumptions to break down.

Because our concern is with bounding rather than point identification, the method is able to tolerate some defiance. To illustrate, suppose that v is homogeneous across individuals, quadratic in a, and normalized such that $v_1(a,\zeta) = b(\zeta) - a$ where b is a constant. The natural action is equal to b for all individuals. Beliefs and ϕ are heterogeneous across individuals, indexed by i. For compactness, label the beliefs $H_i^L := h_i^L p_i^L$, $H_i^+ := (H_i^L + p^T)/(1 + H_i^L p^T)$ and

$H_i^- := (H_i^L - p^T)/(1 - H_i^L p^T)$. Under our assumptions, the actions of interest are given by:

$$a_i^L = b + \phi_i H_i^L \qquad a_i^+ = b + \phi_i H_i^+ \qquad a_i^- = b + \phi_i H_i^-$$

Then, for Bounding to hold on average in the population, we require $E[\phi_i H_i^+] \geq 0 \geq E[\phi_i H_i^-]$, where expectations are over participants. If $p^T \geq p_i^L$ for all individuals, then $H_i^+ \geq 0 \geq H_i^-$, so the conditions are equivalent to weighted averages of ϕ having the correct sign, where the weights are the beliefs. A special case of interest is that where all individuals have identical H_i^+ and H_i^- (this is the case if latent demand (H_i^L) is the same for all individuals, or if $p^T = 1$). Then both conditions reduce to $E[\phi_i] \geq 0$, i.e. Bounding holds if the average participant is a complier.[1]

Now we provide three simple examples where defier behavior causes our key assumptions to break down. First we show that it is possible for Bounding to hold without Monotonicity, second that it is possible for Monotonicity to hold without Bounding, and third that both can fail while retaining well-ordered bounds.

Let all decision makers share $v(a) = -a^2$, so the natural action $a = 0$. 2/3 of the population are compliers with $\phi = \phi_C = 1$ and 1/3 are defiers with $\phi = \phi_D = -1$. Latent demand signals are assumed common within complier/defier groups but different between compliers and defiers. They equal $H_C^L = h_C^L p_C^L$ and $H_D^L = h_D^L p_D^L$ respectively, with corresponding beliefs following the demand treatments equal to:

$$H_i^+ = \frac{H_i^L + p^T}{1 + H_i^L p^T} \qquad H_i^- = \frac{H_i^L - p^T}{1 - H_i^L p^T}$$

We retain the assumption of common p^T. Then the observed average actions under latent demand, positive and negative demand treatments are:

$$E[a^L] = \frac{1}{3}(2H_C^L - H_D^L) \qquad E[a^+] = \frac{1}{3}(2H_C^+ - H_D^+) \qquad E[a^-] = \frac{1}{3}(2H_C^- - H_D^-)$$

Our first example shows that Bounding can hold without Monotonicity. Thus

[1]For Monotonicity to hold on average we require $E[\phi_i(H_i^+ - H_i^L)] \geq 0 \geq E[\phi_i(H_i^L - H_i^-)]$. Since $H_i^+ - H_i^L > 0$ and $H_i^L - H_i^- < 0$, these conditions require that a weighted average of ϕ has the correct sign, where the weights are the belief *changes* induced by the demand treatments. Violations of Monotonicity or, in the extreme case, reversed bounds ($a^- > a^+$), are clear cause for concern. However it is possible for Monotonicity to hold on average while Bounding fails and vice versa.

a Monotonicity failure does not imply a failure of Bounding, but it is a warning sign of the presence of defiers.

EXAMPLE 5 (Bounding without Monotonicity): *Suppose* $H_C^L = 0.5$, $H_D^L = -0.5$ *and* $p^T = 1$. *Then* $E[a^L] = 0.5$, $E[a^+] = 1/3$ *and* $E[a^-] = -1/3$. *Therefore* $E[a^-] < a < E[a^+] < E[a^L]$.

Our second example shows that Bounding can fail while Monotonicity holds. This is possible in the basic model if $p^T < p^L$, but can also be caused by defiance.

EXAMPLE 6 (Monotonicity without Bounding.): *Suppose* $H_C^L = 0.5$, $H_D^L = -0.5$ *and* $p^T = 0.5$. *Then* $E[a^L] = 0.5$, $E[a^+] = \frac{8}{15}$ *and* $E[a^-] = \frac{4}{15}$. *Thus* $a < E[a^-] < E[a^L] < E[a^+]$.

Our third example shows that both Bounding and Monotonicity can fail, while still producing well-ordered bounds (i.e. $a^+ > a^-$).

EXAMPLE 7 (No Bounding or Monotonicity.): *Let* $H_C^L = 0.75$, $H_D^L = 0$, *and* $p^T = 0.75$. *Then* $E[a^L] = 0.5$, $E[a^+] = \frac{39}{100}$ *and* $E[a^-] = 1/4$. *Therefore* $a < E[a^-] < E[a^+] < E[a^L]$.

B5. Extension: learning about ϕ

A possible interpretation of our demand treatments is that they signal not only the direction of the experimenter's objective, but the salience or intensity of her preference over objectives. For instance, "do me a favor" suggests that the choice is important. In this section we extend the model to incorporate this feature, allowing ϕ to depend upon a belief about the "importance" of the objective. We assume that the decision-maker responds more strongly to experimenter demand when they believe that complying with the objective it is more important, and that this belief depends both on latent demand and the demand treatments. Specifically, we now assume that the decision-maker's preferences are:

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

where $g \in \{0, 1\}$ captures whether conforming to h is important (1) or unimportant (0) to the experimenter. ϕ remains the decision-maker's preference for pleasing the experimenter, which is now scaled by g, i.e. the decision-maker internalizes the perceived importance of the objective. We assume that g and

h are believed independent (i.e. direction and importance are independent), so $E[gh|\zeta] = E[g|\zeta]E[h|\zeta]$. We also assume for simplicity is that the decision-maker's prior $E[g] = 0.5$.

Now, ζ contains two signals, $h^L(\zeta)$, defined as before, and $g^L(\zeta) \in \{0, 1\}$, where $E[g|g^L(\zeta)] = E[g|g^L(\zeta), \zeta]$ (i.e. g^L is a sufficient statistic). g^L is believed to equal g with probability $q^L(\zeta) < 1$ and pure independent noise otherwise. We show below that $E[g|g^L(\zeta)] = \frac{1}{2} + q^L\left(g^L - \frac{1}{2}\right)$.

Similarly, a demand treatment is now two signals (h^T, g^T), where h^T is defined as before and $g^T \in \{0, 1, \emptyset\}$. $g^T = \emptyset$ corresponds to the case where no treatment is used, $g^T = 0$ signals to the participant that their action is not important to the experimenter, and $g^T = 1$ signals that it is.

Conditional on sending a demand treatment, g^T is believed to equal g with probability q^T and otherwise be pure noise independent of all other signals. We show below that the Bayesian posterior is:

$$E[g|g^T, g^L(\zeta)] = \frac{\frac{1}{2} + q^L(\zeta)\left(g^L(\zeta) - \frac{1}{2}\right) + q^T\left(g^T - \frac{1}{2}\right) + q^T q^L(\zeta)\left(\mathbb{I}[g^T = g^L(\zeta)] - \frac{1}{2}\right)}{1 + 2q^T q^L(\zeta)\left(\mathbb{I}[g^T = g^L(\zeta)] - \frac{1}{2}\right)}$$

We assume that g^T can be varied independently of h^T and will be held constant within a typical pair of positive and negative demand treatments.

For Bounding to hold, we now need:

$$\phi(\zeta)E[g|g^T, g^L(\zeta)]E[h|h^T = 0, h^L(\zeta)] \leq 0 \leq \phi(\zeta)E[g|g^T, g^L(\zeta)]E[h|h^T = 1, h^L(\zeta)]$$

Since $E[g|g^T, g^L(\zeta)] \geq 0$ our Bounding condition does *not* depend on how the demand treatments affect beliefs about g, all we require is $\phi(\zeta) \geq 0$ and $p^T \geq p^L(\zeta)$ as before.[2]

However, beliefs about g do affect the width of the bounds: sensitivity is increas-

[2]For Monotonicity to hold, we require

$$\phi(\zeta)E[g|g^T, g^L(\zeta)]E[h|h^T = 0, h^L(\zeta)] \leq \phi(\zeta)E[g|g^L(\zeta)]E[h|h^L(\zeta)] \leq \phi(\zeta)E[g|g^T, g^L(\zeta)]E[h|h^T = 1, h^L(\zeta)]$$

We can write

$$\phi(\zeta)\frac{E[h|h^T = 0, h^L(\zeta)]}{E[h|h^L(\zeta)]} \leq \phi(\zeta)\frac{E[g|g^L(\zeta)]}{E[g|g^T, g^L(\zeta)]} \leq \phi(\zeta)\frac{E[h|h^T = 1, h^L(\zeta)]}{E[h|h^L(\zeta)]}$$

We see that $\phi(\zeta) \geq 0$ is necessary but not sufficient for Monotonicity, we also need that $E[g|g^T, g^L(\zeta)]$ is neither "too big" nor "too small" relative to $E[g|g^L(\zeta)]$. Intuitively, if $g^T = 1$ the demand treatments shift all actions further away from the natural action, while if $g^T = 0$. all actions are shifted toward the natural action. $g^T = 1$ and $p^T \geq p^L$ are sufficient for Monotonicity to hold.

ing in $E[g|g^T, g^L(\zeta)]$. The tightest bounds are obtained when $E[g|g^T, g^L(\zeta)] = 0$, which obtains when $g^T = 0$ and $q^T = 1$. More generally, the bounds are tightened by signaling that acting according to the experimenter's objective is not important ($g^T = 0$), or if $g^T = 1$ by minimizing q^T. We suspect that it may be difficult in practice to both strongly signal the direction of the objective (large p^T), which is required for Bounding, and that the objective is not important ($g^T = 0$), so reasonable demand treatments are likely to be those that strongly signal a directional objective while keeping salience low, i.e. large p^T and small q^T with $g^T = 1$.

Derivation of $E[g|g^L(\zeta)]$ and $E[g|g^T, g^L(\zeta)]$

Let the prior belief be $\frac{1}{2}$.

$$
\begin{aligned}
E[g|g^L = y] = Pr(g = 1|g^L = y) &= \frac{A}{B} \\
A &= Pr(g^L = y|g = 1)Pr(g = 1) \\
B &= Pr(g^L = y|g = 1)Pr(g = 1) + Pr(g^L = y|g = 0)Pr(g = 0)
\end{aligned}
$$

Since $Pr(g = j|g^L = y) = \frac{1}{2}(1 - q^L) + q^L \mathbb{I}[y = j]$ and $Pr(g = j) = \frac{1}{2}$ we have

$$
\begin{aligned}
A &= \frac{1}{2}\left(\frac{1}{2}(1 - q^L) + q^L \mathbb{I}[y = 1]\right) \\
&= \frac{1}{2}\left(\frac{1}{2} + q^L\left(g^L - \frac{1}{2}\right)\right) \\
B &= \frac{1}{2}\left[\left(\frac{1}{2}(1 - q^L) + q^L \mathbb{I}[y = 1]\right) + \left(\frac{1}{2}(1 - q^L) + q^L \mathbb{I}[y = 0]\right)\right] = \frac{1}{2}
\end{aligned}
$$

Therefore, $E[g|g^L(\zeta)] = \frac{1}{2} + q^L\left(g^L - \frac{1}{2}\right)$.

Turning to $E[g|g^T, g^L(\zeta)]$, we have assumed that when $g^T = \emptyset$, $E[g|g^T, g^L] = E[g|g^L]$. After observing $g^T \neq \emptyset$, the participant forms a posterior:

$$
\begin{aligned}
E[g|g^T, g^L] = Pr(g = 1|g^T, g^L) &= \frac{A}{B} \\
A &= Pr(g^T = x|g = 1, g^L = y)Pr(g = 1|g^L = y) \\
B &= Pr(g^T = x|g = 1, g^L = y)Pr(g = 1|g^L = y) \\
&+ Pr(g^T = x|g = 0, g^L = y)Pr(g = 0|g^L = y)
\end{aligned}
$$

Using the following

$$
\begin{aligned}
Pr(g^T = x|g = j, g^L = y) &= \frac{1}{2}(1 - q^T) + q^T \mathbb{I}[x = j] \\
Pr(g = j|g^L = y) &= \frac{1}{2}(1 - q^L) + q^L \mathbb{I}[y = j]
\end{aligned}
$$

we have:

$$
\begin{aligned}
A &= \left(\frac{1}{2}(1 - q^T) + q^T \mathbb{I}[x = 1]\right)\left(\frac{1}{2}(1 - q^L) + q^L \mathbb{I}[y = 1]\right) \\
&= \left(\frac{1}{2}(1 - q^T) + q^T g^T\right)\left(\frac{1}{2}(1 - q^L) + q^L g^L\right) \\
&= \frac{1}{2}q^L g^L + \frac{1}{2}q^T g^T - \frac{1}{2}q^T q^L \left(g^L(1 - g^T) + g^T(1 - g^L)\right) \\
&+ \frac{1}{4}(1 - q^T)(1 - q^L) \\
&= \frac{1}{2}q^L g^L + \frac{1}{2}q^T g^T - \frac{1}{2}q^T q^L \left(\mathbb{I}[g^L \neq g^T]\right) \\
&+ \frac{1}{4} - \frac{1}{4}q^T - \frac{1}{4}q^L + \frac{1}{4}q^T q^L \\
&= \frac{1}{2}q^L \left(g^L - \frac{1}{2}\right) + \frac{1}{2}q^T \left(g^T - \frac{1}{2}\right) - \frac{1}{2}q^T q^L \left(1 - \mathbb{I}[g^L = g^T]\right) \\
&+ \frac{1}{4} + \frac{1}{4}q^T q^L \\
&= \frac{1}{4} + \frac{1}{2}q^L \left(g^L - \frac{1}{2}\right) + \frac{1}{2}q^T \left(g^T - \frac{1}{2}\right) \\
&+ \frac{1}{2}q^T q^L \left(\mathbb{I}[g^T = g^L] - \frac{1}{2}\right)
\end{aligned}
$$

$$
\begin{aligned}
B &= \left(\frac{1}{2}(1-q^T)+q^T\mathbb{I}[x=1]\right)\left(\frac{1}{2}(1-q^L)+q^L\mathbb{I}[y=1]\right)\\
&+\left(\frac{1}{2}(1-q^T)+q^T\mathbb{I}[x=0]\right)\left(\frac{1}{2}(1-q^L)+q^L\mathbb{I}[y=0]\right)\\
&=\left(\frac{1}{2}(1-q^T)+q^Tg^T\right)\left(\frac{1}{2}(1-q^L)+q^Lg^L\right)\\
&+\left(\frac{1}{2}(1-q^T)+q^T(1-g^T)\right)\left(\frac{1}{2}(1-q^L)+q^L(1-g^L)\right)\\
&=\frac{1}{2}(1-q^T)q^L+\frac{1}{2}(1-q^L)q^T+\frac{1}{2}(1-q^T)(1-q^L)\\
&+q^Tq^L\mathbb{I}[g^T=g^L]\\
&=\frac{1}{2}+q^Tq^L\left(\mathbb{I}[g^T=g^L]-\frac{1}{2}\right)
\end{aligned}
$$

Therefore,

$$
E[g|g^T,g^L]=\frac{\frac{1}{2}+q^L\left(g^L-\frac{1}{2}\right)+q^T\left(g^T-\frac{1}{2}\right)+q^Tq^L\left(\mathbb{I}[g^T=g^L]-\frac{1}{2}\right)}{1+2q^Tq^L\left(\mathbb{I}[g^T=g^L]-\frac{1}{2}\right)}
$$

B6. Richer beliefs and correlated signals

Researchers sometimes give experimental participants instructions like "there are no right or wrong answers" or "we are only interested in what you think is the best choice." This can be thought of as a demand treatment that demands participants choose the natural action, $a(\zeta)$.

It is straightforward to analyze such treatments in our framework. In this section, we extend the model to allow h to take three values: $\{-1,0,1\}$, where $h=0$ captures the case where the experimenter wants the participant to choose the natural action. We call the action following $h^T=0$, $a^0(\zeta)$.

For simplicity we assume that the participant's prior belief is that each possibility is equally likely (i.e. is true with probability 1/3), so $E[h]=0$. ϵ and η are also believed to take each value with probability 1/3 and are independent. $h^L\in\{-1,0,1\}$ and $h^T\in\{-1,0,1,\emptyset\}$ and p^L and p^T are defined as before. We maintain the assumption that the participant infers nothing when the experimenter does not send a demand treatment ($h^T=\emptyset$).

We show below that the beliefs can be written as:

$$\text{(B1)} \qquad E[h|h^L] = p^L h^L$$

$$\text{(B2)} \quad E[h|h^T = \emptyset, h^L] = p^L h^L$$

$$\text{(B3)} \qquad E[h|h^T, h^L] = \frac{\frac{1}{3}(1-p^T)p^L h^L + \frac{1}{3}(1-p^L)p^T h^T + p^T p^L h^T \mathbb{I}[h^T = h^L]}{\frac{1}{3}\left(1 - p^T p^L\right) + p^T p^L \mathbb{I}[h^T = h^L]}$$

Bounding holds if $E[h|h^T = 1, h^L] \geq 0$ and $E[h|h^T = -1, h^L] \leq 0$. It is straightforward to check that the condition is the same as before: $p^T \geq p^L$.

What purpose, then, do $h^T = 0$ treatments serve? It is natural to think that demanding participants to take the natural action will eliminate demand effects, but under our assumptions, $h^T = 0$ does not in general elicit the natural action. Instead latent demand still influences the participant's action. We have:

$$E[h|h^T = 0, h^L] = \frac{\frac{1}{3}(1-p^T)p^L h^L}{\frac{1}{3}\left(1 - p^T p^L\right) + p^T p^L \mathbb{I}[h^L = 0]}$$

This expression equals zero if $p^T = 1$ (the demand treatment is perfectly informative), or $p^L h^L = 0$ (no latent demand), otherwise it has the same sign as $p^L h^L$. One interpretation is that while the participant takes at face value the experimenter's demand to choose the natural action, he might be unaware of the influence of other design features that nudge him in one direction or another.

Despite this negative result, $h^T = 0$ treatments can still be useful. First, they are informative about the *sign* of the bias due to latent demand. This is because $E[h|h^T = 0, h^L] \in [\min\{E[h|h^L], 0\}, \max\{E[h|h^L], 0\}]$ and therefore $a^0(\zeta) \in [\min\{a^L(\zeta), a(\zeta)\}, \max\{a^L(\zeta), a(\zeta)\}]$.[3] The action taken when $h^T = 0$ lies between the natural action and the action induced by latent demand, because the demand treatment shifts the participant's posterior toward zero.

Second, they can be used to obtain tighter bounds on $a(\zeta)$ if we know the direction of the latent demand effect. Suppose for example we know that $a^L(\zeta) \geq a(\zeta)$ (either from prior information or because we ran a treatment with $h^T = 0$ and verified that $a^0(\zeta) \leq a^L(\zeta)$). Then, the interval $[a^-(\zeta), a^0(\zeta)]$ gives a valid and tighter bound on $a(\zeta)$ than $[a^-(\zeta), a^+(\zeta)]$. Formally $a(\zeta) \in [a^-(\zeta), a^0(\zeta)] \subseteq [a^-(\zeta), a^+(\zeta)]$.[4]

[3] To see this, note that $\left|E[h|h^L] - E[h|h^T = 0, h^L]\right| \geq 0$ and both have the same sign.

[4] We thank Liad Weiss for pointing this out to us.

Finally, there is one important case in which $h^T = 0$ perfectly recovers the natural action, i.e. $a^0(\zeta) = a(\zeta)$. Suppose that instead of assuming that the signals h^T and h^L contain independent shocks, the participant perceives that h^L is a noisy signal of h^T. Formally, he believes that with probability $p^L < 1$, $h^L = h^T$ and with probability $(1 - p^L)$, $h^L = \epsilon$. Then, when h^T and h^L disagree, he knows that h^L is pure noise, when they agree h^L contains no more information than h^T. Hence, the participant disregards h^L after observing h^T and $E[h|h^T, h^L] = p^T h^T$. Then, sending $h^T = 0$ recovers the natural action: $E[h|h^T = 0, h^L] = 0, \forall h^L$. An advantage of our bounds is that they are valid whether or not h^T or h^L are perceived as independent, in other words they are conservative relative to the approach of simply measuring $a^0(\zeta)$.

To summarize, unless the demand treatment is perceived as fully informative ($p^T = 1$), signaling $h^T = 0$ does *not* induce the participant to take the natural action, i.e. $a^0(\zeta) \neq a(\zeta)$. The intuition is that such a treatment does not eliminate all of the influence of latent demand – the decision-maker views both signals as informative and weighs them against one another, so the posterior belief lies between 0 and $E[h|h^L]$. However, because signaling $h^T = 0$ moves actions toward the natural action it can be informative about the *direction* of latent demand. In contrast, in an alternative formulation with non-independent signals, where participants perceive the demand treatments to contain the same information as latent demand but less noise, signaling $h^T = 0$ does elicit the natural action. Thus, demanding the natural action does not necessarily obtain bounds that contain the natural action, while a pair of sufficiently informative positive and negative demand treatments does.

Derivation of beliefs with ternary signals

Recall that now $h \in \{-1, 0, 1\}$, $h^L \in \{-1, 0, 1\}$ and $h^T \in \{-1, 0, 1, \emptyset\}$.

To avoid clutter we suppress dependence on ζ. After observing h^L, the participant forms a posterior $E[h|h^L] = Pr(h = 1|h^L) \times 1 + Pr(h = -1|h^L) \times (-1)$. We can write this as:

$$E[h|h^L = y] = Pr(h = 1|h^L = y) - Pr(h = -1|h^L = y) = \frac{A}{B}$$
$$A = Pr(h^L = y|h = 1)Pr(h = 1) - Pr(h^L = y|h = -1)Pr(h = -1)$$
$$B = Pr(h^L = y|h = 1)Pr(h = 1) + Pr(h^L = y|h = 0)Pr(h = 0)$$
$$+ Pr(h^L = y|h = -1)Pr(h = -1)$$

Since $Pr(h = j|h^L = y) = \frac{1}{3}(1 - p^L) + p^L\mathbb{I}[y = j]$ and $Pr(h = j) = \frac{1}{3}$ we have

$$A = \frac{1}{3}\left[\left(\frac{1}{3}(1 - p^L) + p^L\mathbb{I}[y = 1]\right) - \left(\frac{1}{3}(1 - p^L) + p^L\mathbb{I}[y = -1]\right)\right]$$
$$= \frac{1}{3}p^L\left[\mathbb{I}[y = 1] - \mathbb{I}[y = -1]\right] = \frac{1}{3}p^L h^L$$
$$B = \frac{1}{3}\left[\left(\frac{1}{3}(1 - p^L) + p^L\mathbb{I}[y = 1]\right) + \left(\frac{1}{3}(1 - p^L) + p^L\mathbb{I}[y = 0]\right)\right.$$
$$\left. + \left(\frac{1}{3}(1 - p^L) + p^L\mathbb{I}[y = -1]\right)\right] = \frac{1}{3}$$

So

$$E[h|h^L = y] = p^L h^L \tag{B4}$$

just as before. Turning to beliefs following the demand treatments, as before we assume that when $h^T = \emptyset$, $E[h|h^T, h^L] = E[h|h^L]$. We have:

$$E[h|h^T, h^L] = Pr(h = 1|h^T, h^L) - Pr(h = -1|h^T, h^L) = \frac{A}{B}$$
$$A = Pr(h^T = x|h = 1, h^L = y)Pr(h = 1|h^L = y)$$
$$- Pr(h^T = x|h = -1, h^L = y)Pr(h = -1|h^L = y)$$
$$B = Pr(h^T = x|h = 1, h^L = y)Pr(h = 1|h^L = y)$$
$$+ Pr(h^T = x|h = 0, h^L = y)Pr(h = 0|h^L = y)$$
$$+ Pr(h^T = x|h = -1, h^L = y)Pr(h = -1|h^L = y, h^L = y)$$

Using

$$Pr(h^T = x|h = j, h^L = y) = \frac{1}{3}(1 - p^T) + p^T \mathbb{I}[x = j]$$
$$Pr(h = j|h^L = y) = \frac{1}{3}(1 - p^L) + p^L \mathbb{I}[y = j]$$

we obtain:

$$\begin{aligned}
A = &\left(\frac{1}{3}(1 - p^T) + p^T \mathbb{I}[x = 1]\right)\left(\frac{1}{3}(1 - p^L) + p^L \mathbb{I}[y = 1]\right) \\
&- \left(\frac{1}{3}(1 - p^T) + p^T \mathbb{I}[x = -1]\right)\left(\frac{1}{3}(1 - p^L) + p^L \mathbb{I}[y = -1]\right) \\
&= \frac{1}{3}(1 - p^T)p^L \mathbb{I}[y = 1] + \frac{1}{3}(1 - p^L)p^T \mathbb{I}[x = 1] + p^T p^L \mathbb{I}[x = 1]\mathbb{I}[y = 1] \\
&- \frac{1}{3}(1 - p^T)p^L \mathbb{I}[y = -1] - \frac{1}{3}(1 - p^L)p^T \mathbb{I}[x = -1] - p^T p^L \mathbb{I}[x = -1]\mathbb{I}[y = -1] \\
&= \frac{1}{3}(1 - p^T)p^L h^L + \frac{1}{3}(1 - p^L)p^T h^T + p^T p^L h^T \mathbb{I}[h^T = h^L]
\end{aligned}$$

$$\begin{aligned}
B = &\left(\frac{1}{3}(1 - p^T) + p^T \mathbb{I}[x = 1]\right)\left(\frac{1}{3}(1 - p^L) + p^L \mathbb{I}[y = 1]\right) \\
&+ \left(\frac{1}{3}(1 - p^T) + p^T \mathbb{I}[x = 0]\right)\left(\frac{1}{3}(1 - p^L) + p^L \mathbb{I}[y = 0]\right) \\
&+ \left(\frac{1}{3}(1 - p^T) + p^T \mathbb{I}[x = -1]\right)\left(\frac{1}{3}(1 - p^L) + p^L \mathbb{I}[y = -1]\right) \\
&= \frac{1}{3}(1 - p^T)(1 - p^L) + \frac{1}{3}p^T(1 - p^L)\left(\mathbb{I}[x = 1] + \mathbb{I}[x = 0] + \mathbb{I}[x = -1]\right) \\
&+ \frac{1}{3}p^L(1 - p^T)\left(\mathbb{I}[y = 1] + \mathbb{I}[y = 0] + \mathbb{I}[y = -1]\right) \\
&+ p^T p^L\left(\mathbb{I}[x = 1]\mathbb{I}[y = 1] + \mathbb{I}[x = 0]\mathbb{I}[y = 0] + \mathbb{I}[x = -1]\mathbb{I}[y = -1]\right) \\
&= \frac{1}{3}\left(1 - p^T p^L\right) + p^T p^L \mathbb{I}[h^T = h^L]
\end{aligned}$$

So

$$\text{(B5)} \qquad E[h|h^T, h^L] = \frac{\frac{1}{3}(1 - p^T)p^L h^L + \frac{1}{3}(1 - p^L)p^T h^T + p^T p^L h^T \mathbb{I}[h^T = h^L]}{\frac{1}{3}\left(1 - p^T p^L\right) + p^T p^L \mathbb{I}[h^T = h^L]}$$

B7. Computing confidence intervals

Here we describe how we compute demand-robust confidence intervals. We note that this was not included in the pre-analysis plans.

Confidence intervals for actions

Imbens and Manski (2004) show that asymptotically the probability that the estimate for the upper (lower) bound is lower (higher) than the true value can be ignored when making inference. Thus, one can construct one-sided intervals with confidence level α around both the upper and the lower bound. The 95 percent confidence interval for the true demand-free behavior is thus given by:

$$CI() = [a^-(\zeta) - \overline{C_N}\frac{\widehat{\sigma^-}}{\sqrt{N}},\, a^+(\zeta) + \overline{C_N}\frac{\widehat{\sigma^+}}{\sqrt{N}}]$$

Here, $\widehat{\sigma^-} = \sqrt{\widehat{Var(a^-(\zeta))}}$ and $\widehat{\sigma^+} = \sqrt{\widehat{Var(a^+(\zeta))}}$, and $\overline{C_N}$ satisfies

$$\Phi\left(\overline{C_N} + \sqrt{N}\frac{a^+(\zeta) - a^-(\zeta)}{\max(\widehat{\sigma^-},\, \widehat{\sigma^+})}\right) - \Phi(-\overline{C_N}) = 0.90.$$

The 95 percent confidence interval for the set $[a^-(\zeta), a^+(\zeta)]$ is given by:

$$CI() = [a^-(\zeta) - \overline{C_N}\frac{\widehat{\sigma^-}}{\sqrt{N}},\, a^+(\zeta) + \overline{C_N}\frac{\widehat{\sigma^+}}{\sqrt{N}}],$$

where $\overline{C_N}$ satisfies

$$\Phi\left(\overline{C_N} + \sqrt{N}\frac{a^+(\zeta) - a^-(\zeta)}{\max(\widehat{\sigma^-},\, \widehat{\sigma^+})}\right) - \Phi(-\overline{C_N}) = 0.95.$$

Confidence intervals for treatment effects

We also outline how one can compute confidence intervals for the treatment effects $[a(\zeta_1) - a(\zeta_0)]$ and for the set defined by the upper and lower bounds for treatment effects as given by our demand treatments: $[a(\zeta_1) - a(\zeta_0)] \in [a^-(\zeta_1) - a^+(\zeta_0), a^+(\zeta_1) - a^-(\zeta_0)]$

For simplicity we denote the lower bound, $[a^-(\zeta_1) - a^+(\zeta_0)]$, as T^- and the upper bound, $[a^+(\zeta_1) - a^-(\zeta_0)]$, as T^+. The 95 percent confidence interval for the true demand-free treatment effect is given by:

$$CI() = [T^- - \overline{C_N}\frac{\widehat{\sigma^{T-}}}{\sqrt{N}},\ T^+ + \overline{C_N}\frac{\widehat{\sigma^{T+}}}{\sqrt{N}}].$$

Here, $\widehat{\sigma^{T-}} = \sqrt{\widehat{Var(T^-)}}$ and $\widehat{\sigma^{T+}} = \sqrt{\widehat{Var(T^+)}}$, and $\overline{C_N}$ satisfies

$$\Phi\left(\overline{C_N} + \sqrt{N}\frac{T^+ - T^-}{\max(\widehat{\sigma^{T-}},\ \widehat{\sigma^{T+}})}\right) - \Phi(-\overline{C_N}) = 0.90.$$

The 95 percent confidence interval for the set $[a^-(\zeta_1) - a^+(\zeta_0), a^+(\zeta_1) - a^-(\zeta_0)]$ is as follows:

$$CI() = [T^- - \overline{C_N}\frac{\widehat{\sigma^{T-}}}{\sqrt{N}},\ T^+ + \overline{C_N}\frac{\widehat{\sigma^{T+}}}{\sqrt{N}}],$$

where

$$\Phi\left(\overline{C_N} + \sqrt{N}\frac{T^+(\tau) - T^-(\tau)}{\max(\widehat{\sigma^{T-}},\ \widehat{\sigma^{T+}}}\right) - \Phi(-\overline{C_N}) = 0.95.$$

B8. Controlling for demand

Here we provide derivations for the results in section III.D. We begin with the usual first-order condition, assuming a demand treatment h^T:

$$0 = v_1(a^*(\zeta, h^T), \zeta) + \phi(\zeta)E[h|h^T, h^L(\zeta)]$$

taking the first-order Taylor approximation at the natural action $a(\zeta)$ we obtain:

$$0 \approx \underbrace{v_1(a(\zeta), \zeta)}_{=0} + \phi(\zeta)E[h|h^T, h^L(\zeta)] + [a^*(\zeta, h^T) - a(\zeta)]v_{11}(a(\zeta), \zeta)$$

where the first term is zero by definition of $a(\zeta)$. Rearranging we obtain equation (8):

$$\begin{aligned} a^*(\zeta, h^T) &\approx a(\zeta) - \frac{\phi(\zeta)}{v_{11}(a(\zeta), \zeta)} E[h|h^T, h^L(\zeta)]. \\ &= a(\zeta) + \Phi(\zeta) E[h|h^T, h^L(\zeta)]. \end{aligned}$$

Assume two treatment groups: $\zeta \in \{0, 1\}$, and denote their corresponding demand treatments by h_ζ^T. If no demand treatments are applied ($h_1^T = h_0^T = \emptyset$). Since we are interested in cases where $h_0^T = h_1^T$ we suppress the subscripts. The approximate bias of the treatment effect estimate can be written as:

$$\begin{aligned} Bias &= a^*(1, \emptyset) - a^*(0, \emptyset) - [a(1) - a(0)] \\ &\approx a(1) + \Phi(1) E[h|h^T, h^L(1)] - a(0) - \Phi(0) E[h|h^T, h^L(0)] - [a(1) - a(0)] \end{aligned}$$

Adding and subtracting terms yields:

$$Bias \approx \underbrace{\Phi(1) \left(E[h|h^T, h^L(1)] - E[h|h^T, h^L(0)] \right)}_{\text{Bias due to beliefs}} + \underbrace{\left(\Phi(1) - \Phi(0) \right) E[h|h^T, h^L(0)]}_{\text{Bias due to "responsiveness"}}$$

Fully informative demand treatments

We now show that when demand treatments are fully informative, one can test for bias due to behavioral responsiveness.

$$\begin{aligned} &\underbrace{[a^*(1,1) - a^*(1,0)]}_{\text{Sensitivity } (\zeta = 1)} - \underbrace{[a^*(0,1) - a^*(0,0)]}_{\text{Sensitivity } (\zeta = 0)} \\ &= \underbrace{[a^*(1,1) - a^*(0,1)]}_{\text{Treatment effect } (h^T = 1)} - \underbrace{[a^*(1,0) - a^*(0,0)]}_{\text{Treatment effect } (h^T = -1)} \\ &\approx \left[a(1) + \Phi(1) E[h|1, h^L(1)] - a(0) - \Phi(0) E[h|1, h^L(0)] \right] \\ &- \left[a(1) + \Phi(1) E[h|-1, h^L(1)] - a(0) - \Phi(0) E[h|-1, h^L(0)] \right] \\ &= [\Phi(1) \times 1 - \Phi(0) \times 1] - [\Phi(1) \times -1 - \Phi(0) \times -1] \\ &= 2 \left(\Phi(1) - \Phi(0) \right) \end{aligned}$$

Next, we show that averaging the "positive-positive" and "negative-negative"

treatment effects approximates the true treatment effect

$$\begin{aligned}
&\frac{1}{2}\left([a^*(1,1)-a^*(0,1)]+[a^*(1,-1)-a^*(0,-1)]\right)\\
\approx& a(1)-a(0)+\frac{1}{2}\left[(\Phi(1)-\Phi(0))\times 1+(\Phi(1)-\Phi(0))\times -1\right]\\
=& a(1)-a(0)
\end{aligned}$$

Less informative treatments

For compactness we define notation $H^L(\zeta) \equiv h^L(\zeta)p^L(\zeta)$ and $H^T \equiv h^T p^T$. Our Bounding assumption implies $\left|H^T\right| \geq \left|H^L(\zeta)\right|$.

Consider the expressions for belief differences between treatment and control, first without (Diff^L) and then with (Diff^T) demand treatments. We have

$$\text{Diff}^L \equiv H^L(1) - H^L(0)$$

and:

$$\begin{aligned}
\text{Diff}^T &\equiv \frac{H^L(1)+H^T}{1+H^L(1)H^T} - \frac{H^L(0)+H^T}{1+H^L(0)H^T}\\
&= \frac{(1-H^{T2})}{(1+H^L(1)H^T)(1+H^L(0)H^T)} \times \text{Diff}^L
\end{aligned}$$

We want to find conditions under which $\left|\text{Diff}^T\right| < \left|\text{Diff}^L\right|$, which holds if and only if

$$\frac{(1-H^{T2})}{(1+H^L(1)H^T)(1+H^L(0)H^T)} < 1$$

rearranging we obtain:

$$0 < H^T(H^L(1)+H^L(0)) + H^{T2}(1+H^L(1)H^L(0)) \tag{B6}$$

If $h^T = 1$, (B6) reduces to

$$-\frac{H^L(1)+H^L(0)}{1+H^L(1)H^L(0)} < H^T$$

while if $h^T = -1$ it reduces to

$$-\frac{H^L(1) + H^L(0)}{1 + H^L(1)H^L(0)} > H^T.$$

Since the left hand side lies in the interval $(-1, 1)$ there always exists a sufficiently strong demand treatment (p^T sufficiently large) that (B6) is satisfied. We now evaluate whether there is more we can say. There are X cases to consider. Assume throughout, without loss of generality, that $\left|H^L(1)\right| > \left|H^L(0)\right|$.

1) Suppose the latent demand beliefs have the same sign as each other ($h^L(1) = h^L(0) = h^L$) and the same sign as the demand treatment ($h^L = h^T$). Then it is easy to verify that (B6) holds for all H^T. Intuitively, when the latent demand beliefs have the same sign, additional information that further reinforces those beliefs has a greater effect on the one that is less certain, reducing the gap between them.

2) The latent demand beliefs have the same sign as each other ($h^L(1) = h^L(0) = h^L$) and the opposite sign to the demand treatment ($h^L = -h^T$). Assume $h^T = 1$ and $h^L = -1$ (the opposite case is symmetric). We know that (B6) holds for sufficiently strong H^T, we will ask if our Bounding assumption is sufficient. We show that it is not, by contradiction. Suppose Bounding holds exactly, i.e. $H^T = -H^L(1)$. Then, by the premise that (B6) is satisfied:

$$\begin{aligned} -\frac{H^L(1) + H^L(0)}{1 + H^L(1)H^L(0)} &< -H^L(1) \\ \frac{1 + \frac{H^L(0)}{H^L(1)}}{1 + H^L(1)H^L(0)} &< 1 \end{aligned}$$

which holds if and only if:

$$\begin{aligned} \frac{H^L(0)}{H^L(1)} &< H^L(1)H^L(0) \\ 1 &< H^L(1)^2 \end{aligned}$$

a contradiction since $H^L(1) < 1$. Thus in this case the condition for demand treatments to reduce bias is stronger than Bounding.

3) The latent demand beliefs have opposite signs ($h^L(1) = -h^L(0)$), and the stronger belief ($H^L(1)$) has the same sign as h^T, i.e. $h^T = h^L(1)$. Focus again on the case where $h^T = 1$ (the opposite case is symmetric). We require:

$$-\frac{H^L(1) + H^L(0)}{1 + H^L(1)H^L(0)} < H^T$$

It is easy to see that the condition is always satisfied since the left-hand side is negative.

4) The latent demand beliefs have opposite signs ($h^L(1) = -h^L(0)$), and the stronger belief has the opposite sign to h^T, i.e. $h^T = -h^L(1)$. Focus again on the case where $h^T = 1$ (the opposite case is symmetric). We know that (B6) holds for sufficiently strong H^T, we will ask if our Bounding assumption is sufficient. Thus let $H^T = -H^L(1)$ (bounding holds exactly). We require:

$$\begin{aligned}
-\frac{H^L(1) + H^L(0)}{1 + H^L(1)H^L(0)} &< -H^L(1) \\
\frac{1 + \frac{H^L(0)}{H^L(1)}}{1 + H^L(1)H^L(0)} &< 1 \\
\frac{H^L(0)}{H^L(1)} &< H^L(1)H^L(0) \\
1 &> H^L(1)^2
\end{aligned}$$

which is satisfied. Thus Bounding is sufficient for (B6) to hold.

B9. Structural estimation

This section outlines step by step how the parameters are constructed in our NLLS estimation of the structural model in section III.E.

Data and parameter adjustments

First, we follow DP exactly in rounding effort scores to the nearest 100 (except for those in range $[1, 49]$ which we round to 25). This is because incentives were paid per 100 points, and we wish to avoid modeling effort choices that lie between two 100 point thresholds. We refer the reader to DP for further details.

Second, we make a couple of adjustments pre and post-estimation. First, we divide the rounded scores by 100. In other words, if effort a is measured in points,

we compute $a' = a/100$ which is measured in hundreds of points. Second, we multiply the incentive, ζ, which is measured in cents per point, by 100 to express it as $\zeta' = 100\zeta$ which is measured in cents per 100 points. These transformations were helpful in achieving convergence of the estimator, which otherwise occasionally suffered from underflow problems. However they change the interpretation of the parameters. Specifically, the intrinsic motivation parameter s and the preference for pleasing the experimenter, ϕ, will both be measured in units equivalent to cents per 100 points, while the cost function parameters will be expressed for effort measured in hundreds of points.

To aid comparability with DP we therefore re-transform the parameters after estimation. DP present their estimates of incentive parameters (which in our case are s and ϕ) in the same units, cents per 100 points, so we do not need to correct them. k and γ are reported for effort measured in points, so we transform our estimates for comparability. We derive the adjustments as follows. First, for the power cost function, we have:

$$U = (s + \zeta + \phi E[h|h^T, h^L])a - \frac{ka^{1+\gamma}}{1+\gamma}$$

Let $a' = \frac{a}{100}$ and $\zeta' = 100\zeta$. Then:

$$\begin{aligned} U &= \left(s + \frac{\zeta'}{100} + \phi E[h|h^T, h^L]\right) 100a' - \frac{k(100a')^{1+\gamma}}{1+\gamma} \\ &= \left(100s + \zeta' + 100\phi E[h|h^T, h^L]\right) a' - \frac{k(100a')^{1+\gamma}}{1+\gamma} \end{aligned}$$

giving rise to first-order condition:

$$\begin{aligned} 0 &= \left(100s + \zeta' + 100\phi E[h|h^T, h^L]\right) - ka'^{\gamma} 100^{1+\gamma} \\ a' &= \left(\frac{100s + \zeta' + 100\phi E[h|h^T, h^L]}{k100^{1+\gamma}}\right)^{\frac{1}{\gamma}} \\ \log(a') &= \frac{1}{\gamma} \log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{k^*}\right) \end{aligned}$$

where $s^* = 100s$, $\phi^* = 100\phi$ and $k^* = 100^{1+\gamma}k$. We leave s^* and ϕ^*, (which are in equivalent units to cents per 100 points) untransformed for comparability with DP. In the tables we report $k = k^*/100^{1+\gamma}$ and its standard error, computed via

the delta method.

For the exponential cost function we have:

$$\begin{aligned} U &= (s + \zeta + \phi E[h|h^T, h^L])a - \frac{k}{\gamma}\exp(\gamma a) \\ &= (s^* + \zeta' + \phi^* E[h|h^T, h^L])a' - \frac{k}{\gamma}\exp(100\gamma a') \end{aligned}$$

implying first-order condition:

$$\begin{aligned} 0 &= s^* + \zeta' + \phi^* E[h|h^T, h^L] - 100k\exp(100\gamma a') \\ a' &= \frac{1}{100\gamma}\log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{100k}\right) \\ &= \frac{1}{\gamma^*}\log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{k^*}\right) \end{aligned}$$

where $s^* = 100s$, and $\phi^* = 100\phi$ as before, while $\gamma^* = 100\gamma$, $k^* = 100k$. In the tables we report $\gamma = \gamma^*/100$ and $k = k^*/100$.

Error term

To allow for the observed heterogeneity in effort, we follow DP in assuming heterogeneous effort costs, as follows. Let the cost of effort under power utility equal $ka^{1+\gamma}(1+\gamma)^{-1}\exp(-\gamma\epsilon)$ where $\epsilon \sim N(0, \sigma_\epsilon^2)$. Then our FOC becomes

$$\begin{aligned} 0 &= \left(100s + \zeta' + 100\phi E[h|h^T, h^L]\right) - ka'^{\gamma}100^{1+\gamma}\exp(-\gamma\epsilon) \\ a' &= \left(\frac{100s + \zeta' + 100\phi E[h|h^T, h^L]}{k100^{1+\gamma}}\right)^{\frac{1}{\gamma}}\exp(\epsilon) \\ \log(a') &= \frac{1}{\gamma}\log\left(\frac{100s + \zeta' + 100\phi E[h|h^T, h^L]}{k100^{1+\gamma}}\right) + \epsilon \end{aligned}$$

where ϵ becomes the error term in our NLLS routine. For the exponential cost, we follow DP and assume effort cost is $k\gamma^{-1}\exp(\gamma a)\exp(-\gamma\epsilon)$. Then our FOC

becomes

$$
\begin{aligned}
0 &= s^* + \zeta' + \phi^* E[h|h^T, h^L] - 100k \exp(100\gamma a') \exp(-\gamma\epsilon) \\
a' &= \frac{1}{100\gamma} \log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{100k}\right) + \frac{\epsilon}{100} \\
&= \frac{1}{\gamma^*} \log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{k^*}\right) + \epsilon^*
\end{aligned}
$$

where $\epsilon^* = \epsilon/100$ forms the error term in our estimation.

Estimating equation

Finally, in our estimation we sometimes need to estimate the product $\phi^* E[h|h^L]$. We estimate this product directly, then transform by dividing by ϕ^*. Specifically, we estimate the following:

$$
\begin{aligned}
y_i = \frac{1}{\beta_0} \log \big[& \zeta_i' + \beta_1 + \beta_2(\text{pos\_demand}_i - \text{neg\_demand}_i) \\
& + \beta_3 \times \text{no\_demand}_i \times \text{incentive\_0c}_i + \beta_4 \times \text{no\_demand}_i \times \text{incentive\_1c}_i \\
& + \beta_5 \times \text{no\_demand}_i \times \text{incentive\_4c}_i \big] - \frac{1}{\beta_0} \log(\beta_6) + \varepsilon_i
\end{aligned}
$$

where $y = \log(a')$ or a' respectively, pos_demand, neg_demand and no_demand are dummies for our positive, negative and no demand treatments, while incentive_Xc is a dummy for the treatment with X cents per 100 points. Parameters are as follows: $\beta_0 = \gamma$ or γ^* respectively, $\beta_1 = s^*$, $\beta_2 = \phi^*$, $\beta_3 = \phi^* E[h|h^L(\zeta = 0)]$, $\beta_4 = \phi^* E[h|h^L(\zeta = 1)]$, $\beta_5 = \phi^* E[h|h^L(\zeta = 4)]$ and $\beta_6 = k^*$. We then compute the three values for $E[h|h^L]$ by dividing by β_2, i.e. β_3/β_2, β_4/β_2 and β_5/β_2. γ and k are computed by the transformations outlined above. Standard errors are computed by the delta method. In the specification where we restrict latent demand to be equal for the 1 cent and 4 cent treatments we impose $\beta_4 = \beta_5$.

Extrapolation

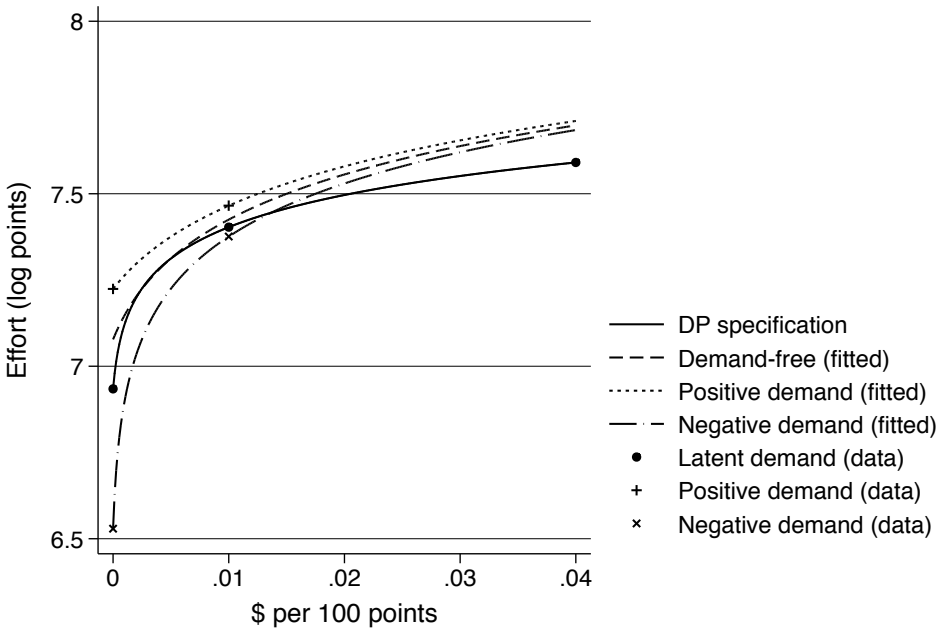
Our large estimates of $h^L(4)p^L(4)$ reflect an out-of-sample extrapolation as the main model parameters are estimated from the 0 and 1 cent treatment groups. Figure B1 illustrates this for the power cost case. Points correspond to mean (log) effort for each treatment group. The figure then plots (a) predicted effort using

the DP specification, which fits the "no demand" data only (parameters taken from Column 1 of table 4), and (b) predicted effort using the exactly identified model (Column 3 of table 4), for each case of zero demand $E[h] = 0$, strong positive demand ($E[h] = 1$) and strong negative demand ($E[h] = -1$). The estimation then recovers the latent demand beliefs by comparing observed effort to predicted effort when demand is zero.

It is clear from the figure that the extrapolation from model (b) to the 4 cent effort treatment is not perfect, and the observed behavior lies outside the limits implied by $E[h] \in [-1, 1]$. This is the reason for the large negative fitted value for beliefs at this point.

Another thing that is clear from the figure is how the curvature of the effort cost function determines the imputed latent beliefs, which may explain the difference in imputed beliefs between the power and exponential cost functions. The sign of imputed beliefs depends on whether the "no demand" point lies above or below the curve, so changes in curvature can flip the sign of these estimates.

Figure B1. Structural estimation: fitted values



Note: Figure displays mean (log) effort for each treatment used in the structural estimation, and fitted values from the estimated models.

B10. Using the method in practice

In this section we work through an example to illustrate the methods we have developed. We do this with the help of Figure B2, which represents the 10 data points available to a researcher who has applied no demand, weak and strong treatments to a control group (ζ_0) and a treatment group (ζ_1). To avoid confusion, we will label the actions under weak positive and negative demand as $a^+(\zeta)$ and $a^-(\zeta)$, while strong (assumed to be fully informative) positive and negative demand actions are defined as $a^{++}(\zeta)$ and $a^{--}(\zeta)$. In our example the conventional treatment effect is given by $a^L(\zeta_1) - a^L(\zeta_0)$ (or point H minus point C).

1) We can use our strong treatments to construct bounds for actions: In Figure B2 the bounds on action one are defined by points A and E for the control group, $[a^{--}(\zeta_0), a^{++}(\zeta_0)]$, and by points F and J for the treatment group, $[a^{--}(\zeta_1), a^{++}(\zeta_1)]$.

2) Similarly, we can construct bounds using the weak treatments, which are defined by points B and D for the control group, $[a^-(\zeta_0), a^+(\zeta_0)]$, and by points G and I for the treatment group, $[a^-(\zeta_1), a^+(\zeta_1)]$. Under the assumption that our demand treatments are more informative than underlying latent demand, these bounds contain the natural action.

3) We can analogously also define strong bounds for treatment effects. The upper bound is given by the comparison between respondents in the treatment group that receive strong positive demand treatments, and respondents in the control group that receive strong negative demand treatments. In Figure B2 this corresponds to the difference between points J and A: $a^{++}(\zeta_1) - a^{--}(\zeta_0)$. The lower bound is given by the comparison between respondents in the treatment group that receive strong negative demand treatments, and respondents in the control group that receive strong positive demand treatments, given by the difference points F and E: $[a^{--}(\zeta_1) - a^{++}(\zeta_0)$. The bounds are formally defined as follows: $[a^{--}(\zeta_1) - a^{++}(\zeta_0), a^{++}(\zeta_1) - a^{--}(\zeta_0)]$.

4) Similarly, we can construct weak bounds for treatment effects, by applying weak instead of strong treatments. The bounds are given by: $[a^-(\zeta_1) -$

$a^+(\zeta_0), a^+(\zeta_1) - a^-(\zeta_0)]$. In Figure B2 the upper and lower bounds are given by the difference between points I and B, and points G and D, respectively.

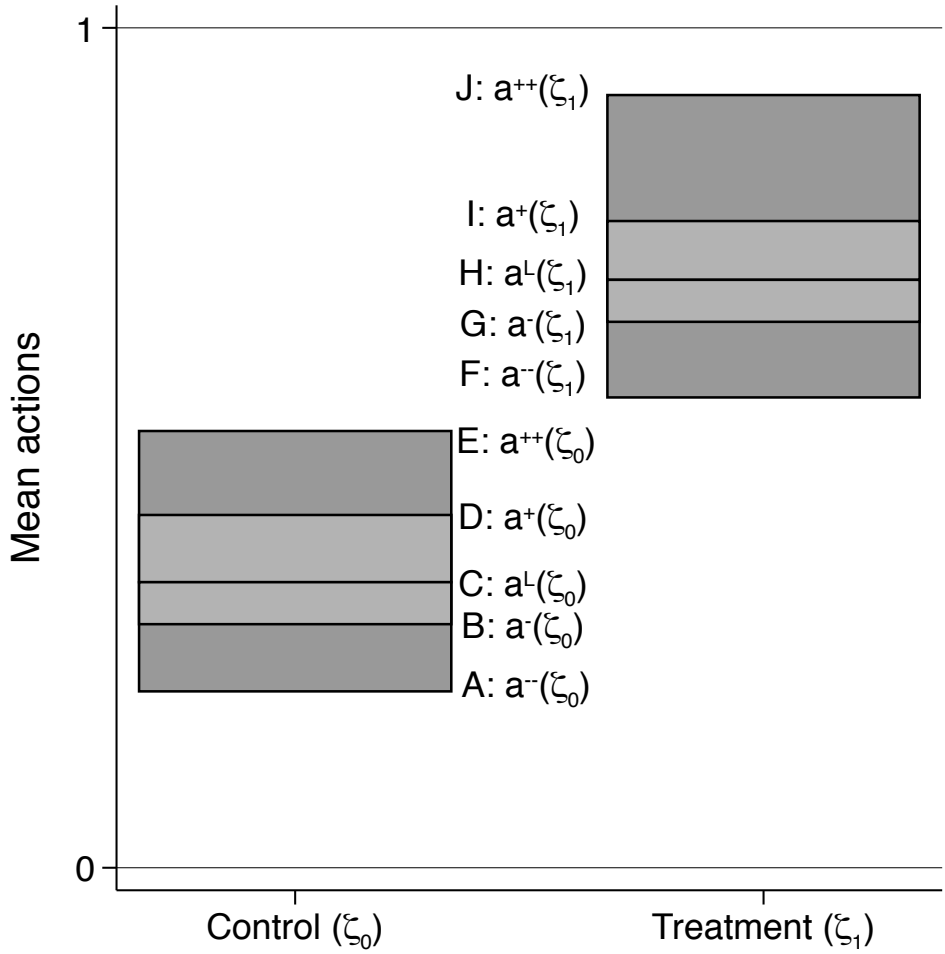
5) An alternative to creating bounds, is to "control for demand effects". Under the assumption that demand treatments are fully informative, provided responsiveness to demand treatments does not differ across treatment groups, we can point identify treatment effects that are not biased by demand effects. In this specific case, we could apply strong positive or strong negative demand treatments to both the treatment and the control group: In Figure B2 the estimates are given by the difference between points J and E: $(a^{++}(\zeta_1) - a^{++}(\zeta_0))$, or F and A: $(a^{--}(\zeta_1) - a^{--}(\zeta_0))$.

6) If responsiveness to fully informative demand treatments differs significantly across treatment arms, our point estimates from employing same-signed demand treatments are still biased. However, by the symmetry of the Taylor approximation, we can approximate the treatment effect using the midpoints of the bounds generated the strong demand treatments. In B2 this corresponds to comparing the average of A and E to the average of J and F: $0.5 * [(a^{++}(\zeta_1) + a^{--}(\zeta_1)) - (a^{++}(\zeta_0) + a^{--}(\zeta_0))]$.

7) Our approach of "controlling for demand effects" can also be extended to weak treatments. In Section 3.4 we outline the conditions under which this approach reduces bias. First, we compare respondents in the treatment and control group who all receive weak positive or weak negative demand treatments. In Figure B2 the positive-positive point estimate is defined by points I and D: $(a^+(\zeta_1) - a^+(\zeta_0))$, while the negative-negative estimate is comes from points F and B: $(a^-(\zeta_1) - a^-(\zeta_0))$.

8) Finally, fully informative demand treatments can be used to eliminate nuisance parameters due to unobservable beliefs, facilitating the estimation of structural models. Structural estimation leverages points A, E, F, and J to estimate model parameters, and uses those to impute the latent demand beliefs at points C and H.

FIGURE B2. USING THE METHOD IN PRACTICE: EXAMPLE



Note: Figure displays mean actions under different treatment conditions and different demand treatments. Point A is given by respondents in the control group who receive the negative strong demand treatment: $a^{--}(\zeta_0)$; Point B is given by respondents in the control group who receive the negative weak demand treatment: $a^{-}(\zeta_0)$; Point C is defined by respondents in the control group who receive no demand treatment: $a^{L}(\zeta_0)$; Point D is given by respondents in the control group who receive the positive weak demand treatment: $a^{+}(\zeta_0)$; Point E is given by respondents in the control group who receive the positive strong demand treatment: $a^{++}(\zeta_0)$. Points F to J are defined analogously for respondents in the treatment group (ζ_1).

C. Pre-specified Tables and Figures

This section works through the pre-specified analysis for each experiment, presenting summaries of the raw data and conducting hypothesis tests.

1) Pre-analysis plan 1 described experiment 1, which was conducted on MTurk with strong demand treatments and both real and hypothetical stakes, on the dictator game, investment game and convex time budget.

2) Pre-analysis plan 2 described experiment 2, which was conducted on MTurk with weak demand treatments and both real and hypothetical stakes, on the dictator game and investment game.

3) Pre-analysis plan 3 described experiment 3, which was conducted on MTurk with strong demand treatments, real stakes and the real-effort task.

4) Pre-analysis plan 4 described experiment 4, which was conducted on the representative panel with both strong and weak demand treatments, real stakes, and the dictator game and investment game.

5) Pre-analysis plan 5 described experiments 5 and 6, which were conducted on MTurk with strong and weak demand treatments and collected data for the remaining games (experiment 6 collected real-effort data and experiment 5 collected the other games).

6) Pre-analysis plan 6 described experiment 7, which were conducted on MTurk with strong demand treatments, varied within-participant, on the dictator game and investment game.

The majority of the hypothesis tests for each pre-analysis plan are presented in a single table format (e.g. Table C3). The top half of these tables report regression coefficients and standard errors, and the bottom half reports p-values (and adjusted p-values) on the pre-specified hypothesis tests.

When conducting multiple tests within a family of hypotheses we also report false-discovery rate corrected p-values. These are used when a) testing for a positive effect (on actions or beliefs) of the positive demand treatment, negative effect of the negative demand treatment and overall effect; and b) when testing for heterogeneity across games within an experiment.

We deviate from the pre-analysis plans in two minor ways, which are inconsequential for the results.

1) As described in section II.A of the paper, we only pre-specified sample exclusions in the main real-effort experiment 3 (to match those used by DellaVigna and Pope). For consistency, we decided to apply the same restrictions to all other games. The only binding restriction was the dropping of participants who submitted multiple responses in a given experiment, amounting to less than 0.5 percent of our sample.

2) In experiments for which we collected data without demand treatments ("no demand" conditions), we pre-specified to standardize actions by the mean and standard deviation of this group. However, experiments 5 and 6 only collected positive and negative demand conditions. For consistency, therefore, we instead always standardize by the mean and standard deviation of the negative demand treatment group. This amounts to a simple linear transformation of the data.

In addition, some of the analysis in the paper was not described in the pre-analysis plans: the bounding of treatment effects, the computation of confidence intervals on the bounds, and the structural analysis.

C1. Pre-analysis Plan 1

- Table C1 and Figure C1 summarize the means, standard errors, and corresponding 95 percent confidence intervals from experiment 1 across all 18 treatment arms. Table C2 displays the game-level regressions based on the raw data showing the control mean from the "no demand condition" as well as the coefficients on the positive demand treatment indicator and the negative demand treatment indicator.

- Balance tests for this experiment are in Table D1 in Section D and indicate that there are no imbalances.

- Table C3 displays the main effects of the positive and negative demand treatment as well as heterogeneous treatment effects by gender, attention and whether choices are hypothetical or incentivized. This table also summarizes the results for the tests we had pre-specified in the pre-analysis plan.

  – Column 1 of Table C3 shows that people increase their actions in response to the positive demand treatments ($p<0.001$), decrease their

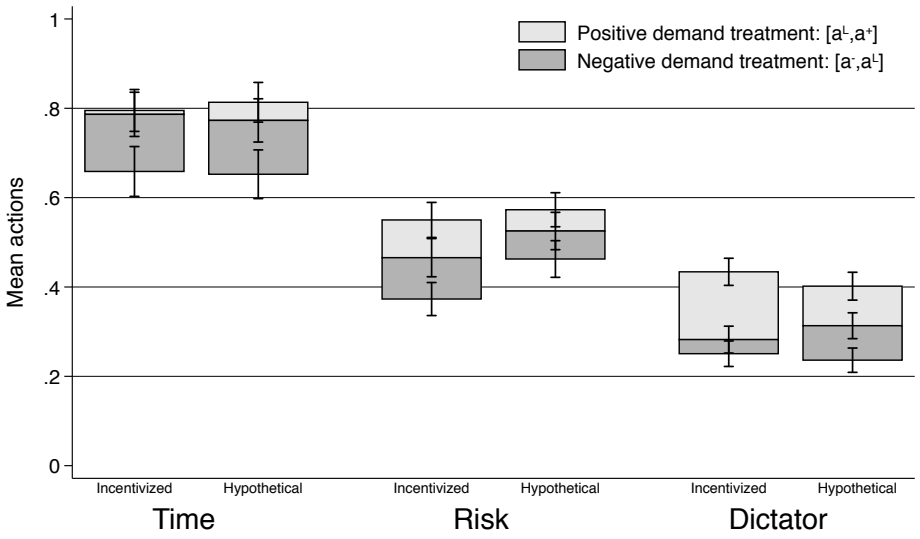
actions in response to the negative demand treatments ($p<0.001$) and that the overall response to demand is non-zero ($p<0.001$). False-discovery rate corrected p-values reach the same conclusion.

- Next, in column 2 of Table C3 we show that there is no significant treatment heterogeneity depending on whether choices are hypothetical or incentivized ($p=0.24$).
- In column 3, we test whether there are any systematic gender differences in response to demand. Pooling across all tasks, measured sensitivity was higher for women than for men ($p=0.099$).
- In column (4) we test whether attention moderates the response to demand treatments. We find stronger responses to the demand treatments for more attentive respondents ($p=0.102$).
- In column (5) we examine heterogeneity across games. We find that overall sensitivity in the dictator game is significantly higher than sensitivity in the time preference measure and the risk game ($p<0.01$). We find no significant difference in sensitivity in the time preference measure and the risk game ($p=0.552$). False-discovery rate corrected p-values reach the same conclusion. An omnibus test of differences across all games highlights that responses significantly differ between games ($p<0.001$).

- Table C4 explores how people's beliefs about whether the experimenter wanted (column 1) or expected (column 2) a high action. Table C4 shows the results for the tests we pre-specified.
  - People in the positive demand condition are more likely to think that the experimenter wanted a high action ($p<0.001$) and that the experimenter expected a high action ($p<0.001$).
  - People in the negative demand condition are less likely to think that the experimenter wanted a high action ($p<0.001$) and that the experimenter expected a high action ($p<0.001$).
  - Overall, people in the positive demand condition are significantly more likely to think that the experimenter wanted a high action or expected a high action compared to people in the negative demand condition ($p<0.001$).

– False-discovery rate corrected p-values reach the same conclusions.

Figure C1. Overview of raw data: Experiment 1



Note: This figure summarizes the mean actions and corresponding 95 confidence intervals from experiment 1 across all 18 treatment arms

Table C1—Overview of raw data: experiment 1

	Time		Risk Aversion		Dictator Game	
	Incentivized	Hypothetical	Incentivized	Hypothetical	Incentivized	Hypothetical
Unconditional Means						
Positive demand: Mean	0.795	0.813	0.550	0.573	0.434	0.402
Positive demand: SD	0.379	0.357	0.300	0.316	0.253	0.258
No demand: Mean	0.786	0.773	0.466	0.525	0.282	0.313
No demand: SD	0.386	0.392	0.340	0.335	0.246	0.230
Negative demand: Mean	0.659	0.652	0.373	0.463	0.251	0.236
Negative demand: SD	0.437	0.440	0.300	0.327	0.225	0.206
Observations	727	757	728	764	770	733

Note:

This table summarizes the raw actions from experiment 1 across all 18 treatment arms.

Table C2—Game-level regressions: experiment 1

	Time	Risk Aversion	Dictator Game
Positive demand	0.025	0.067	0.121
	(0.024)	(0.021)	(0.015)
Negative demand	-0.124	-0.079	-0.054
	(0.026)	(0.021)	(0.015)
Control Mean	0.779	0.496	0.297
Observations	1484	1492	1503

Note: This table shows the effect of the positive and negative demand treatment at the game level based on the raw actions (pooling across incentivized and unincentivized choices).

Table C3—Strong Demand (Experiment 1)

	(1)	(2)	(3)	(4)	(5)
Positive Demand	0.242	0.191	0.280	0.017	0.457
	(0.035)	(0.049)	(0.048)	(0.189)	(0.058)
Negative Demand	-0.247	-0.257	-0.269	-0.206	-0.203
	(0.036)	(0.051)	(0.049)	(0.175)	(0.055)
Positive demand × interactant		0.103	-0.072	0.236	
		(0.070)	(0.070)	(0.192)	
Negative demand × interactant		0.021	0.044	-0.044	
		(0.072)	(0.072)	(0.179)	
Interactant		-0.096	-0.046	-0.082	
		(0.051)	(0.051)	(0.141)	
Positive Demand × Risk					-0.258
					(0.085)
Negative Demand × Risk					-0.031
					(0.083)
Positive Demand × Time					-0.393
					(0.085)
Negative Demand × Time					-0.114
					(0.087)
Constant	-0.149	-0.101	-0.125	-0.070	-0.343
	(0.025)	(0.035)	(0.034)	(0.139)	(0.040)
Interactant		Monetary Incentive	Male	Passed attention check	
Adjusted R^2	0.041	0.041	0.041	0.041	0.052
Positive demand ≤ 0	0.000				
Adjusted p-value	0.010				
Negative demand ≥ 0	0.000				
Adjusted p-value	0.010				
Positive demand = negative demand	0.000				
Adjusted p-value	0.010				
(Positive demand - negative demand)* interaction = 0		0.240	0.099	0.102	
Risk*(pos - neg) = Time*(pos - neg)					0.552
Adjusted p-value					0.283
Risk*(positive demand - negative demand) = 0					0.006
Adjusted p-value					0.011
Time*(positive demand - negative demand) = 0					0.001
Adjusted p-value					0.006
Joint F-test					.001
Observations	4479	4479	4479	4479	4479

Note: This table summarizes the results from experiment 1. The outcome variable (action chosen) is standardized at the game level using the mean and standard deviation of the negative demand group. Robust standard errors are in parentheses. Lower section of the table reports p-values on pre-specified hypothesis tests.

Table C4—Beliefs about the experimental objective and hypothesis: Strong Demand

	Belief: Want High	Belief: Expect High
Positive - Negative	0.278	0.181
	(0.017)	(0.018)
Adjusted p-value	[0.001]	[0.001]
Positive - Neutral	0.161	0.143
	(0.017)	(0.018)
Adjusted p-value	[0.001]	[0.001]
Negative - Neutral	-0.116	-0.038
	(0.018)	(0.018)
Adjusted p-value	[0.001]	[0.006]
Mean (No Demand)	0.543	0.451
Observations	4479	4479

Note: The outcome variables take value one if the respondents believed that the experimenter wanted (column 1) or expected (column 2) a high action and zero if they wanted or expected a low action.

- Table C5 and Figure C2 summarize the means, standard errors, and corresponding 95 confidence intervals from experiment 2 across all 12 treatment arms. Table C6 displays the game-level regressions based on the raw data showing the control mean from the "no demand condition" as well as the coefficients on the positive demand treatment indicator and the negative demand treatment indicator.

- Balance tests for this experiment are in Table D2 in Section D. We find a slight imbalance for an indicator taking value 1 for those in part-time employment. Table C8 shows the main results controlling for this indicator.

- Table C7 displays the main effects of the positive and negative demand treatment as well as heterogeneous treatment effects by gender, attention and whether choices are hypothetical or incentivized. This table also summarizes the results of the tests we pre-specified.

  - Column (1) of Table C7 shows that people increase their actions in response to the positive demand treatments ($p<0.001$), but do not significantly decrease their actions in response to the negative demand treatments ($p=0.221$). The overall sensitivity in response to demand is non-zero ($p<0.001$). False-discovery rate corrected p-values reach the same conclusions.

  - In column (2) we show that there is no significant treatment heterogeneity depending on whether choices are hypothetical or incentivized ($p=0.313$).

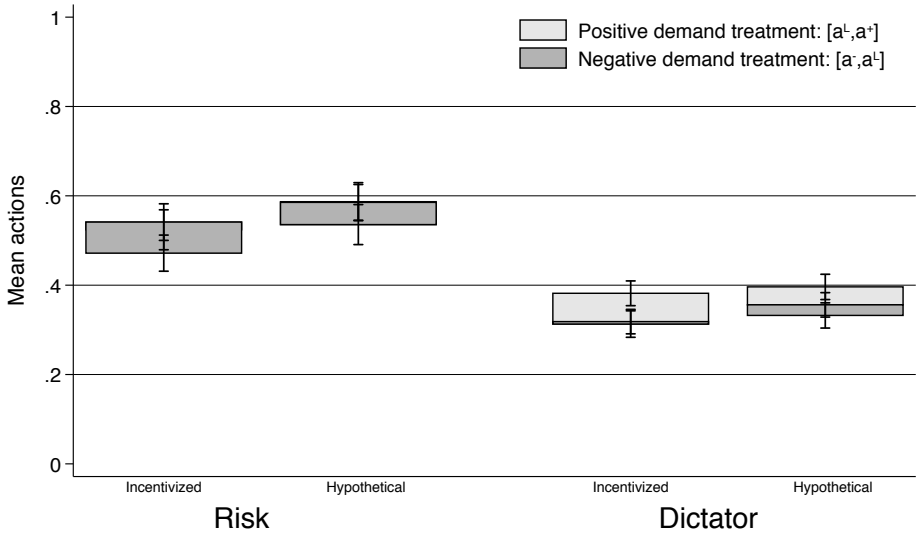
  - In column (3), we find no significant treatment heterogeneity in response to demand between men and women ($p=0.252$).

  - In column (4) we test whether attention moderates the response to demand treatments. We find no significant heterogeneity by attention ($p=0.530$).

  - In column (5) we examine heterogeneity across games. We find that overall sensitivity in the dictator game was significantly higher in the risk game ($p=0.046$).

- Table C10 explores how people's beliefs about whether the experimenter wanted (column 1) or expected (column 2) a high action. Table C10 shows the results for the tests we pre-specified.
  - People in the positive demand condition are more likely to think that the experimenter wanted a high action ($p<0.001$) and that the experimenter expected a high action ($p<0.001$).
  - People in the negative demand condition are less likely to think that the experimenter wanted a high action ($p<0.001$) and that the experimenter expected a high action ($p<0.001$).
  - Overall, people in the positive demand condition are significantly more likely to think that the experimenter wanted a high action or expected a high action compared to people in the negative demand condition ($p<0.001$).
  - False-discovery rate corrected p-values reach the same conclusions.
- In Table C11 we report results confirming that people in the incentive condition are more likely to believe that the task involved real money ($p<0.001$).
- In Table C9 we test for differences between strong and weak demand pooling data from experiment 1 and 2. We find that the overall sensitivity to strong demand is significantly higher pooling across games (($p<0.001$), as seen in column (1)), for the dictator game (($p<0.001$), as seen in column (2)), and for the risk game (($p<0.001$), as seen in column (3)). False-discovery rate corrected p-values reach the same conclusions.
- Table C12 shows that there was no differential attrition across treatment arms.

Figure C2. Overview of raw data: Experiment 2



Note: This figure summarizes the mean actions and corresponding 95 confidence intervals from experiment 2 across all 12 treatment arms

Table C5—Overview of raw data: experiment 2

	Risk Aversion		Dictator Game	
	Incentivized	Hypothetical	Incentivized	Hypothetical
Unconditional Means				
Positive demand: Mean	0.524	0.587	0.382	0.396
Positive demand: SD	0.348	0.335	0.222	0.224
No demand: Mean	0.541	0.536	0.313	0.356
No demand: SD	0.339	0.350	0.234	0.215
Negative demand: Mean	0.472	0.585	0.318	0.332
Negative demand: SD	0.317	0.325	0.226	0.219
Observations	739	734	758	719

Note: This table summarizes the raw action data from experiment 2 across all 12 treatment arms.

Table C6—Game-level regressions: experiment 2

	Risk Aversion	Dictator Game
Positive demand	0.017	0.055
	(0.022)	(0.014)
Negative demand	-0.008	-0.009
	(0.021)	(0.014)
Control Mean	0.539	0.334
Observations	1473	1477

Note: This table shows the effect of the positive and negative demand treatment at the game level based on the raw action data (pooling across incentivized and unincentivized choices).

Table C7—Weak Demand (Experiment 2)

	(1)	(2)	(3)	(4)	(5)
Positive Demand	0.127	0.153	0.085	0.067	0.206
	(0.043)	(0.060)	(0.056)	(0.116)	(0.054)
Negative Demand	-0.032	0.037	-0.029	-0.023	-0.036
	(0.042)	(0.060)	(0.055)	(0.109)	(0.054)
Pos. demand × interactant		-0.054	0.090	0.070	-0.155
		(0.085)	(0.086)	(0.124)	(0.085)
Neg. demand × interactant		-0.138	-0.006	-0.010	0.012
		(0.084)	(0.085)	(0.118)	(0.083)
Interactant		-0.066	-0.032	-0.217	0.192
		(0.060)	(0.061)	(0.081)	(0.060)
Interactant		Monetary Incentive	Male	Passed attention check	Risk
Adjusted R-squared	0.005	0.009	0.004	0.009	0.011
Pos. demand ≤ 0	0.001				
Adjusted p-value	0.010				
Neg. demand ≥ 0	0.221				
Adjusted p-value	0.070				
Pos. demand = neg. demand	0.000				
Adjusted p-value	0.010				
(Pos. - neg.) × interactant = 0		0.313	0.252	0.530	0.046
Observations	2950	2950	2950	2950	2950

Note: This table summarizes the results from experiment 2. The outcome variable (action chosen) is standardized at the game level using the mean and standard deviation of the negative demand group. Robust standard errors are in parentheses. Lower section of the table reports p-values on pre-specified hypothesis tests.

Table C8—Weak Demand (Experiment 2): controlling for imbalances

	(1)	(2)	(3)	(4)	(5)
Positive Demand	0.128	0.154	0.086	0.069	0.208
	(0.043)	(0.060)	(0.056)	(0.115)	(0.054)
Negative Demand	-0.032	0.037	-0.029	-0.023	-0.036
	(0.042)	(0.060)	(0.055)	(0.109)	(0.054)
Pos. demand × interactant		-0.054	0.089	0.069	-0.155
		(0.085)	(0.086)	(0.124)	(0.085)
Neg. demand × interactant		-0.139	-0.006	-0.010	0.012
		(0.084)	(0.085)	(0.118)	(0.083)
Interactant		-0.066	-0.030	-0.217	0.193
		(0.060)	(0.061)	(0.081)	(0.060)
Interactant		Monetary Incentive	Male	Passed attention check	Risk
Adjusted R-squared	0.004	0.009	0.004	0.009	0.011
Pos. demand ≤ 0	0.001				
Adjusted p-value	0.010				
Neg. demand ≥ 0	0.222				
Adjusted p-value	0.070				
Pos. demand = neg. demand	0.000				
Adjusted p-value	0.010				
(Pos. - neg.) × interactant = 0		0.307	0.255	0.538	0.045
Observations	2950	2950	2950	2950	2950

Note: This table summarizes the results from experiment 2. The outcome variable (action chosen) is standardized at the game-level using the mean and standard deviation of the negative demand group. Robust standard errors are in parentheses. Here we control for an indicator taking value 1 for those in part-time employment due to imbalance on this variable. Lower section of the table reports p-values on pre-specified hypothesis tests.

Table C9—Comparing experiments 1 and 2

	(1)	(2)	(3)
Positive Demand=1	0.127	0.206	0.051
	(0.043)	(0.054)	(0.065)
Experiment 1=1	-0.137	-0.140	-0.128
	(0.043)	(0.056)	(0.064)
Positive Demand=1 × Experiment 1=1	0.203	0.251	0.148
	(0.060)	(0.080)	(0.090)
Negative Demand=1	-0.032	-0.036	-0.024
	(0.042)	(0.054)	(0.063)
Negative Demand=1 × Experiment 1=1	-0.182	-0.167	-0.211
	(0.059)	(0.077)	(0.088)
Constant	-0.105	-0.203	-0.011
	(0.030)	(0.039)	(0.046)
Sample	All	Dictator Game	Investment
Adjusted R^2	0.034	0.056	0.021
H_0 : (Positive Demand - Negative Demand)*Interaction = 0	0.000	0.000	0.000
Adjusted p-value	0.001	0.001	0.001
Observations	5945	2980	2965

Note: This table uses action data from the investment game and dictator game from experiments 1 (strong demand treatments) and 2 (weak demand treatments), standardized at the game-experiment level using the mean and standard deviation of the negative demand treatment group. The dummy experiment 1 takes value 1 for respondents from experiment 1. Column (1) pools the data from both games, column (2) uses dictator game data and column (3) investment game data. Adjusted p-values are corrected for false-discovery rate across the three tests.

Table C10—Beliefs about the experimental objective and hypothesis: Weak Demand (Experiment 2)

	Belief: Want High	Belief: Expect High
Positive - Negative	0.332	0.402
	(0.021)	(0.020)
Adjusted p-value	[0.001]	[0.001]
Positive - Neutral	0.171	0.217
	(0.022)	(0.022)
Adjusted p-value	[0.001]	[0.001]
Negative - Neutral	-0.161	-0.185
	(0.022)	(0.020)
Adjusted p-value	[0.001]	[0.001]
Mean (No Demand)	0.485	0.392
Observations	2950	2950

Note: This table uses data from all respondents who completed experiment 2. The outcome variables take value one if the respondents believed that the experimenter wanted (column 1) or expected (column 2) a high action and zero if they wanted or expected a low action.

Table C11—Beliefs about whether the experiment is incentivized

	(1) Belief: Real Money
Monetary Incentive	0.367
	(0.016)
Control Mean	0.139
R^2	0.153
Observations	2950

Note: This table uses data from all respondents who completed experiment 2. The outcome variable takes value one if the respondent believes that the tasks in the experiment involve real money and value zero otherwise. Monetary incentive takes value 1 for respondents whose choices were incentivized, and takes value 0 for respondents whose choices were hypothetical.

Table C12—Attrition across treatment arms

	(1) Finished
Positive Demand	0.00601
	(0.003)
Negative Demand	0.00104
	(0.003)
Mean (no demand)	0.990
R^2	0.00145
Observations	2964

Note: This table uses data from all respondents who started experiment 2. Finished takes value one for all respondents who completed the experiment.

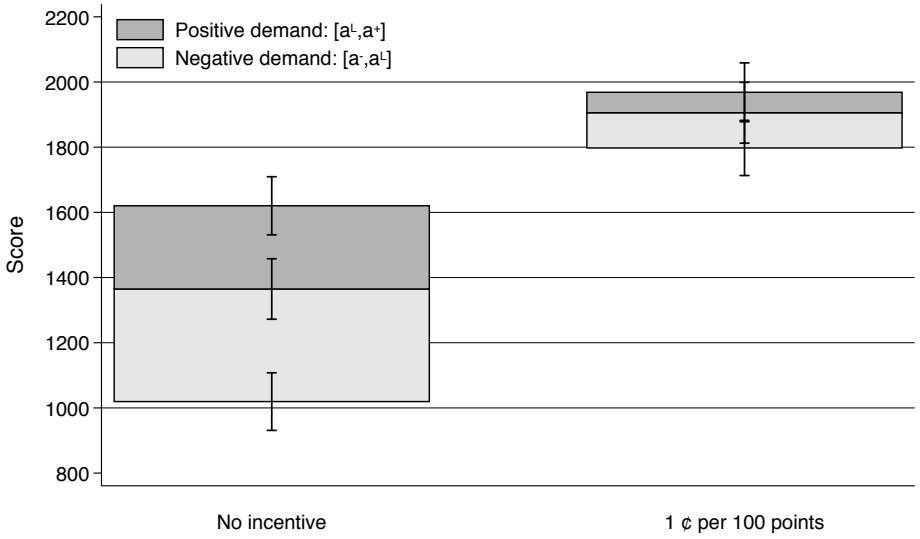
C3. Pre-analysis Plan 3

- Balance tests for this experiment are in Table D3 in Section D. We found a slight imbalance on an indicator for Hispanic race. Table C15 shows our main results controlling for this variable.
- Table C13 and Figure C3 summarizes the means, standard errors, and corresponding 95 percent confidence intervals from experiment 3 across all 6 treatment arms (excluding the 4 cent treatment that was used only for structural estimation).
- Table C14 displays the main effects of the positive and negative demand treatment as well as heterogeneous treatment effects by gender, and whether people are paid a 1-cent bonus or no bonus. This table summarizes the results for the main pre-specified tests.
  - Column (1) shows that people increase their effort in response to the positive demand treatments ($p<0.001$), decrease their effort in response to the negative demand treatments ($p<0.001$). Moreover, the overall sensitivity in response to demand is non-zero ($p<0.001$). False-discovery rate corrected p-values reach the same conclusions.
  - In column (2), we test for systematic differences by incentive level in response to the demand treatments. Sensitivity was higher in the no-incentive condition compared to the 1-cent incentive condition ($p<0.001$).

  - In column (3), we test for gender differences in response to the demand treatments. Sensitivity was not significantly different for women than for men (p=0.946).

- Table C16 explores beliefs about whether the experimenter wanted (column 1) or expected (column 2) a high action.
  - People in the positive demand condition are more likely to think that the experimenter wanted a high action (p<0.001) and that the experimenter expected a high action (p<0.001).
  - People in the negative demand condition are less likely to think that the experimenter wanted a high action (p<0.001) and that the experimenter expected a high action (p<0.001).
  - Overall, people in the positive demand condition are significantly more likely to think that the experimenter wanted a high action or expected a high action compared to people in the negative demand condition (p<0.001).
- False-discovery rate corrected p-values reach the same conclusions.
- In Table C17, we show that there is no differential attrition across treatment arms.

FIGURE C3. OVERVIEW OF RAW DATA: EXPERIMENT 3



Note: This figure summarizes the mean actions (expressed in points scored) and corresponding 95 confidence intervals from experiment 3 across all 6 treatment arms.

TABLE C13—OVERVIEW OF RAW DATA: EXPERIMENT 3

	Effort	
	1-cent bonus	No bonus
Unconditional Means		
Positive demand: Mean	0.492	0.405
Positive demand: SD	0.179	0.177
No demand: Mean	0.476	0.341
No demand: SD	0.184	0.182
Negative demand: Mean	0.449	0.255
Negative demand: SD	0.162	0.176
Observations	714	731

Note: This table summarizes the raw action data from experiment 3 across all 6 treatment arms.

Table C14—Effort (z-scored) with strong demand

	(1)	(2)	(3)
Positive Demand	0.209	0.333	0.313
	(0.061)	(0.085)	(0.107)
Negative Demand	-0.305	-0.450	-0.198
	(0.061)	(0.085)	(0.103)
Positive demand × interactant		-0.249	-0.182
		(0.123)	(0.129)
Negative demand × interactant		0.305	-0.191
		(0.121)	(0.126)
Interactant		0.082	0.136
		(0.088)	(0.093)
Constant	0.068	0.027	-0.009
	(0.044)	(0.061)	(0.078)
Interactant		1-cent incentive	Male
Adjusted R^2	0.046	0.061	0.046
Positive demand ≤ 0	0.000		
Adjusted p-value	0.010		
Negative demand ≥ 0	0.000		
Adjusted p-value	0.010		
Positive demand = negative demand	0.000		
Adjusted p-value	0.010		
(Positive demand - negative demand)* interaction = 0		0.000	0.946
Observations	1445	1445	1445

Note: This table summarizes the results from experiment 3. The outcome variable (action chosen) is standardized at the incentive treatment level using the mean and standard deviation of the negative demand group. Lower section of the table reports p-values on pre-specified hypothesis tests.

Table C15—Effort (z-scored) with strong demand: with control for imbalance

	(1)	(2)	(3)
Positive Demand	0.221	0.344	0.319
	(0.061)	(0.085)	(0.107)
Negative Demand	-0.299	-0.444	-0.200
	(0.061)	(0.085)	(0.103)
Positive demand × interactant		-0.248	-0.173
		(0.122)	(0.128)
Negative demand × interactant		0.305	-0.177
		(0.120)	(0.126)
Interactant		0.081	0.127
		(0.088)	(0.092)
Constant	0.050	0.009	-0.022
	(0.045)	(0.061)	(0.078)
Interactant		1-cent incentive	Male
Adjusted R^2	0.049	0.064	0.049
Positive demand ≤ 0	0.000		
Adjusted p-value	0.010		
Negative demand ≥ 0	0.000		
Adjusted p-value	0.010		
Positive demand = negative demand	0.000		
Adjusted p-value	0.010		
(Positive demand - negative demand)* interaction = 0		0.000	0.974
Observations	1445	1445	1445

Note: This table summarizes the results from experiment 3. The outcome variable (action chosen) is standardized at the incentive treatment level using the mean and standard deviation of the negative demand group. Here we control for an indicator taking value 1 for Hispanics as we found an imbalance for this variable across demand treatment arms. Lower section of the table reports p-values on pre-specified hypothesis tests.

Table C16—Beliefs: Effort with strong demand

	Belief: Want High	Belief: Expect High
Positive - Negative	0.459	0.414
	(0.027)	(0.028)
Adjusted p-value	[0.001]	[0.001]
Positive - Neutral	0.170	0.190
	(0.026)	(0.028)
Adjusted p-value	[0.001]	[0.001]
Negative - Neutral	-0.289	-0.224
	(0.031)	(0.031)
Adjusted p-value	[0.001]	[0.001]
Mean (No Demand)	0.688	0.640
Observations	1445	1445

Note: The outcome variables take value one if the respondents believed that the experimenter wanted (column 1) or expected (column 2) a high action and zero if they wanted or expected a low action.

Table C17—Attrition across treatment arms

	(1) Finished
Positive Demand	-0.000449
	(0.009)
Negative Demand	0.00679
	(0.010)
Mean (no demand)	0.990
R^2	0.000366
Observations	1739

Note: Finished takes value one for all respondents who completed the experiment.

- Balance tests for this experiment are in Table D4 in Section D and indicate that there are no imbalances.

- Table C18 and Figure C4 summarize the means, standard errors, and corresponding 95 percent confidence intervals from experiment 4 across all treatment arms. Table C19 displays the game-level regressions based on the raw data showing the control mean from the "no demand condition" as well as the coefficients on the positive demand treatment indicator and the negative demand treatment indicator.

- Table C20 displays the main effects of the positive and negative demand treatment as well as heterogeneous treatment effects by strong vs. weak demand treatment, gender, attention and game. This table also summarizes the results for the main pre-specified tests.

  - Column (1) shows that people increase their actions in response to the positive demand treatments ($p<0.001$), decrease their actions in response to the negative demand treatments ($p<0.001$). Moreover, the overall sensitivity in response to demand is non-zero ($p<0.001$). False-discovery rate corrected p-values reach the same conclusions.
  - Pooling across games, column (2) finds that sensitivity was significantly higher in the strong treatments than the weak treatments ($p<0.001$).
  - Pooling across games and demand treatments, column (3) finds sensitivity was significantly higher for women than for men ($p=0.014$).
  - Pooling across games and demand treatments, column (4) finds that sensitivity was significantly higher for attentive respondents than for inattentive respondents ($p<0.001$).
  - Pooling across demand treatments, column (5) finds that sensitivity in the dictator game was significantly higher than sensitivity in the risk game ($p=0.001$).

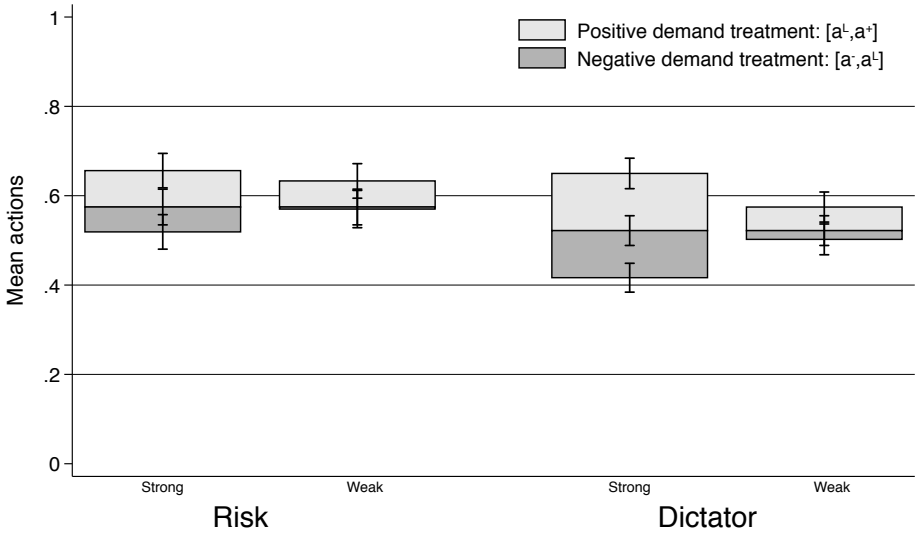
- Table C22 explores how people's beliefs about whether the experimenter wanted (column 1) or expected (column 2) a high action. Table C22 shows the results for the tests we pre-specified.

  - People in the positive demand condition are more likely to think that the experimenter wanted a high action ($p<0.001$) and that the experimenter expected a high action ($p<0.001$).
  - People in the negative demand condition are less likely to think that the experimenter wanted a high action ($p<0.001$) and that the experimenter expected a high action ($p<0.001$).
  - Overall, people in the positive demand condition are significantly more likely to think that the experimenter wanted a high action ($p<0.001$) or expected a high action compared to people in the negative demand condition ($p<0.001$).
  - False-discovery rate corrected p-values reach the same conclusions.

- Table C21 examines demand sensitivity by population and shows that there were no systematic differences ($p=0.602$) when pooling across games.

- Table C23 shows that there was no differential attrition across treatment arms.

FIGURE C4. OVERVIEW OF RAW DATA: EXPERIMENT 4



Note: This figure summarizes the means and corresponding 95 confidence intervals from experiment 4 across all treatment arms.

Table C18—Overview of raw data: experiment 4

	Risk Aversion		Dictator Game	
	Strong	Weak	Strong	Weak
Unconditional Means				
Positive demand: Mean	0.656	0.633	0.650	0.575
Positive demand: SD	0.341	0.337	0.300	0.292
No demand: Mean	0.575	0.575	0.522	0.522
No demand: SD	0.358	0.358	0.289	0.289
Negative demand: Mean	0.519	0.570	0.416	0.502
Negative demand: SD	0.331	0.351	0.286	0.291
Observations	900	880	896	862

Note: This table summarizes the raw data from experiment 4 across all treatment arms.

Table C19—Game-level regressions: experiment 4

	Risk Aversion	Dictator Game
Positive demand	0.070	0.091
	(0.025)	(0.021)
Negative demand	-0.031	-0.065
	(0.025)	(0.021)
Control Mean	0.575	0.522
Observations	1468	1465

Note: This table shows the effect of the positive and negative demand treatment at the game level based on the raw data (pooling across strong and weak demand treatments).

Table C20—Representative Sample with strong and weak demand treatments (Experiment 4)

	(1)	(2)	(3)	(4)	(5)
Positive Demand	0.281	0.193	0.322	0.244	0.555
	(0.055)	(0.063)	(0.063)	(0.061)	(0.064)
Negative Demand	-0.159	-0.037	-0.224	-0.083	-0.033
	(0.055)	(0.064)	(0.061)	(0.061)	(0.064)
Pos. demand × interactant		0.175	-0.084	0.113	-0.545
		(0.064)	(0.064)	(0.064)	(0.062)
Neg. demand × interactant		-0.237	0.139	-0.222	-0.257
		(0.063)	(0.064)	(0.063)	(0.063)
Interactant		Strong demand treatment	Male	Passed attention check	Risk
Adjusted R-squared	0.031	0.038	0.033	0.035	0.060
Pos. demand ≤ 0	0.000				
Adjusted p-value	0.010				
Neg. demand ≥ 0	0.002				
Adjusted p-value	0.010				
Pos. = neg. demand	0.000				
Adjusted p-value	0.010				
(Pos. - neg.) × interactant = 0		0.000	0.014	0.000	0.001
Observations	2933	2933	2933	2933	2933

Note: This table summarizes the results from experiment 4. The outcome variable (action chosen) is standardized at the game level using the mean and standard deviation of the negative demand group. Lower section of the table reports p-values on pre-specified hypothesis tests.

Table C21—Comparing Representative and MTurk Samples

	(1)	(2)	(3)
Positive Demand=1	0.261	0.423	0.095
	(0.043)	(0.057)	(0.064)
Representative Sample=1	0.522	0.851	0.208
	(0.054)	(0.076)	(0.075)
Positive Demand=1 × Representative Sample=1	0.021	-0.079	0.114
	(0.070)	(0.097)	(0.097)
Negative Demand=1	-0.148	-0.042	-0.251
	(0.042)	(0.056)	(0.061)
Negative Demand=1 × Representative Sample=1	-0.011	-0.202	0.160
	(0.069)	(0.096)	(0.096)
Constant	-0.226	-0.344	-0.111
	(0.031)	(0.041)	(0.045)
Sample	All	Dictator Game	Investment
Adjusted R^2	0.093	0.165	0.041
H_0 : (Positive Demand - Negative Demand)*Repres. Sample = 0	0.602	0.149	0.592
Adjusted p-value	0.813	0.813	0.813
Observations	5928	2993	2935

Note: This table uses data from the incentivized MTurk respondents from experiments 1 and 2 and the representative online panel (experiment 4). Representative Sample is a dummy variable taking value 1 for respondents from the representative online panel and value zero for the MTurk respondents. Lower section of the table reports p-values on pre-specified hypothesis tests.

Table C22—Beliefs about the experimental objective and hypothesis: Representative Sample

	Belief: Want High	Belief: Expect High
Positive - Negative	0.206	0.205
	(0.020)	(0.020)
Adjusted p-value	[0.001]	[0.001]
Positive - Neutral	0.068	0.091
	(0.024)	(0.025)
Adjusted p-value	[0.001]	[0.001]
Negative - Neutral	-0.138	-0.114
	(0.025)	(0.025)
Adjusted p-value	[0.001]	[0.001]
Mean (No Demand)	0.602	0.511
Observations	2933	2933

Note: The outcome variables take value one if the respondents believed that the experimenter wanted (column 1) or expected (column 2) a high action and zero if they wanted or expected a low action.

Table C23—Attrition across treatment arms: Experiment 4

	(1) Finished
Positive Demand	0.000710
	(0.004)
Negative Demand	-0.000400
	(0.004)
Mean (no demand)	0.990
R^2	0.0000390
Observations	2952

Note: This table uses data from all respondents who started experiment 4. Finished takes value one for all respondents who completed the experiment.

C5. Pre-analysis Plan 5

This plan encompasses experiments 5 and 6 and pre-specified the collecting of all incentivized MTurk data together by demand treatment type, to present results in single tables and figures.

- Balance tests for the experiments are Tables D5 and D6 in Section D and indicate that there are no imbalances.
- Figure 2 and Tables 2 and 1 (included in the paper) summarize the raw data and sensitivities across games.
- Next, we consider whether sensitivity to weak demand treatments differs across games. In Table C24 we show little evidence of statistically significant differences in sensitivity across all games (p=0.241 with effort tasks included, p=0.437 when excluded).
- Table C24 shows that there are large differences in sensitivity across games in response to strong demand both when all 11 games are considered and when the effort tasks are excluded (p<0.001).
- In Table C25 we conduct a pooled test with all MTurk experiments examining whether sensitivity varies between strong and weak demand treatments. We find a larger response to strong compared to weak demand treatments (p<0.001).

Table C24—Differences in response to demand across games

	(1)	(2)
Positive Demand=1	1.058	0.289
	(0.133)	(0.125)
Ambiguity	0.149	0.007
	(0.110)	(0.102)
DG	-0.153	-0.248
	(0.103)	(0.089)
Effort: incentive	0.332	0.085
	(0.104)	(0.100)
Effort: no incentive	-0.056	0.049
	(0.105)	(0.101)
Lying	0.241	0.015
	(0.123)	(0.102)
Risk	-0.138	-0.195
	(0.103)	(0.095)
Time	0.100	0.021
	(0.113)	(0.099)
Trust	0.124	0.015
	(0.109)	(0.105)
UG 1	0.137	0.015
	(0.118)	(0.104)
UG 2	0.230	0.015
	(0.118)	(0.100)
Positive Demand=1 × Ambiguity	-0.596	-0.116
	(0.165)	(0.161)
Positive Demand=1 × DG	-0.364	-0.049
	(0.156)	(0.146)
Positive Demand=1 × Effort: incentive	-0.829	-0.211
	(0.158)	(0.156)
Positive Demand=1 × Effort: no incentive	-0.275	-0.352
	(0.157)	(0.161)
Positive Demand=1 × Lying	-0.454	-0.247
	(0.178)	(0.161)
Positive Demand=1 × Risk	-0.530	-0.133
	(0.156)	(0.155)
Positive Demand=1 × Time	-0.709	-0.277
	(0.164)	(0.158)
Positive Demand=1 × Trust	-0.495	-0.213
	(0.165)	(0.163)
Positive Demand=1 × UG 1	-0.374	-0.132
	(0.172)	(0.168)
Positive Demand=1 × UG 2	-0.308	-0.008
	(0.173)	(0.161)
Constant	-0.367	-0.015
	(0.087)	(0.072)
Treatment	Strong	Weak
Adjusted R^2	0.102	0.012
P-value(Omnibus F-Test)	0.000	0.241
Adjusted p-values	0.001	0.191
P-value(Omnibus F-Test): without effort tasks	0.001	0.437
Adjusted p-values	0.001	0.279
Observations	4800	4450

Note: Outcome variable (action chosen) is standardized at the game level. We pool all real stakes MTurk observations across all experiments. Column (1) presents results from the strong demand treatments and column 2 presents results from the weak demand treatments.

Table C25—Differences in response to strong vs. weak demand treatments

	(1) Z-scored behavior
Strong × Positive Demand	0.471
	(0.042)
Positive demand	0.133
	(0.030)
R^2	0.0455
Observations	9250

Note: Outcome variable (action chosen) is standardized at the game level. We pool all real stakes MTurk observations across all experiments.

- Balance tests for this experiment are in Table D7 in Section D. We found a slight imbalance for men and people in part-time employment. Therefore, in Table C26 we show the main results controlling for an indicator taking value 1 for men and an indicator taking value 1 for people in part-time employment.

- Figures 3 (included in the paper) and C5 present the raw data graphically. Figure 3 plots task 2 actions against task 1 actions – points above the 45 degree line correspond to increases in actions. Figure C5 plots the distributions of changes of actions between task 1 and task 2 (positive difference means action increased in task 2).

- Table A2 (included in the paper) Panel A summarizes behavior in tasks 1 and 2. The relevant columns are headed "Within," Task 1 choices are labeled "no demand" and Task 2 choices are either labeled Positive or Negative demand. Panel B of Table A2 shows the sensitivities computed from raw or standardized (at the game level) task 2 actions.

- Table A2 compares sensitivity estimates and raw choices from the "within" experiment (experiment 7) and the incentivized MTurk "between experiment" with strong demand treatments (experiment 1). We do not find statistically significant differences in sensitivity.

- Table A5 (in the main web Appendix) documents for each game and demand treatment the number of strict compliers and strict defiers. Defiance rates were very low at around 5 percent.

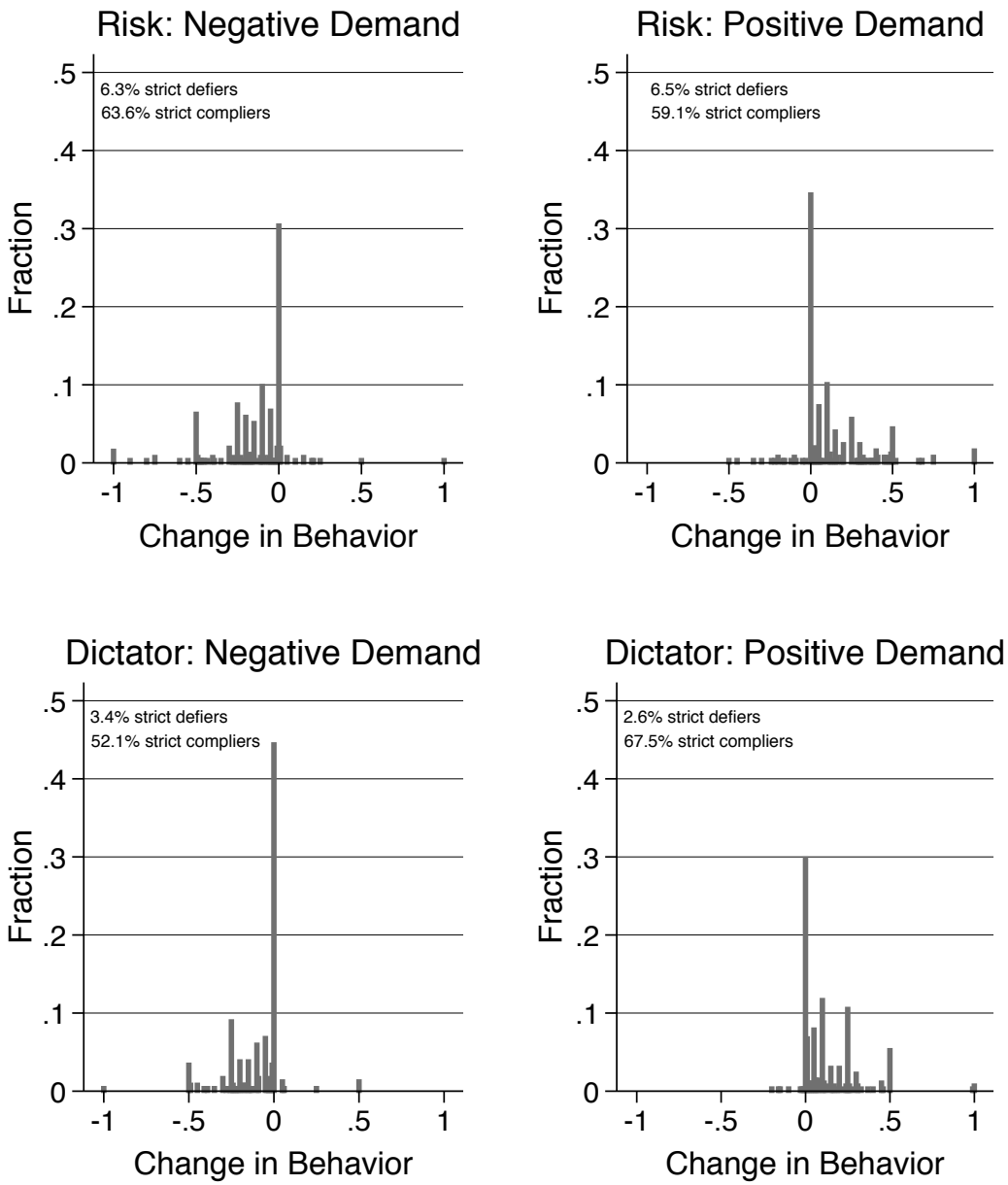
- Table A5 also displays the average change in action between tasks 1 and 2 for each treatment arm and for compliers and defiers separately.

Table C26—Within Design (Experiment 7): with controls for imbalance

	Dictator			Risk		
	Within	Between	Difference	Within	Between	Difference
Panel A: Unconditional Means						
Positive demand	0.384	0.434	-0.045	0.560	0.550	0.010
	(0.017)	(0.015)	(0.023)	(0.021)	(0.020)	(0.029)
No demand	0.273	0.282	-0.005	0.448	0.466	-0.023
	(0.011)	(0.015)	(0.019)	(0.015)	(0.022)	(0.026)
Negative demand	0.195	0.251	-0.044	0.318	0.373	-0.058
	(0.014)	(0.014)	(0.021)	(0.019)	(0.019)	(0.027)
Panel B: Sensitivity (positive - negative)						
Raw data	0.182	0.190	-0.007	0.244	0.177	0.068
	(0.023)	(0.021)	(0.031)	(0.029)	(0.027)	(0.040)
Z-score	0.763	0.745	-0.005	0.715	0.520	0.197
	(0.095)	(0.086)	(0.126)	(0.084)	(0.080)	(0.117)
Panel C: Monotonicity						
Positive - Neutral (z-score)	0.514	0.617	-0.110	0.377	0.248	0.137
	(0.044)	(0.088)	(0.128)	(0.041)	(0.087)	(0.124)
Negative - Neutral (z-score)	-0.380	-0.128	-0.252	-0.427	-0.272	-0.155
	(0.045)	(0.086)	(0.122)	(0.042)	(0.084)	(0.119)
Observations	499	770	1269	500	728	1228

Note: This table uses data from the within design (experiment 7) and incentivized choices from the dictator game and the investment game in experiment 1. These experiments employ strong demand treatments. Here we control for an indicator taking value 1 for men and another indicator value taking value 1 for people in part-time employment. as we had found an imbalance for this variable across demand treatment arms

Figure C5. Distribution of Responses: Within Design



Note: This figure uses MTurk data from experiment 7 and displays the distribution of changes in behavior (in task 2 compared to task 1) to our strong demand treatments.

D. Balance tables, summary statistics, and attrition

Table D1—Balance Table: Experiment 1 (Strong Demand)

	No demand	Pos. demand	Neg. demand	P-value(Pos. demand - no demand)	P-value(Neg. demand - no demand)	P-value(Pos. demand - neg. demand)	Observations
Male	0.511	0.519	0.497	0.639	0.442	0.215	4479
Income	51560.364	52459.736	53429.878	0.387	0.068	0.346	3995
Age	36.226	36.464	36.414	0.560	0.653	0.904	4479
Household Size	3.711	3.649	3.626	0.225	0.097	0.640	4479
White	0.773	0.785	0.774	0.429	0.974	0.451	4479
Black	0.070	0.067	0.071	0.676	0.927	0.612	4479
Hispanic	0.053	0.057	0.055	0.592	0.824	0.757	4479
Asian	0.079	0.063	0.075	0.091	0.703	0.194	4479
Full-time employment	0.484	0.508	0.522	0.194	0.039	0.430	4479
Part-time employment	0.128	0.120	0.115	0.541	0.280	0.632	4479
Unemployed	0.144	0.133	0.130	0.409	0.269	0.770	4479
Bachelor Degree	0.353	0.369	0.389	0.360	0.044	0.264	4479
Conservative	0.230	0.238	0.242	0.641	0.459	0.778	4441
Number of HITs	9393.289	9217.178	8651.406	0.762	0.202	0.321	4479
Joint							

Note: In this table we present evidence on the experimental integrity in experiment 1. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.9110. The p-value of the joint F-test when comparing covariates in the positive and no-demand demand condition is 0.6965. The p-value of the joint F-test when comparing covariates in the negative and no-demand demand condition is 0.2402.

Table D2—Balance Table: Experiment 2 (Weak Demand)

	No demand	Pos. demand	Neg. demand	P-value(Pos. demand - no demand)	P-value(Neg. demand - no demand)	P-value(Pos. demand - neg. demand)	Observations
Male	0.466	0.467	0.477	0.949	0.630	0.678	2950
Income	51017.241	51410.405	51949.829	0.757	0.455	0.664	2612
Age	35.909	35.871	35.227	0.940	0.171	0.197	2950
Household Size	3.696	3.686	3.757	0.880	0.346	0.275	2950
White	0.785	0.761	0.749	0.194	0.060	0.568	2950
Black	0.069	0.076	0.076	0.525	0.517	0.994	2950
Hispanic	0.054	0.051	0.057	0.756	0.723	0.507	2950
Asian	0.066	0.070	0.087	0.705	0.083	0.178	2950
Full-time employment	0.493	0.464	0.467	0.199	0.236	0.915	2950
Part-time employment	0.130	0.099	0.125	0.033	0.742	0.071	2950
Unemployed	0.101	0.139	0.129	0.010	0.056	0.493	2950
Bachelor Degree	0.367	0.353	0.376	0.524	0.662	0.283	2950
Conservative	0.273	0.254	0.243	0.342	0.131	0.583	2927
Number of HITs	5854.863	5642.157	5306.841	0.703	0.314	0.529	2950

Note: In this table we present evidence on the experimental integrity in experiment 2. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.7084. The p-value of the joint F-test when comparing covariates in the positive and no-demand demand condition is 0.2332. The p-value of the joint F-test when comparing covariates in the negative and no-demand demand condition is 0.4838.

Table D3—Balance Table: Experiment 3 (Effort Experiment with strong demand)

	No demand	Pos. demand	Neg. demand	P-value(Pos. demand - no demand)	P-value(Neg. demand - no demand)	P-value(Pos. demand - neg. demand)	Observations
Male	0.558	0.574	0.536	0.598	0.439	0.235	1691
Income	33596.974	32213.115	32457.983	0.170	0.265	0.822	1691
Age	37.444	37.359	36.618	0.906	0.251	0.352	1691
Household Size	3.750	3.783	3.771	0.690	0.790	0.894	1691
White	0.752	0.783	0.761	0.217	0.749	0.411	1691
Black	0.109	0.084	0.084	0.149	0.152	0.999	1691
Hispanic	0.055	0.025	0.046	0.006	0.493	0.070	1691
Asian	0.065	0.072	0.074	0.634	0.555	0.914	1691
Full-time employment	0.509	0.498	0.536	0.707	0.364	0.241	1691
Part-time employment	0.125	0.125	0.107	0.993	0.337	0.387	1691
Unemployed	0.105	0.121	0.107	0.380	0.886	0.502	1691
Bachelor Degree	0.395	0.355	0.370	0.155	0.382	0.623	1691
Republican	0.250	0.289	0.273	0.139	0.382	0.585	1691

Note: In this table we present evidence on the integrity of the randomization in experiment 3. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.9171. The p-value of the joint F-test when comparing covariates in the positive and no-demand demand condition is 0.1012. The p-value of the joint F-test when comparing covariates in the negative and no-demand demand condition is 0.4845.

Table D4—Balance Table: Experiment 4 (Representative Sample)

	No demand	Pos. demand	Neg. demand	P-value(Pos. demand - no demand)	P-value(Neg. demand - no demand)	P-value(Pos. demand - neg. demand)	Observations
Male	0.488	0.485	0.468	0.912	0.446	0.428	2933
Income	68357.264	65309.037	67175.470	0.253	0.662	0.398	2882
Age	47.972	46.899	47.879	0.195	0.911	0.147	2933
Household Size	3.332	3.312	3.333	0.752	0.983	0.692	2926
White	0.801	0.772	0.784	0.159	0.399	0.505	2927
Black	0.070	0.069	0.062	0.968	0.518	0.457	2927
Hispanic	0.051	0.064	0.062	0.269	0.379	0.799	2927
Asian	0.043	0.061	0.062	0.104	0.079	0.866	2927
Full-time employment	0.499	0.485	0.496	0.566	0.888	0.604	2933
Part-time employment	0.074	0.078	0.091	0.768	0.218	0.263	2933
Unemployed	0.068	0.050	0.052	0.132	0.188	0.818	2933
Bachelor Degree	0.331	0.352	0.330	0.368	0.975	0.262	2933
Conservative	0.350	0.352	0.351	0.921	0.962	0.951	2797

Note: In this table we present evidence on the integrity of the randomization in experiment 4. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.7455. The p-value of the joint F-test when comparing covariates in the positive and no-demand demand condition is 0.4909. The p-value of the joint F-test when comparing covariates in the negative and no-demand demand condition is 0.6390.

Table D5—Balance Table: Experiment 5 (Many Task experiment)

	Pos. demand	Neg. demand	P-value(Pos. demand - neg. demand)	Observations
Male	0.453	0.472	0.168	5045
Income	53324.385	52746.322	0.460	4478
Age	37.328	37.207	0.714	5045
Household Size	3.711	3.654	0.159	5045
White	0.770	0.776	0.620	5045
Black	0.078	0.072	0.434	5045
Hispanic	0.048	0.048	0.956	5045
Asian	0.075	0.078	0.773	5045
Full-time employment	0.513	0.516	0.819	5045
Part-time employment	0.115	0.113	0.830	5045
Unemployed	0.126	0.140	0.147	5045
Bachelor Degree	0.376	0.372	0.745	5045
Conservative	0.263	0.257	0.636	5019
Number of HITs	9381.041	8544.300	0.055	5045

Note: In this table we present evidence on the integrity of the randomization in experiment 5. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.1990.

Table D6—Balance Table: Experiment 6 (Effort Experiment with weak demand treatments)

	Pos. demand	Neg. demand	P-value(Pos. demand - neg. demand)	Observations
Male	0.547	0.556	0.803	769
Income	32317.708	32532.468	0.861	769
Age	37.339	37.545	0.807	769
Household Size	3.732	3.681	0.626	769
White	0.755	0.730	0.422	769
Black	0.083	0.083	0.991	769
Hispanic	0.055	0.073	0.306	769
Asian	0.081	0.075	0.780	769
Full-time employment	0.552	0.527	0.491	769
Part-time employment	0.128	0.094	0.132	769
Unemployed	0.125	0.122	0.902	769
Bachelor Degree	0.432	0.379	0.134	769
Conservative	0.266	0.325	0.078	764

Note: In this table we present evidence on the integrity of the randomization in experiment 6. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.2562.

Table D7—Balance Table: Experiment 7 (Within Design)

	Pos. demand	Neg. demand	P-value(Pos. demand - neg. demand)	Observations
Male	0.545	0.610	0.038	999
Income	53645.374	54549.763	0.604	876
Age	34.465	34.439	0.970	999
Household Size	3.533	3.540	0.938	999
White	0.732	0.743	0.696	999
Black	0.078	0.078	0.995	999
Hispanic	0.064	0.049	0.300	999
Asian	0.090	0.109	0.317	999
Full-time employment	0.520	0.569	0.118	999
Part-time employment	0.133	0.092	0.043	999
Unemployed	0.137	0.125	0.592	999
Bachelor Degree	0.408	0.386	0.475	999
Conservative	0.241	0.233	0.776	994

Note: In this table we present evidence on balance for experiment 7. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.043.

Table D8—Summary Statistics: Pooled across all experiments

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.50	0.50	0.00	0.00	1.00	18866
Income	52105.70	32773.39	45000.00	5000.00	225000.00	17303
Age	38.25	13.02	35.00	17.00	116.00	18866
Household Size	3.63	1.40	3.00	2.00	13.00	18859
White	0.77	0.42	1.00	0.00	1.00	18860
Black	0.07	0.26	0.00	0.00	1.00	18860
Hispanic	0.05	0.22	0.00	0.00	1.00	18860
Asian	0.07	0.26	0.00	0.00	1.00	18860
Full-time employment	0.50	0.50	1.00	0.00	1.00	18866
Part-time employment	0.11	0.32	0.00	0.00	1.00	18866
Unemployed	0.12	0.32	0.00	0.00	1.00	18866
Bachelor Degree	0.37	0.48	0.00	0.00	1.00	18866
Conservative	0.27	0.44	0.00	0.00	1.00	16942
Number of HITs	8215.31	14921.14	2500.00	750.00	75000.00	12474

Note: This table summarizes the main covariates of all respondents across all 6 experiments.

Table D9—Summary Statistics: Experiment 1 (Strong demand)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.51	0.50	1.00	0.00	1.00	4479
Income	52481.85	26617.99	55000.00	5000.00	100000.00	3995
Age	36.37	11.26	33.00	19.00	88.00	4479
Household Size	3.66	1.40	3.00	2.00	11.00	4479
White	0.78	0.42	1.00	0.00	1.00	4479
Black	0.07	0.25	0.00	0.00	1.00	4479
Hispanic	0.06	0.23	0.00	0.00	1.00	4479
Asian	0.07	0.26	0.00	0.00	1.00	4479
Full-time employment	0.50	0.50	1.00	0.00	1.00	4479
Part-time employment	0.12	0.33	0.00	0.00	1.00	4479
Unemployed	0.14	0.34	0.00	0.00	1.00	4479
Bachelor Degree	0.37	0.48	0.00	0.00	1.00	4479
Conservative	0.24	0.43	0.00	0.00	1.00	4441
Number of HITs	9091.59	15766.32	2500.00	750.00	75000.00	4479

Note: This table summarizes the main covariates of all respondents in experiment 1.

Table D10—Summary Statistics: Experiment 2 (Weak demand)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.47	0.50	0.00	0.00	1.00	2950
Income	51460.57	26145.92	55000.00	5000.00	100000.00	2612
Age	35.67	11.09	33.00	19.00	81.00	2950
Household Size	3.71	1.43	3.00	2.00	13.00	2950
White	0.77	0.42	1.00	0.00	1.00	2950
Black	0.07	0.26	0.00	0.00	1.00	2950
Hispanic	0.05	0.23	0.00	0.00	1.00	2950
Asian	0.07	0.26	0.00	0.00	1.00	2950
Full-time employment	0.47	0.50	0.00	0.00	1.00	2950
Part-time employment	0.12	0.32	0.00	0.00	1.00	2950
Unemployed	0.12	0.33	0.00	0.00	1.00	2950
Bachelor Degree	0.37	0.48	0.00	0.00	1.00	2950
Conservative	0.26	0.44	0.00	0.00	1.00	2927
Number of HITs	5600.34	12081.52	1500.00	750.00	75000.00	2950

Note: This table summarizes the main covariates of all respondents in experiment 2.

Table D11—Summary Statistics: Experiment 3 (Effort Experiment: Strong demand)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.56	0.50	1.00	0.00	1.00	1691
Income	32877.00	17304.76	35000.00	5000.00	85000.00	1691
Age	37.19	12.32	36.00	21.00	70.00	1691
Household Size	3.77	1.39	4.00	2.00	12.00	1691
White	0.76	0.43	1.00	0.00	1.00	1691
Black	0.09	0.29	0.00	0.00	1.00	1691
Hispanic	0.04	0.20	0.00	0.00	1.00	1691
Asian	0.07	0.25	0.00	0.00	1.00	1691
Full-time employment	0.51	0.50	1.00	0.00	1.00	1691
Part-time employment	0.12	0.33	0.00	0.00	1.00	1691
Unemployed	0.11	0.31	0.00	0.00	1.00	1691
Bachelor Degree	0.38	0.48	0.00	0.00	1.00	1691
Republican	0.27	0.44	0.00	0.00	1.00	1691

Note: This table summarizes the main covariates of all respondents in experiment 3.

Table D12—Summary Statistics: Experiment 4 (Representative sample)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.48	0.50	0.00	0.00	1.00	2933
Income	66658.57	52862.72	62500.00	7500.00	225000.00	2882
Age	47.50	16.39	47.00	17.00	116.00	2933
Household Size	3.32	1.26	3.00	2.00	13.00	2926
White	0.78	0.41	1.00	0.00	1.00	2927
Black	0.07	0.25	0.00	0.00	1.00	2927
Hispanic	0.06	0.24	0.00	0.00	1.00	2927
Asian	0.06	0.23	0.00	0.00	1.00	2927
Full-time employment	0.49	0.50	0.00	0.00	1.00	2933
Part-time employment	0.08	0.28	0.00	0.00	1.00	2933
Unemployed	0.05	0.23	0.00	0.00	1.00	2933
Bachelor Degree	0.34	0.47	0.00	0.00	1.00	2933
Conservative	0.35	0.48	0.00	0.00	1.00	2797

Note: This table summarizes the main covariates of all respondents in experiment 4.

Table D13—Summary Statistics: Experiment 5 (Many task experiment)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.46	0.50	0.00	0.00	1.00	5045
Income	53034.84	26174.26	55000.00	5000.00	100000.00	4478
Age	37.27	11.72	34.00	17.00	88.00	5045
Household Size	3.68	1.44	3.00	2.00	13.00	5045
White	0.77	0.42	1.00	0.00	1.00	5045
Black	0.07	0.26	0.00	0.00	1.00	5045
Hispanic	0.05	0.21	0.00	0.00	1.00	5045
Asian	0.08	0.27	0.00	0.00	1.00	5045
Full-time employment	0.51	0.50	1.00	0.00	1.00	5045
Part-time employment	0.11	0.32	0.00	0.00	1.00	5045
Unemployed	0.13	0.34	0.00	0.00	1.00	5045
Bachelor Degree	0.37	0.48	0.00	0.00	1.00	5045
Conservative	0.26	0.44	0.00	0.00	1.00	5019
Number of HITs	8966.40	15468.91	2500.00	750.00	75000.00	5045

Note: This table summarizes the main covariates of all respondents in experiment 5.

Table D14—Summary Statistics: Experiment 6 (Effort Experiment: Weak demand)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.55	0.50	1.00	0.00	1.00	769
Income	32425.23	16975.09	35000.00	5000.00	85000.00	769
Age	37.44	11.73	35.00	21.00	70.00	769
Household Size	3.71	1.46	3.00	2.00	10.00	769
White	0.74	0.44	1.00	0.00	1.00	769
Black	0.08	0.28	0.00	0.00	1.00	769
Hispanic	0.06	0.24	0.00	0.00	1.00	769
Asian	0.08	0.27	0.00	0.00	1.00	769
Full-time employment	0.54	0.50	1.00	0.00	1.00	769
Part-time employment	0.11	0.31	0.00	0.00	1.00	769
Unemployed	0.12	0.33	0.00	0.00	1.00	769
Bachelor Degree	0.41	0.49	0.00	0.00	1.00	769
Conservative	0.30	0.46	0.00	0.00	1.00	764

Note: This table summarizes the main covariates of all respondents in experiment 6.

Table D15—Summary Statistics: Experiment 7 (Within Design)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.58	0.49	1.00	0.00	1.00	999
Income	54081.05	25778.96	55000.00	5000.00	100000.00	876
Age	34.45	10.73	31.00	19.00	83.00	999
Household Size	3.54	1.39	3.00	2.00	13.00	999
White	0.74	0.44	1.00	0.00	1.00	999
Black	0.08	0.27	0.00	0.00	1.00	999
Hispanic	0.06	0.23	0.00	0.00	1.00	999
Asian	0.10	0.30	0.00	0.00	1.00	999
Full-time employment	0.54	0.50	1.00	0.00	1.00	999
Part-time employment	0.11	0.32	0.00	0.00	1.00	999
Unemployed	0.13	0.34	0.00	0.00	1.00	999
Bachelor Degree	0.40	0.49	0.00	0.00	1.00	999
Conservative	0.24	0.43	0.00	0.00	1.00	994

Note: This table summarizes the main covariates of all respondents in experiment 7.

Table D16—Attrition overview by task in the strong demand experiments

	Finished: Time	Finished: Risk	Finished: Ambiguity Aversion	Finished: Effort 0 cent bonus	Finished: Effort 1 cent bonus	Finished: Lying	Finished: Dictator Game	Finished: Ult. Game 1	Finished: Ult. Game 2	Finished: Trust Game 1	Finished: Trust Game 2
Panel A: Unconditional Means											
Positive demand	1.000	1.000	0.995	0.972	0.968	0.995	1.000	1.000	0.995	0.990	1.000
	(0.000)	(0.000)	(0.005)	(0.010)	(0.011)	(0.005)	(0.000)	(0.000)	(0.005)	(0.007)	(0.000)
No demand	0.996	1.000		0.941	0.980		1.000				
	(0.004)	(0.000)		(0.015)	(0.009)		(0.000)				
Negative demand	0.992	0.996	1.000	0.984	0.970	1.000	0.996	1.000	0.990	1.000	1.000
	(0.006)	(0.004)	(0.000)	(0.008)	(0.011)	(0.000)	(0.004)	(0.000)	(0.007)	(0.000)	(0.000)
Panel B: Differential attrition											
Positive - Negative	0.008	0.004	-0.005	-0.012	-0.002	-0.005	0.004		0.005	-0.010	0.000
	(0.006)	(0.004)	(0.005)	(0.013)	(0.016)	(0.005)	(0.004)		(0.008)	(0.007)	(0.000)
Positive - Neutral	0.004			0.031	-0.012						
	(0.004)			(0.018)	(0.014)						
Negative - Neutral	-0.004	-0.004		0.043	-0.009		-0.004				
	(0.007)	(0.004)		(0.017)	(0.014)		(0.004)				
Observations	730	729	405	757	734	366	771	409	424	384	371

Note: In Panel A we present the proportion of respondents who completed the experiment in the positive, negative and no-demand treatment arms respectively. In Panel B we assess whether there was differential attrition across treatment arms by examining differences in completion rates across demand treatment arms.

Table D17—Attrition overview by task in the weak demand experiments

	Finished: Time	Finished: Risk	Finished: Ambiguity Aversion	Finished: Effort 0 cent bonus	Finished: Effort 1 cent bonus	Finished: Lying	Finished: Dictator Game	Finished: Ult. Game 1	Finished: Ult. Game 2	Finished: Trust Game 1	Finished: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.995	1.000	0.990	0.955	0.941	0.995	0.996	1.000	1.000	1.000	1.000
	(0.005)	(0.000)	(0.007)	(0.015)	(0.017)	(0.005)	(0.004)	(0.000)	(0.000)	(0.000)	(0.000)
No demand		0.993					0.992				
		(0.005)					(0.006)				
Negative demand	0.991	0.992	0.995	0.965	0.960	1.000	0.993	0.995	1.000	1.000	0.985
	(0.007)	(0.006)	(0.005)	(0.013)	(0.014)	(0.000)	(0.005)	(0.005)	(0.000)	(0.000)	(0.009)
Panel B: Differential attrition											
Positive - Negative	0.004	0.008	-0.005	-0.010	-0.019	-0.005	0.004	0.005		0.000	0.015
	(0.008)	(0.006)	(0.009)	(0.019)	(0.022)	(0.005)	(0.007)	(0.005)		(0.000)	(0.009)
Positive - Neutral		0.007					0.004				
		(0.005)					(0.007)				
Negative - Neutral		-0.001					0.001				
		(0.008)					(0.008)				
Observations	425	743	393	404	401	413	763	361	411	352	349

Note: In Panel A we present the proportion of respondents who completed the experiment in the positive, negative and no-demand treatment arms respectively. In Panel B we assess whether there was differential attrition across treatment arms by examining differences in completion rates across demand treatment arms.

E. Citations for Experimental Tasks

Our respondents complete one of the following tasks: a dictator game (Kahneman, Knetsch and Thaler, 1986); a risky investment game (Gneezy and Potters, 1997), without or with ambiguity; a convex time budget task (Andreoni and Sprenger, 2012); a trust game (first or second mover, Berg, Dickhaut and McCabe, 1995); an ultimatum game (first or second mover, Güth, Schmittberger and Schwarze, 1982); a lying game (Fischbacher and Föllmi-Heusi, 2013); and a real effort task with or without performance pay (DellaVigna and Pope, 2017, DellaVigna and Pope, 2018).

*

REFERENCES

Andreoni, James, and Charles Sprenger. 2012. "Estimating Time Preferences from Convex Budgets." *The American Economic Review*, 102(7): 3333–56.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior*, 10(1): 122–142.

DellaVigna, Stefano, and Devin Pope. 2017. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy, forthcoming.*

DellaVigna, Stefano, and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies*, 85(2): 1029–1069.

Fischbacher, Urs, and Franziska Föllmi-Heusi. 2013. "Lies in Disguise—an Experimental Study on Cheating." *Journal of the European Economic Association*, 11(3): 525–547.

Gneezy, Uri, and Jan Potters. 1997. "An Experiment on Risk Taking and Evaluation Periods." *The Quarterly Journal of Economics*, 631–645.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior & Organization*, 3(4): 367–388.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler. 1986. "Fairness and the Assumptions of Economics." *Journal of Business*, S285–S300.