NBER WORKING PAPER SERIES

MEASURING AND BOUNDING EXPERIMENTER DEMAND

Jonathan de Quidt
Johannes Haushofer
Christopher Roth

Working Paper 23470
http://www.nber.org/papers/w23470

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2017

We are grateful to Johannes Abeler, Stefano Caria, Rachel Cassidy, Tom Cunningham, Elwyn Davies, Stefano DellaVigna, Thomas Graeber, Don Green, Alexis Grigorieff, Johannes Hermle, 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 would like to thank seminar participants at Bergen, Berlin, Busara, LSE, Oxford, Stockholm, 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 views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measuring and Bounding Experimenter Demand
Jonathan de Quidt, Johannes Haushofer, and Christopher Roth
NBER Working Paper No. 23470
June 2017
JEL No. B41,C91,C92

ABSTRACT

We propose a technique for assessing robustness of behavioral measures and treatment effects to experimenter demand effects. The premise is that by deliberately inducing demand in a structured way we can measure its influence and construct plausible bounds on demand-free behavior. We provide formal restrictions on choice that validate our method, and a Bayesian model that microfounds them. Seven pre-registered experiments with eleven canonical laboratory games and around 19,000 participants demonstrate the technique. We also illustrate how demand sensitivity varies by task, participant pool, gender, real versus hypothetical incentives, and participant attentiveness, and provide both reduced-form and structural analyses of demand effects.

Jonathan de Quidt
Institute for International Economic Studies
Stockholm University
106 91 Stockholm
Sweden
jonathan.dequidt@iies.su.se

Johannes Haushofer
Woodrow Wilson School
Princeton University
427 Peretsman-Scully Hall
Princeton, NJ 08540
and Busara Center for Behavioral Economics,
Nairobi, Kenya
and also NBER
haushofer@princeton.edu

Christopher Roth
Keble College Oxford
Oxford OX1 3PG
christopher.roth@economics.ox.ac.uk

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/1248

An online appendix is available at
http://www.nber.org/data-appendix/w23470

1 Introduction

A basic concern in experimental work with human participants is that, knowing that she is being experimented on, the participant may change her behavior. Specifically, participants may try to infer the experimenter's objective from their treatment, and then act accordingly (Rosenthal, 1966; Zizzo, 2010). For instance, participants who believe the experimenter 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 experimental 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, and shroud the interpretation of treatment effects.

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 an effort task, we tell some participants "you will do us a favor if you work more than you normally would," and others "you will do us a favor if you work less than you normally would." 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 that maximizes the participant's utility as derived from the experimental payoffs. However, the participant is also motivated to take actions he perceives will "please" the experimenter by conforming to her research objectives. He tries to infer those objectives from the design features and distorts his action accordingly, biasing the experimenter's estimates. Our demand treatments attempt to manipulate those beliefs to identify an interval that contains the natural action. We remain agnostic about *why* the participant wishes to please the experimenter; motives may include altruism, a desire to conform, a misguided attempt to contribute

to science, or an expectation of reciprocity from the experimenter (Orne, 1962).

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 games and preference measures.[1] In each game we employ positive and negative "demand treatments" which tell participants that they will "do us a favor" if they choose a higher or lower action than they normally would.

Responses to these demand treatments are substantial and vary across tasks. The difference in average (standardized) behavior between our positive and negative demand treatment groups ranges from approximately 0.25 standard deviations for incentivized real effort to 1 standard deviation for trust game second movers. In an application to treatment effect estimation, we also derive bounds on the real effort response to performance pay. The bounds we obtain exclude zero, but are quite wide, ranging from a 0.25 to a 1.35 standard deviation increase in effort. Overall, the results suggest significant potential for bias due to demand effects.

Nevertheless, it is reasonable to think that explicitly asking subjects to "do us a favor" is a strong manipulation relative to realistic demand effects in typical experiments. Theory implies a trade-off in applying our method. "Strong" demand manipulations provide reliable but wide bounds on demand-free behavior, because they shift participants' beliefs a lot even if the underlying influence of demand is small. Many researchers will be satisfied with weaker, less conservative manipulations that yield tighter intervals. In another series of experiments, we employ demand treatments in which we simply signal an experimental hypothesis to participants, without explicitly demanding that they conform to it. Specifically, we tell them "we expect that participants who are shown these instructions will [work, invest, ...] more/less than they normally would." We find a much more moderate response to these treatments, and consequently obtain much narrower bounds, ranging from around 0.1 to 0.3 standard deviations.

[1]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). Our data come from US-based Amazon Mechanical Turk (MTurk) participants and a US nationally representative online panel.

Our approach invokes a weak form of the "no defiers" assumption familiar from estimation of local average treatment effects (Angrist and Imbens, 1994). Specifically, we require that when we demand a higher action not "too many" participants choose a strictly lower action, and vice versa. We provide evidence supporting this assumption with a within-participant manipulation in which we measure the same participants' actions first without, then with our demand treatments. Under slightly stronger assumptions, this design enables us to classify participants as compliers or defiers. We find that only around 5 percent of our participants are strict defiers that respond in the opposite direction to our treatments. We discuss how such within designs can be used to extract additional information about natural actions, correct the bounds for defier behavior, and reduce the cost of applying our method.

Next, following the basic approach of DellaVigna and Pope (2016), we illustrate how our demand treatments can be used in conjunction with a structural model to obtain unconfounded estimates of structural parameters, measure the representative participant's value of conforming to the experimenter's wishes, and predict demand-free choices. We estimate for the effort task that the value of pleasing the experimenter is equivalent to increasing the monetary incentives offered by around 20 percent.

Finally, we examine several moderators of sensitivity to experimenter demand. First, we find that women respond more to our demand treatments than men. Second, surprisingly, we find little evidence that sensitivity to our demand manipulations varies between incentivized and hypothetical choice. Third, we find some evidence that more attentive respondents responded more strongly. Fourth, we compare behavior between two participant pools—Amazon MTurk workers and a US representative online panel—and find very similar responsiveness.

We contribute to the small literature discussing experimenter demand effects (Shmaya and Yariv, 2016; Zizzo, 2010), demand characteristics (Orne, 1962), and obedience to the experimenter (Milgram, 1963). List (2007) and Bardsley (2008) argue that behavior in the dictator game is to a large degree an artefact of the experimental situation, showing that adding the option to take money in a dictator game dramatically reduces giving.

We also contribute to the literature which examines the effects of anonymity

on social behavior in the laboratory (Barmettler et al., 2012; Levitt and List, 2007; Hoffman et al., 1996, 1994). Barmettler et al. (2012) find no evidence that experimenter-participant anonymity affects behavior in standard social preference measures, while other studies document that non-anonymity in the lab can increase pro-social behavior (List et al., 2004). We also relate to work that explores the principal-agent relationship between experimenter and participant (Chassang et al., 2012; Shmaya and Yariv, 2016).

Third, our paper contributes to the literature discussing whether lab behavior generalizes to the field (Levitt and List, 2007; Harrison and List, 2004). List (2006) finds that behavior in the lab environment can be at odds with behavior in the field, which could be due to differences in demand effects in the lab setting.

Fourth, we relate to the growing literature on the effects of social pressure on economic, and social behavior, for example, charitable giving (DellaVigna et al., 2012) and voting (DellaVigna et al., 2017; Gerber et al., 2008).[2]

The paper proceeds as follows: in Section 2 we set up a simple theoretical model of experimenter demand that motivates our approach. In Section 3 we describe the data and our experimental design. In Section 4, we present the main experimental results and structural estimates. In Section 5 we examine heterogeneous effects. Section 6 concludes. An extensive set of web appendices provides additional results and theory.

2 Theory

We model a decision-maker (he) who has preferences over the outcomes induced by his action $a \in \mathbb{R}$ in an experiment. a can be continuous or discrete, but for simplicity we focus on the case of continuous actions with a natural ordering (more/less effort, investment, giving). The analysis extends naturally to the case where a is the probability of a binary choice. While throughout the analysis we treat a as the choice of a representative agent, it is straightforward to reinterpret as a population or group mean

[2]This literature is also linked to work on moral suasion and pro-social behavior (Dal Bó and Dal Bó, 2014).

action, and our conditions as applying to average actions.

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. Given ζ, the optimal "natural" (experimenter-absent) action is $a(\zeta)$.

The experimenter (she) is interested in either i) measuring a specific action $a(\zeta)$ (e.g., the level of giving out of an endowment), or ii) 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. Knowing that he is a participant in an experiment, the decision-maker changes his action according to his belief about the experimenter's wishes or objectives. 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. The logic of a randomized experiment is to induce orthogonal variation in a treatment so as to estimate its influence purged of confounds. If the influence of latent demand is orthogonal to the treatment, the treatment effect is not biased. To see this, note that the treatment effect can be decomposed as follows:

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

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.

Example 1. Consider two variants on the classic Dictator game, in which the participant is told to choose what fraction of \$10 to give to another

participant. In variant 0, she 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). In both scenarios, absent any motive for pleasing the experimenter she would prefer to give \$4, so the true treatment effect is $a(\zeta_1) - a(\zeta_0) = \$0$. However, in variant 0 she infers that the experimenter wants her to be generous, so she gives \$5, while in variant 1 she infers that the experimenter wants her to be selfish, so she gives zero. Then $a^L(\zeta_0) - a(\zeta_0) = \$1$ and $a^L(\zeta_1) - a(\zeta_1) = -\$4$, so $a^L(\zeta_1) - a^L(\zeta_0) = -\$5$ and we spuriously identify a treatment effect that is in reality a demand effect.

2.1 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 $a^+(\zeta)$. For illustrative purposes we assume for now that there exists just one type of positive and negative demand treatment, and discuss treatments that differ in intensity below.

Our first substantive assumption is a basic monotonicity condition:

Assumption 1 (*Monotone demand treatment effects*)**.** $a^-(z) \leq a^L(z) \leq a^+(z)$, $\forall z \in Z$.

Assumption 1 requires that a deliberate attempt by the experimenter to demand an increase in the action does not decrease it, and vice versa. It has a natural counterpart in the literature on local average treatment effects (Angrist and Imbens, 1994): the assumption rules out "defier" behavior whereby participants demanded to increase their actions, decrease them, and vice versa. While this is a strong assumption, we note that we only require it to hold for *average* actions, which is weaker than the standard no-defiers assumption. Moreover, Assumption 1 is testable (for average behavior) at $t = \zeta$ because we can test whether a^-, a^L and a^+ are correctly ordered. We perform this test for some of our applications, discussed below.

Our next assumption is central to our bounding exercises, and simply

amounts to assuming that the demand treatments are capable of bounding the natural action of interest:

Assumption 2 (*Bounding*)**.** $a^-(\zeta) \leq a(\zeta) \leq a^+(\zeta)$.

Assumption 2 allows us to bound estimates of mean actions and treatment effects. It implies the following:

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

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

This assumption therefore delivers the central result than we can use demand-inducing treatments to obtain bounds on mean actions (equation 2) and treatment effects (equation 3).

These bounds are the main objects of interest for our analysis, but 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. Now we provide an assumption that enables us to make such statements.

Definition (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: $S(\zeta) \geq \left|a^L(\zeta) - a(\zeta)\right|$.[3]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 under ζ_0 (or equivalently, our bounds on $a(\zeta_1)$ are wider than those on $a(\zeta_0)$).

However, as sensitivity only gives us an upper bound on the latent demand effect, it could be that the true latent demand effect is larger under ζ_0. Our third assumption, Monotone Sensitivity, allows us to make comparative statements about magnitudes.

[3]Proof: Assumption 1 gives $a^L(\zeta) \in [a^-(\zeta), a^+(\zeta)]$ while assumption 2 gives $a(\zeta) \in [a^-(\zeta), a^+(\zeta)]$. Taken together this implies $\left|a^L(\zeta) - a(\zeta)\right| \leq a^+(\zeta) - a^-(\zeta)$.

Definition (Comparison classes)**.** A comparison class $Z^C \subseteq Z$ is a set of environments for which 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 $z \in Z^C$.*

This assumption allows us to make comparative statements about latent demand effects between environments. Natural candidates for comparison classes include environments that differ only in a small number of attributes. For instance, an experimenter might be interested in testing whether demand effects are larger in one participant pool compared to another, or under one incentive scheme compared to another.[4] We derive some comparison classes below in the more structured context of our Bayesian model.[5]

Finally, we describe how demand treatments might be used to extract estimates of a true treatment effect of interest. If the researcher is willing to assume i) monotone sensitivity, and ii) that the latent demand effects under ζ_1 and ζ_0 have the same sign (i.e, $a^L(\zeta_1) - a(\zeta_1) \geq 0 \Leftrightarrow a^L(\zeta_0) - a(\zeta_0) \geq 0$), then, if the following difference-in-differences condition holds: $(a^+(\zeta_1) - a^-(\zeta_1)) - (a^+(\zeta_0) - a^-(\zeta_0)) = 0$, $a^L(\zeta_1) - a^L(\zeta_0)$ identifies the true treatment effect. Even if the condition does not hold, it can be used to sign the bias due to demand. Suppose the condition is positive, implying $S(\zeta_1) > S(\zeta_0)$. Monotone sensitivity then implies that $\left|a^L(\zeta_1) - a(\zeta_1)\right| > \left|a^L(\zeta_0) - a(\zeta_0)\right|$. Then, knowing the sign of the latent demand effect under ζ_1 enables the researcher to sign the bias due to experimenter demand: If

[4] A necessary condition for comparisons to be interesting is that actions are measured in the same units. Translating actions into a common scale (for example, "standardizing" the data to measure actions as multiples of the standard deviation) is one way to achieve this. An environment can belong to multiple comparison classes, and all comparison classes may be singletons.

[5] Some experimenters might be willing to assume monotone sensitivity without bounding. In that case the demand treatments can be informative about the scale of experimenter demand in environment ζ without bounding the true action (e.g. it could be that $S(\zeta)$ is some fixed proportion of $\left|a^L(\zeta) - a(\zeta)\right|$). Then, a natural use of $S(\zeta)$ is to provide information about "how bad" experimenter demand would have to be to threaten a certain interpretation of the data. She might compute objects like $m = \left|a^L(\zeta) - \alpha\right| / S(\zeta)$ in order to make statements like "latent demand would have to be m multiples of $S(\zeta)$ to be consistent with $a(\zeta) = \alpha$." Analogous approaches exist for bounding bias due to sample selection or violation of the exclusion restriction in IV, eg. Conley et al. (2012), Nevo and Rosen (2012) and Altonji et al. (2005).

$a^L(\zeta_1) - a(\zeta_1) > 0$ the experiment overestimates the treatment effect of interest, if $a^L(\zeta_1) - a(\zeta_1) < 0$ it underestimates it.

2.2 Bayesian model

In this section we provide a simple model of a decision-maker who is subject to demand effects, so as to provide intuition for our main assumptions and precise 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, over which the decision-maker has preferences. We compactly describe his payoff from action a in environment ζ by $v(a, \zeta)$ where v captures both the payoff structure (mapping from actions to outcomes) and preferences (mapping from outcomes to utility). We assume that v is strictly concave and differentiable, so the natural action $a(\zeta)$ solves $v_1(a(\zeta), \zeta) = 0$.

Example 2. Effort provision: A risk-neutral decision-maker chooses effort a. Effort is rewarded by a piece-rate ζ and the cost of effort is $c(a)$. Then $v(a, \zeta) = \zeta a - c(a)$ and $a(\zeta) = c'^{-1}(\zeta)$.

2.2.1 Latent demand

Demand enters preferences as follows. Upon observing the experiment and treatment, the decision-maker makes an inference about an unobservable parameter, $h \in \{-1, 1\}$. If $h = -1$, he believes the experimenter benefits from him taking low actions, while if $h = 1$ he believes she benefits from high actions.[6] His preference for pleasing the experimenter is captured by a preference parameter ϕ, which we allow to depend upon ζ, having in mind that ϕ might depend on the identity of the experimenter (e.g. the decision-maker might have different attitudes toward a researcher and a firm) or decision-maker (e.g. women might have different attitudes than men). $\phi(\zeta)$ might also vary with other environment features such as the salience of the potential benefit to the experimenter or how important the

[6]We think of the decision-maker perceiving the experimenter's preference over his actions directly, preferring actions toward one or other extreme, rather than the experimental outcomes outcomes induced by those actions.

participant believes his actions are in achieving the experimenter's objectives. 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. See e.g. Orne (1962) for discussion.

We assume utility is separable in payoffs and demand, in the following form:

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

where $E[h|\zeta] = Pr(h=1|\zeta)\times 1 + Pr(h=-1|\zeta)\times(-1) = 2Pr(h=1|\zeta) - 1$. The optimal action $a^L(\zeta)$ 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 a) the decision-maker assigns equal likelihood to the preferred action being high or low ($E[h|\zeta] = 0$), or when he does not care about the experimenter's objectives ($\phi = 0$).

We assume that the decision-maker's prior over h is $h_0 = 0$, so in the absence of any new information about h he chooses $a(\zeta)$. The relation between actions and beliefs is captured by $da^L(\zeta)/dE[h|\zeta] = -\phi/v''(a,\zeta)$, which has the same sign as ϕ.

We model the decision-maker's learning about h as follows. The environment ζ includes a signal $h^L(\zeta) \in \{-1,1\}$ which the decision-maker believes is a sufficient statistic for h, i.e. it contains all of the information in ζ about h, so $E[h|\zeta] = E[h|h^L(\zeta)]$. 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 negative or positive one with equal probability). We impose that $p^L(\zeta) \in [0,1)$, so the latent demand signal can never be perfectly informative.[7] It is straightforward to see (and we show in the Appendix) that:

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

[7]This assumption avoids the possibility that both latent demand and the demand treatment are seen as perfectly informative, but contradictory.

The decision-maker's belief depends on the experimental treatment in two ways. First, the treatment determines the sign of $h^L(\zeta)$, i.e. whether it induces a belief that the experimenter wants a high or low action. This determines the *direction* of the latent demand effect. Second, the *strength* of the latent demand effect depends on $p^L(\zeta)$, which measures the informativeness of the signal to be about h. [8]

In an effort to reduce demand confounds it is common practice to randomize treatments between participants and ensure participants are not informed about treatments that they are not exposed to. Intuitively this makes it harder for them to form conjectures about the experimental objective, thereby reducing the strength of learning and reducing its correlation with the treatment variation. In our notation, information about other treatments enters into ζ and would affect both h^L and p^L.

2.2.2 Demand treatments

We now assume that the experimenter has the option to send a "demand treatment" signal $h^T \in \{-1, 1, \emptyset\}$, which is either positive ($h^T = 1$), negative ($h^T = -1$), or no signal ($h^T = \emptyset$). These signals deliberately direct the decision-maker toward a high or low action by changing his belief about h. We assume that if the experimenter does not send a signal (no demand treatment), the decision-maker does not update his belief about h, i.e. he does not draw any inference from the absence of a demand treatment. This assumption is reasonable as at present demand treatments are rarely used in experiments.

We maintain throughout that ζ (and hence $h^L(\zeta)$, $p^L(\zeta)$ 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. 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

[8] As an aside, we note that the setup nests "natural field experiments" (Harrison and List, 2004) that induce no demand because the participant does not know they are in an experiment.

the selection of appropriate demand treatments.[9] If this does not hold then bounding may fail because the demand treatments alter the natural action itself: $a(\zeta(\emptyset)) \notin [a(\zeta(-1)), a(\zeta(1))]$. In section 2.3.6 and the appendix we extend the model to allow ϕ to depend upon h^T (i.e. the demand treatments change the motive for pleasing the experimenter) and show that the bounding condition remains unchanged.

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 negative and positive one with equal probability. η and ϵ are believed to be independent (we return to this assumption below). Based on h^T, the decision-maker updates his beliefs about h. We show in the Appendix that the Bayesian posterior is:

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

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.

2.2.3 Assumptions

We now provide the formal connection from the Bayesian model to the assumptions outlined in Section 2.1.

First, Assumption 1 (monotone demand treatment effects) states that a positive demand treatment increases the action (relative to no demand treatment) and the negative demand treatment decreases it. In the Bayesian model the conditions for this relationship to hold are $\phi(E[h|h^T = 1, h^L(\zeta)] - E[h|h^L]) \geq 0$ and $\phi(E[h|h^T = -1, h^L(\zeta)] - E[h|h^L]) \leq 0$. It is straightforward to see (and we show in the Appendix) that except for the trivial case $p^T = 0$, these conditions are satisfied if and only if $\phi \geq 0$, i.e. the participant has a weak preference for pleasing the experimenter (and therefore does not "defy" the perceived demand).

Proposition 1. *Assumption 1 (monotone demand treatment effects) holds for all p^T if and only if $\phi \geq 0$.*

[9] For example our "do us a favor" treatments may be unsuited to an experiment studying altruism toward the experimenter.

Second, Assumption 2 (bounding) states that the demand treatments provide bounds on the true action. In the Bayesian model, the action is larger or smaller than $a(\zeta)$ when $\phi E[h|h^T, h^L] \gtreqless 0$. Given that $\phi \geq 0$ by Assumption 1, we need $E[h|h^T = 1, h^L] \geq 0$ and $E[h|h^T = -1, h^L] \leq 0$. This is obviously 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.

Proposition 2. *Assumption 2 (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. Since latent demand and sensitivity can vary for multiple reasons, there is no simple condition that guarantees when this assumption will and will not hold. In Appendix C.4 we work out the following cases:

1. 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.

2. We analyze monotone sensitivity when variation in demand effects is driven by differences in the utility function, v, deriving specific conditions when v is additively or multiplicatively separable and providing examples.

3. We show that monotone sensitivity holds for variation driven by inattention to experimenter demand.

4. We show that monotone sensitivity does *not* hold in general for variation driven by differences in beliefs.

We use these findings when interpreting our results on heterogeneity.

2.3 Discussion and extensions

2.3.1 "Strong" and "weak" demand treatments

As shown above, the bounding assumption holds when $p^T \geq p^L(\zeta)$. Thus far we have assumed that there is only one demand treatment (with a positive and negative variant), but in reality there are many different ways to signal a desire for high or low actions. How should the experimenter choose?

Observe that the width of the bounds $[a^-(\zeta), a^+(\zeta)]$ is increasing in p^T.[10] Therefore the tightest bounds, subject to satisfying the bounding condition, are obtained when $p^T = p^L(\zeta)$. In other words, we want the "least informative" demand treatment, conditional on it being more informative than the latent signal. Thus, there exists a trade-off between the informativeness of the demand signal and the tightness of bounds: one can use demand treatments that strongly signal the experimenter's objective, giving confidence in the (wide) bounds obtained, or use more subtle manipulations to obtain tighter bounds, at the risk of failing to bound the true action or treatment effect.

In our applications, we use two demand treatments. The first, "strong" treatment explicitly tells participants what we wish: "You will do us a favor if you [...] more/less than you normally would." The second, "weak" treatment hints at a hypothesis, but does not explicitly tell the participant what we want them to do: "We expect that participants who are shown these instructions will [...] more/less than they normally would." We view the first treatment as being more informative about the experimenter's wishes (i.e., carrying a higher p^T) than the second, therefore generating more reliable but wider bounds.

2.3.2 Fewer demand treatments

It may not always be necessary to construct two-sided bounds. One such case is when the researcher has a strong prior about the direction of latent demand. For example, if they believe $a^L(\zeta) < a(\zeta)$ they might use

[10] Proof: $dE[h|h^T, h^L]/dp^T = h^T \left(1 - h^{L2} p^{L2}\right) \left(1 + h^L p^L h^T p^T\right)^{-2}$, which has the same sign as h^T, so a^+ is increasing and a^- decreasing in p^T.

only a positive demand treatment and construct bounds $[a^L(\zeta), a^+(\zeta)]$. Alternatively, they may only be interested in one bound. If they only wish to obtain a lower bound on a treatment effect $a(\zeta_1) - a(\zeta_0)$, they might measure only $a^+(\zeta_0)$ and $a^-(\zeta_1)$.

2.3.3 Discrete actions

The analysis easily extends to discrete and possibly un-ordered actions, such as accepting/rejecting a contract, selecting a lottery from a choice list or choosing bundles from a menu. Typically the experimenter will then be interested in the probability that a given option is chosen, and may be concerned that this probability is influenced by participants' beliefs about what the experimenter *wants* them to choose. Demand treatments can be readily constructed to manipulate those beliefs and obtain bounds, for instance telling participants "you will do us a favor if you (do not) choose option j."

2.3.4 Heterogeneity

Thus far we assumed a representative agent and made assumptions about his behavior. However, the approach naturally extends to the case where participants are heterogeneous and the experimenter is interested in average behavior or average treatment effects. If our non-parametric assumptions 1 and 2 hold for all agents individually, then we can simply reinterpret the natural action a and observed actions a^L, a^+ and a^- as representing mean behaviors and our approach remains valid. Intuitively, if we can bound all individuals' natural actions, then we also bound the mean of those actions.

An important source of heterogeneity is in participants' beliefs about the experimenter's wishes, whereby $E_i[h|h^L]$ takes on different values for different individuals i. This could be because $p_i^L(\zeta)$ (perceived precision of the signal) varies across individuals, $h_i^L(\zeta)$ (perceived direction of demand), or both. However, since the bounding condition depends only on $p^T \geq p^L(\zeta)$ and not the direction h^L, provided the inequality is satisfied each individual's natural action will be bounded by a^- and a^+ and the average natural action is bounded. A simple sufficient condition that guarantees bounding is $p^T \geq \max_i p_i^L(\zeta)$.

2.3.5 Defiers

Our monotone demand treatment effects assumption requires that there are no "defiers" who do the opposite of what they believe the experimenter wishes. In the model, such individuals possess a $\phi < 1$. Bounding then fails for these individuals, because $a^- \geq a^+$. However, we may still be able to bound population average behavior if the number of defiers and their responsiveness are relatively small.

To illustrate, we focus on the special case where preferences v and beliefs $E[h|h^L]$, $E[h|h^T, h^L]$ are identical for all individuals. In other words, all participants believe they know what the experimenter wants, but vary in their desire to conform. This means that the natural action is the same for all individuals and given by the solution to $v_1(a(\zeta), \zeta) = 0$. We label the beliefs H^L, H^+ and H^-. We further restrict preferences to be quadratic in actions, so $v_1(a, \zeta) = b - ca$ where b and c are constants that may depend on ζ. Normalizing c to equal 1, the natural action is equal to b for all individuals, while the action when beliefs equal H is $b + \phi H$. We therefore have:

$$Ea^L = b + H^L E\phi_i \qquad Ea^+ = b + H^+ E\phi_i \qquad Ea^- = b + H^- E\phi_i$$

Monotone demand treatment effects is satisfied on average if $Ee^L \in [Ea^-, Ea^+]$, which holds if and only if $E\phi_i \geq 0$. Therefore, testing for monotone demand treatment effects is equivalent to testing whether ϕ is positive on average. Bounding will hold provided $H^+ E\phi_i \geq 0$ and $H^- E\phi_i \leq 0$, which follows if $p^T \geq p^L(\zeta)$ and $E\phi_i \geq 0$.

This case is simple because of the quadratic preferences assumption, which implies that compliers and defiers respond symmetrically in opposite directions. More generally we would require conditions on the joint distribution of preferences and beliefs such that that the response by the compliers "outweighs" that of the defiers.

2.3.6 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. In appendix C.5 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. We show that the key condition for bounding is still $p^T \geq p^L$, but that demand treatments that are perceived to signal more importance generate wider bounds.

2.3.7 Extension: 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." Such instructions can be naturally thought of as a demand treatment that attempts to demand participants choose the natural action, $a(\zeta)$. It is straightforward to analyze such treatments in our framework. In Appendix C.6 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 show that 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, therefore the posterior belief lies between 0 and $E[h|h^L]$.[11] However, because signaling $h^T = 0$ moves actions toward the natural action it can be informative about the *direction* of latent demand. We also show that 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.

In sum, demanding the natural action is not guaranteed to obtain bounds that contain the natural action, while a pair of sufficiently strong positive and negative demand treatments does.

[11] One interpretation of latent influences in this model is "implicit" influence – the participant is not fully aware of the influence of latent demand and therefore unable to fully ignore it.

2.4 Inference

The theory tells us how to measure bounds on actions and treatment effects. Since these objects can be estimated experimentally, we may also wish to perform inference on the bounds themselves, or on the underlying parameters contained by the bounds. In Appendix C.7 we show how to do this, following Imbens and Manski (2004). We also provide a Stata package that allows calculation of demand-robust confidence intervals for mean behavior and treatment effects.

3 Sample and experimental design

We conducted seven experiments in total to test the method and to provide estimates of demand sensitivity on a wide range of experimental tasks. We conduct all of our experiments online, primarily because the very large number of treatments would be infeasible to implement in the laboratory.[12] We purposefully designed the experiments to maximize comparability. For all experiments except the effort task, choice sets are similar in that they can be expressed as real numbers from 0 to 1; we pay the same show-up fee; use a similar subject pool, mode of collection (online experiments with MTurk and online panel), and response mode (sliders); and stake sizes are as similar as possible.[13]

In Table 7 we summarize the key design features of each experiment. We describe the sample, which games were used, which demand treatments we employed, and whether choices involved real stakes or were hypothetical. More details on the experimental designs and the exact experimental instructions can be found in the pre-analysis plans, as well as the experimental instructions in the online appendix.[14]

[12] One might surmise that demand effects are more difficult to induce in online settings, which provide greater anonymity than in-person lab settings. Our bounds might therefore be underestimates. We leave to future work to compare demand effect bounds in online and in-person lab settings.

[13] For the effort task, we replicated the design of DellaVigna and Pope (2016). The primary differences are a higher show-up fee and a different response mode (effort).

[14] The online appendix is available at nber.org.

3.1 Participant populations

We conducted six experiments with approximately 16,000 participants on Amazon Mechanical Turk (MTurk), and one experiment with 3,000 respondents using a representative online panel of the US population. MTurk is an online labor marketplace that is frequently used by academics to recruit participants for experiments. It is attractive because it offers researchers a large and diverse pool of workers that have been shown to be more attentive to instructions than college students (Hauser and Schwarz, 2016).

To participate in our MTurk experiment, people had to live in the U.S, have an overall approval rating of more than 95%, and have completed more than 500 tasks on MTurk. We excluded prior participants when recruiting for experiments 2, 3 and 7. However, in 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.[15]

Most workers on MTurk are experienced in taking surveys, which is a potential threat to the external validity of our results (Chandler et al., 2014). We therefore conducted one additional experiment using a representative online sample, whose participants are less experienced with social science experiments.[16] This sample of 3,000 respondents is representative of the US population in terms of region, age, income, and gender.

3.2 Preanalysis plans

Each experiment is described in a preanalysis plan (PAP) posted online prior to launch.[17] Each PAP outlines the analysis for an individual experiment. For brevity and ease of exposition, in this paper, we pool the data, i.e. present and compare all tasks side by side, rather than experiment by experiment. However, the analysis follows what was pre-specified, with a

[15] Our results are virtually unchanged by the exclusion of respondents that completed more than one of our experiments; results available upon request.

[16] We collect data on this sample through an online survey in collaboration with the market research company, *Research Now*. This provider has been used in previous research, for example by Almås et al. (2016).

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

few minor exceptions.[18] In addition, online appendix E presents all of the pre-specified analyses for each experiment.

3.3 Summary statistics

The experiments were run between May 2016 and May 2017. In Tables A.40 to A.46 in the online appendix, we provide details on the sample characteristics of our respondents from both MTurk and the representative sample. In addition, in Tables A.32 to A.38 in the online appendix, we present the pre-specified balance tables for all of these experiments. Table A.43 highlights that respondents from the online panel are representative of the US population by gender, income, age, and region, and in terms of other observables, such as education and race. Attrition was low, with less than 2 percent of participants dropping out of our experiments on average. Importantly, there was no differential attrition across the different demand treatment arms. Tables A.14 and A.15 in the online appendix summarize attrition rates at the game level for the strong and weak demand experiments, respectively.

4 Applying the method

4.1 Bounding natural actions

Our first set of experiments attempts to measure the upper and lower bound of behavior using our strong demand treatments, which explicitly tell participants they will "do us a favor" by taking a higher or lower action than normal. Our respondents complete one of the following games: a dictator game, an investment game (to measure risk preferences; (Gneezy and Potters, 1997)), convex time budgets (to measure time preferences; (Andreoni and Sprenger, 2012)), a trust game (first mover or second mover; (Berg et al., 1995)), an ultimatum game (first mover or second mover; (Güth et al., 1982)), a lying game (Fischbacher and Föllmi-Heusi, 2013), a measure of

[18]For example, in experiment 1 we prespecified three pairwise tests for differences in behavior between the dictator, risk and time tasks. Expanding to eleven tasks would entail 55 such tests, which seems excessive and which we therefore do not conduct. However, the results are available upon request.

ambiguity aversion based on the risk preference task, and a real effort task (DellaVigna and Pope, 2016).

As an example, in the dictator game, participants in the positive demand condition are given the following message: "You will do us a favor if you give more to the other participant than you normally would".[19] Participants in the negative demand condition receive the following message: "You will do us a favor if you give less to the other participant than you normally would." In Table 8, we describe the key design features and demand instructions for each game.

For a subset of games,[20] we also measure behavior in a no-demand condition in which participants receive no demand manipulation, to test Assumption 1, monotone demand treatment effects. In addition, in the dictator game, convex time budgets, and the investment game, half of our respondents made choices for real stakes, while half made hypothetical choices for the same stake sizes. For the remaining games, all choices are incentivized.

Table 1 and Figures 1 and 2 summarize the response to the strong demand treatments for each of the games, restricting the sample to MTurk respondents and incentivized choices. In Panel A of Table 1, we show the unconditional mean actions in the different demand treatment arms.[21] Our objects of interest are mean behavior in the positive demand condition, $a^+(\zeta)$, mean behavior in the negative demand condition, $a^-(\zeta)$, and mean behavior in the no-demand condition, $a^L(\zeta)$.

In Panel B of Table 1, we display our sensitivity measure $(a^+(\zeta)-a^-(\zeta))$ for each of the 11 games, in both raw and z-scored units.[22] Behavior is responsive to our strong demand treatments across tasks, and sensitivity is

[19]Our instructions are close to the ones used by Binmore et al. (1985): "You will be doing us a favor if you simply set out to maximize your winnings". Deutsch et al. (1967) employ a similar approach, telling participants "I want you to to earn as much money as you can regardless of how much the other earns". Such instructions have been criticized precisely because they risk of inducing experimenter demand (Thaler, 1988; Zizzo, 2010).

[20]The dictator game, convex time budgets, the investment game, and the two real effort tasks.

[21]For most of the games the action set lies between 0 and 1; if not we rescale the action space to the $[0, 1]$ interval.

[22]We z-score our outcome variables at the game level, using the mean and s.d. for the negative demand group (Kling et al., 2007).

significantly different from zero in all tasks. Sensitivity is particularly high in the dictator game, for second movers in the trust game and ultimatum games, and for unincentivized effort.

In Panel C of Table 1, we examine the monotone demand treatment effects assumption: $a^+(\zeta) \geq a^L(\zeta) \geq a^-(\zeta)$. We estimate the following equation, in which POS_i takes value one for people in the positive demand condition and value zero otherwise, and where NEG_i takes value one for people in the negative demand condition and value zero otherwise:

$$ZY_i = \pi_0 + \pi_1 POS_i + \pi_2 NEG_i + \varepsilon_i \tag{8}$$

As can be seen in Panel C of Table 1, we find support for the assumption. In all cases the positive demand treatment increased actions, and the negative treatment decreased them on average. In most cases the differences are statistically significant.

[Insert Table 1 and Figures 1 2]

4.2 Bounding treatment effects

In our real effort experiment, which replicates a subset of the treatments in DellaVigna and Pope (2016), participants earned points by alternately pressing two keyboard buttons for 10 minutes. In one treatment arm, respondents were told that their score "will not affect [their] payment," while in the second they received one cent per 100 points. For some participants we added our demand treatments to these instructions, telling workers "you will do us a favor if you work harder/less hard than you normally would." We can apply our method to estimate bounds on the treatment effect of performance pay on effort provision.[23]

Panel A in Figure 3 and Panel A in Table 2 summarize the sensitivity of effort to our strong demand treatments, and how this sensitivity depends on incentives. We find that individuals who receive no bonus are substantially more sensitive to our demand treatments.

[23]In addition to these main treatment arms, 250 additional individuals completed the "no demand condition" for a reward of 4 cents per 100 points. This treatment arm is used in the structural estimation below.

Panel B in Figure 3 and Panel B in Table 2 display 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 of the treatment effect ($a^-(1) - a^+(0)$). The bounds we estimate are quite wide, ranging from 0.25 to 1.35 standardized units. However, the lower bound (0.25) is statistically significantly different from zero. This implies that even if behavior in the unincentivized condition is biased by extreme negative latent demand (i.e. $a(0) = a^+(0)$), while behavior in the incentivized condition is strongly positively biased (i.e. $a(1) = a^-(1)$), we can still support the conclusion that incentives increase effort.

[Insert Table 2 and Figure 3]

4.3 Tighter bounds with weak demand treatments

While the bounds we obtain using the strong demand treatments satisfy the monotone demand treatment effects assumption, they are wide, meaning that we cannot rule out many possible values for $a(\zeta)$. We therefore conduct additional experiments to obtain less conservative bounds, at the cost of a less plausible bounding assumption.[24]

Our weak demand treatments take a similar form to the strong treatments, but rather than explicitly reporting an objective ("you will do us a favor"), we simply reveal an experimental hypothesis to the participants, without demanding that they act to confirm it. For example, in the investment game, participants were told that "We expect that participants who are shown these instructions will invest more/less in the project than they normally would," with the phrasing modified accordingly for other tasks. Table 9 describes design features and the demand instructions for each game.

As before, for a subset of games (the dictator and investment games), we collect data without the demand treatment to test for monotone demand treatment effects. In addition, for the dictator game and the investment

[24]We employed weak demand treatments in experiments 2, 4, 5 and 6. A more detailed description can be found in Table 7.

game, half of our respondents' choices are incentivized, and half hypothetical. For all other tasks all choices are incentivized.

We present results from the MTurk experiments with weak demand treatments and incentivized choices. In Panel A of Table 3 we display the unconditional means in the different demand treatment arms. In Figure 2 we plot these values with confidence intervals. Our objects of interest are mean behavior in the positive demand condition, $a^+(\zeta)$, mean behavior in the negative demand condition, $a^-(\zeta)$, and mean behavior in the no-demand condition, $a^L(\zeta)$.

Panel B of Table 3 and Figure 1 display the sensitivities by game, $(a^+(\zeta) - a^-(\zeta))$.[25] For convex time budgets, the effort tasks, the lying task, and the trust game first mover, we find sensitivities to weak demand below 0.10 standard deviations.[26] We find stronger responses (between 0.20–0.25 standard deviations) for the dictator game, the ultimatum game second mover, and the trust game second mover. Sensitivity in the investment game under risk and uncertainty, as well as for the ultimatum game first mover, is approximately 0.17 standard deviations.

In Panel C of Table 3 we examine the monotone demand treatment effect assumption, finding that most demand treatments have the correct sign and are significant. We estimate a small negative effect of the positive demand treatment in the investment game, and a small positive effect of the negative treatment in the dictator game, though neither estimate is statistically significant.

[Insert Table 3]

4.4 Confidence intervals

We compute confidence intervals for (a) the bounds themselves, and (b) for the parameters contained by those bounds (an action or treatment effect), following Imbens and Manski (2004). Confidence intervals for the parameter are slightly tighter than for the bounds, which reflects the fact

[25] We again z-score our outcome variables at the paradigm-incentive level, using the mean and s.d. for the negative demand group (Kling et al., 2007).

[26] Our minimum detectable effect sizes range from between 0.20 to .34 standard deviations.

that the parameter cannot lie at both bounds simultaneously.[27] In Table A.4 we present confidence intervals for all 11 games, and in Table A.5 we present confidence intervals for the treatment effect in the effort experiment. Details on how confidence intervals are computed are given in the online appendix C.7.

4.5 Structural estimates

Under additional parametric and identifying assumptions, our demand treatments permit structural estimation of demand-free model parameters (v), as well as ϕ and the latent demand beliefs. Knowing v allows the researcher to make predictions about behavior in demand-free environments. Knowing ϕ allows them to quantify the importance of experimenter demand relative to v. 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 demand effects can be easily incorporated in the model of DellaVigna and Pope (2016) (DP), we can exactly follow their approach to structural estimation.

In their experiment, 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, $c(a)$ is a cost of effort function, ζ is a piece rate, and s is an intrinsic motivation parameter — workers may work for no pay because they enjoy the task. DP solve the first order condition and estimate the model parameters using nonlinear least squares (NLLS).[28]

Adding demand to this utility function gives:

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

[27] But for a coverage correction derived by Imbens and Manski, the $1 - \alpha$ confidence interval on the parameter corresponds to the $1 - 2\alpha$ interval on the bounds.

[28] They also employ a minimum distance estimation procedure. We stick to NLLS for brevity.

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:

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

Second, an exponential form $c(a) = k\exp(\gamma a)/\gamma$, with corresponding 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.[29] However, we have 13 parameters in total: s, k, γ, $\phi(0)$, $\phi(1)$, $\phi(4)$, $p^L(0)$, $p^L(1)$, $p^L(4)$, $h^L(0)$, $h^L(1)$, $h^L(4)$, and p^T.[30] We therefore need to impose some further restrictions.

First we assume that ϕ is fixed: $\phi(0) = \phi(1) = \phi(4) = \phi$. In other words, varying incentives do not change the participants' desire to please the experimenter. Second, we assume $p^T = 1$. By assumption this is not justified for our weak demand treatments, so we focus on the strong treatments, where the assumption is more plausible. This assumption implies that $E[h|h^T, h^L] = h^T$. We are therefore able to identify ϕ, s, γ, and k just using the four demand treatment groups. Third, since $E[h|h^L(\zeta)] = p^L(\zeta)h^L(\zeta) \in [-1, 1]$, we can treat it as a single parameter whose sign identifies h^L and whose magnitude identifies $p^L(\zeta)$. 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 mone-

[29] 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 as we explain below, for estimation we need to impose the parameter restriction $p^T = 1$, which we do not believe is satisfied in the weak treatments.

[30] In principle p^T might also vary with ζ and h^T. Our model presented above rules this out by assumption.

tary incentives are present, i.e. $p^L(1)h^L(1) = p^L(4)h^L(4)$, in which case we are overidentified.

While we estimate the same model as DP, the 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 treatments. When latent demand is present, the model parameters are identified from the demand treatments, and the neutral treatments identify the latent demand beliefs. This also means that for our purposes the neutral treatment with four cent incentives is not necessary for identification of the core parameters.

We follow DP in estimating equation (12) in logs and equation (13) in levels, using nonlinear least squares.[31] Estimation results are presented in Table 4. Columns 1–3 correspond to the power cost function and columns 4–6 to the exponential cost function. In each case we first mirror DP by estimating s, γ, and k using only the neutral treatments, assuming that there is no latent demand.[32] Second, we include all treatment groups and impose that latent demand depends only on whether monetary incentives are present. 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 demand playing an equivalent role to an incentive of 1 cent per 100 points, while $\phi = 1$ means that a worker who is certain the experimenter wants high effort works as if her incentives were increased by 1 cent per 100 points, relative to someone who does not know the experimenter's wishes.

Our first finding is an important potential role for experimenter demand. Our estimates of s and ϕ are quite large and of similar magnitude in all specifications, taking values equivalent to 0.2–0.5 cents per 100 points. Assuming a value of ϕ of around 0.25 would imply that switching from extreme negative to extreme positive demand (a change in $E[h|h^L]$ from -1 to 1) increases effort by as much as increasing the incentive by 0.5 cents per

[31]Since the piece rates are per 100 points, we follow DP in rounding scores to the nearest 100. See the online appendix D for further details of the estimation.

[32]The parameter estimates we obtain are quite different from those of DP. One possible explanation is that DP estimate their main specification from 0, 1, and 10 cent treatments, while we use 0, 1, and 4 cents, which may discipline the curvature of the effort cost function less. Additionally, while we like they recruited our participants on MTurk, our experiments were conducted some time after theirs, so the participant pool may have changed somewhat.

100 points. Our estimates of $E[h|h^L]$ are mostly negative, consistent with latent demand decreasing effort (though in the exponential cost case we estimate a positive value in the 1 cent treatment). However, the estimates are noisy and typically not significantly different from zero. We also estimate that in the 4 cent treatment, $E[h|h^L(4)] \approx -6.5$, which contradicts the theory (beliefs should lie in $[-1, 1]$), though we note that -1 lies well within the 95% confidence interval. We do not have demand treatment groups who were paid 4 cents, so the effort cost function is extrapolated out of sample to this group, which may help explain the poor fit of the model.

Our second finding is that allowing for demand can be quantitatively important for other parameter estimates. This is most noticeable when comparing the estimates of s when we do and do not allow for demand effects; the estimates are an order of magnitude smaller in the latter case (columns 1 and 4). Our estimates imply a negative latent demand effect which is instead attributed to low intrinsic motivation.

Finally, the structural estimates enable us to go beyond bounding to back out predicted demand-free behavior. To do this we plug our parameter estimates back into the first order conditions, fixing $E[h|h^L] = 0$. Results are presented in table A.3 in the online appendix. Since most of our estimated latent demand effects are negative, predicted demand-free effort is usually higher than observed effort, sometimes considerably so.

[Insert Table 4]

4.6 Measuring defiance

Valid bounds on average behavior require that there are not "too many" defiers: individuals who try to do the opposite of what they believe the experimenter wants. This is an identifying assumption in our basic approach which randomizes participants into different demand treatments, exposing them to either a positive, negative, or no demand treatment. Our seventh experiment is designed to assess the reasonableness of the assumption, by collecting within-participant data on behavior first without, and then with a demand treatment. 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. Note that in this sense, our setup differs from the "potential outcomes" framework, in which defier behavior is usually unobservable.

The structure of the experiment is as follows. Participants ($N = 1002$) on MTurk are told that they will complete two tasks and that they will be paid according to the choice made in one of them, selected by chance. Half play the dictator game twice, and the other half the investment game. They first complete the task without any demand treatment. Then, they complete the same task again, but with the addition of the strong positive or negative demand treatment. We thus have four groups, split by dictator/investment game and positive/negative demand treatment in task 2.

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 depending on ϕ. 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. We do however caution against over-interpreting the data, for two reasons. First, in the theory we assume that the environment ζ, and therefore the natural action, $a(\zeta)$, is independent of the demand treatment, h^T. This is a stronger assumption in our within design that reveals to participants that the response to h^T is itself part of the analysis, and could change their interpretation of ζ.[33] 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 either reveal or conceal their defier/complier identity.

Our main findings are presented in Figures 4 and 5, which show the distributions of the changes in behavior between tasks 1 and 2. Only about 5 percent of our respondents are strict defiers. About 30 percent do not change their behavior at all in response to our strong demand treatments,

[33]We suspect that this is not a serious concern in practice: participants presumably infer that our interest is in showing people respond to our demand treatments by changing their actions in task 2 relative to task 1. Compliers who wish to conform to this objective would then increase their action while defiers would decrease their action, thus we would arrive at the correct complier/defier classification.

while the remaining 65 percent respondents strictly comply with our demand treatments. The proportions are similar for the dictator and the investment game. These findings suggest that there is very little defiance in practice, in support of our method.

Table 10 mirrors our prior results, presenting mean actions and sensitivities estimated from the within design, alongside the equivalent objects from the earlier experiments (using incentivized choices from MTurk participants). Interestingly, the sensitivities estimated from the second stage of the within design are very similar in magnitude to those from the between design, and if anything slightly larger.[34] This suggests that the concerns about the within design outlined above may not be serious in practice. It also suggests that researchers may be able to simply and relatively cheaply obtain bounds using within-participant demand treatments, avoiding the need to recruit new subjects.

We also highlight several other potential uses of the within-participant data. First, they can be used to construct "defier-robust" bounds. For defiers, $a(\zeta) \in [a^+(\zeta), a^-(\zeta)]$, so if the proportion of compliers is c we can construct defier-robust bounds equal to $[cE[a^-(\zeta)|\phi \geq 0] + (1-c)E[a^+(\zeta)|\phi < 0], cE[a^+(\zeta)|\phi \geq 0] + (1-c)E[a^-(\zeta)|\phi < 0]]$. The "defier-robust" bounds as well as the standard bounds are displayed in Table A.2. Second, our assumption that $p^L(\zeta) < p^T \leq 1$ guarantees that those with $\phi \neq 0$ will change their action in response to a demand treatment. Thus, participants who do not respond at all reveal that $\phi = 0$, in which case $a^L(\zeta) = a(\zeta)$, i.e. we can identify the natural action for these participants. Third, researchers might be intrinsically interested in comparing the observable characteristics and behavior of compliers and defiers.

4.7 Beliefs

In each experiment, after participants had completed the relevant task, we collected simple, unincentivized belief data. Specifically, we asked two questions. First, we asked "What do you think is the result that the researchers of this study want to find?" and second, "What do you think

[34] We also analyze the behavior of compliers and defiers separately, in Table A.1 in the online appendix, comparing the average change in action between task 1 and 2 for each group separately.

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. The main purpose of these belief measures was a manipulation check, to ascertain that participants' beliefs responded as expected to the demand treatments.[35] Naturally, these measures may be subject to their own demand bias.

A natural interpretation of the belief responses is that participants report a high belief if their posterior $E[h|h^T, h^L]$ 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 A.10, A.11, A.12, and A.13 (in the online appendix). They confirm that our treatments moved average responses in the anticipated direction. For example, in the dictator game, 65% of the strong positive demand treatment group reported that the researchers "want to find that on average people give a large share of the $1 to the other person" (alternative: "small"). Under the strong negative demand treatment only 24% gave this response. The corresponding figures are 54% and 23% for the weak demand treatment.

Overall, the levels of beliefs and magnitudes of shifts in beliefs are similar for the strong and weak treatments.[36] This result implies that both treatments were successful in fixing the sign of participants' posteriors, and that therefore bounding holds for both treatments.

A final use of the belief data is as an alternative (not pre-specified) measure of attentiveness: we can classify as attentive those participants reporting the "correct" belief about the experimental objective in response to our demand treatments. A.5 shows that sensitivity to our strong demand treatments is high — around 1 standard deviation — among participants considered attentive by this measure. Sensitivity for inattentive partici-

[35]While in principle once could collect richer belief measures and incentivize responses, we opted for this simple approach because asking for fine-grained beliefs about *our own* motivations seemed quite unnatural, and because there was no objective truth against which to score. Techniques do exist for belief scoring without an objective truth, e.g. Prelec (2004).

[36]In some cases, the weak treatments shifted beliefs by more than the strong treatments, though note that not all strong and weak treatments were conducted in the same experiments.

pants is close to zero and never significant.

5 Heterogeneity

Does sensitivity to demand treatments vary with design and participant characteristics? In this section, we examine heterogeneous responses to our strong and weak demand treatments 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 (e.g., whether greater sensitivity among one gender reflects a greater influence of latent demand for that gender) depends upon whether Monotone Sensitivity holds for the subset of environments under consideration, i.e. whether they belong to the same comparison class. We show in Appendix C.4 that variation in incentives, attention, and the preference for pleasing the experimenter, ϕ (which may differ by gender or participant pool), plausibly form valid bases for comparison classes.

5.1 Incentivized vs. hypothetical choices

In the dictator game, investment game, and the convex time budgets that we conducted on MTurk, we randomly assigned participants to make either hypothetical or incentivized choices. In what follows, we test whether making an incentivized rather than a hypothetical choice affects participants' response to our strong demand treatments. To do so, we regress our standardized outcome variables, pooled across games, on a demand treatment indicator, POS_i, taking value one for people in the positive demand condition and value zero for people in the negative demand condition; an indicator M_i, taking value one for the incentive condition; and their interaction:[37]

$$ZY_i = \beta_0 + \beta_1 POS_i + \beta_2 M_i \times POS_i + \beta_3 M_i + \varepsilon_i \tag{14}$$

[37] We standardize the outcome variable at the game-incentive level.

Results are presented in Figure 4 and Panel B of Table 5. Interestingly, participants making hypothetical or incentivized choices responded very similarly to experimenter demand.[38] This result is somewhat surprising, since deviations from the natural action are presumably less costly in hypothetical choice. Possibly this finding reflects that even our incentivized choices involve relatively low stakes, and it is possible that we would see a difference at higher stakes. Our results relate to previous work examining the effects of incentives on behavior in the lab (Camerer et al., 1999).

However, we note that in the effort experiment we do see lower sensitivity when effort is incentivized than when unincentivized (note that unincentivized effort is conceptually different from hypothetical choice).[39]

[Insert Figure 6]

5.2 Gender and attention

We measure participant gender and attentiveness in all tasks. We define a participant as attentive if they passed an attention screener at the beginning of the task.[40] To examine heterogeneous responses to our strong demand treatments by these variables, we estimate the following equation:

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

where H_i is the dimension of heterogeneity of interest.

First, we test whether the normalized sensitivity to experimenter demand differs for men and women using a dummy, $Male_i$, taking value one for males.[41] As can be seen in Table 5 and Figure 7, which show data

[38] As we discuss in section 2, we can meaningfully compare effect sizes across the respondents making incentivized choices rather than hypothetical choices if the monotone sensitivity assumption holds. In Appendix C.4, we show that this is the case if utility is additively or multiplicatively separable in incentives.

[39] In the model in section 4.5, effort is additively separable in incentives, satisfying the condition for Monotone Sensitivity in Appendix C.4.

[40] The screener 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 ignore this possible interpretation and instead choose two specific options. The assumption is that attentive respondents read the question and follow the concealed instruction, while inattentive respondents do not.

[41] We normalize the outcome variable at the game level.

from incentivized MTurk respondents, we find that females respond more strongly to the strong demand treatments than males.[42,43]

[Insert Figure 7]

As we show in Appendix C.4, if the difference in sensitivity is driven by differences in willingness to please the experimenter, then it is indicative of stronger latent demand effects for females. However, males and females might also hold different beliefs about the experimental objective, in which case Monotone Sensitivity would not hold.[44]

We also examine whether respondents who did not pass an attention screener at the beginning of our experiments respond differently to our demand treatments from respondents who passed. Table 5 and Figure 8 show higher sensitivity among attentive MTurkers, but this effect is not significant and noisily measured as only 10 percent of our MTurk sample were inattentive.[45]

[Insert Figure 8 and Table 5]

However, in the representative online panel we have enough variation to examine heterogeneous effects by attention with sufficient statistical power. As can be seen in Table A.26, we find evidence that participants who passed the screener were significantly more sensitive to the demand treatments than those who did not. The estimated difference between attentive and inattentive respondents from the representative sample is quite similar to that of MTurkers. If we pool the data from MTurk and the representative online panel, we find significantly different sensitivities to our demand treatments by attention. As discussed in the online appendix C.4, variation in sensitivity generated by inattention to experimenter demand satisfies the

[42]Our finding is related to the large literature on gender differences in preferences (Croson and Gneezy, 2009).

[43]We find similar results for the representative online panel, as can be seen in Table A.7.

[44]Using the belief measure described in section 4.7, we do not find any evidence that males and females in the no-demand condition hold different beliefs, nor do they update their beliefs differently in response to the demand treatments.

[45]We normalize the outcome at the game level.

monotone sensitivity assumption.[46,47]

We also examine heterogeneous responses to the weak demand treatments, though we have less power to detect differences because of the lower overall sensitivity to these treatments. We find no significant differences in sensitivity by incentives, attention, gender, education or experience. These results are summarized in Tables 6 and A.9.

5.3 MTurk vs. representative online panel

Some experimental social scientists are concerned that MTurk workers are experienced research participants and may behave differently to a more representative participant pool. Moreover, MTurkers need to maintain a high "approval" rating and may therefore be especially motivated to please the researcher (Berinsky et al., 2014).[48]

To address such concerns, and to test an additional dimension of heterogeneity, we replicated the MTurk dictator game and investment game experiments with a representative online survey panel, whose participants are less experienced experimental participants. We randomly assigned respondents to either a positive weak demand treatment, a negative weak demand treatment, a positive strong demand treatment, a negative strong demand treatment or no demand treatment. All of our respondents' choices in this experiment were incentivized, and the stake size was the same as in the MTurk experiment.

In Table 5 and Figure 9, we test for differences in sensitivity in both the pooled dictator and investment games, and for each game separately. Representative panel participants responded more strongly in the risk game and less strongly in the dictator game (significant at 10%); the pooled test

[46]Consistent with our model, the belief data suggest that attentive respondents update their beliefs about the experimental objective more strongly in response to our demand treatments.

[47]We also examine heterogeneous sensitivities to our strong demand treatments along two additional dimensions that we had not pre-specified, education and prior experience with experiments measured by the number of HITs previously completed on MTurk. We find no evidence that more experienced or more educated respondents react more to our demand treatments. These results are summarized in Table A.8.

[48]Recruiters on MTurk have the option to reject unsatisfactory work, and recruiters can screen out workers with high rejection rates. However, we believe it is well-known that researchers rarely reject work; we never did.

finds a small and non-significant difference in sensitivity. Thus, MTurkers are not differentially susceptible to experimenter demand.[49]

6 Conclusion

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

We find that behavior in eleven canonical economic games is quite sensitive to our strong demand manipulations that explicitly signal an experimental objective, generating bounds of up to 1 standard deviation in width. Much tighter bounds are obtained using weak demand treatments in which we signal only an experimental hypothesis. We expect that the latter are a more realistic measure of the magnitude of demand effects in the typical experiments. Since our method is vulnerable to "defiers," individuals who respond in the opposite direction to our manipulations, we conduct an experiment to measure it, and find very low defiance rates of approximately 5 percent.

We also show how to analyze demand effects structurally, following the approach of DellaVigna and Pope (2016). Our estimates suggest a utility from pleasing the experimenter roughly equivalent to increasing the monetary incentives offered by 20 percent. We leverage the structural model to extract predictions for demand-free behavior, or "natural actions" in our terminology.

[49]Respondents from MTurk and the representative online panel might exhibit different willingness to please the experimenter, but they could also have different beliefs or attentiveness. In line with the above finding that respondents from the online panel are less attentive, they also updated their beliefs about the experimental objective less than MTurkers. However, focusing on the subsample of attentive respondents from MTurk and the representative online panel, we find that respondents from the representative online panel update their beliefs more. This suggests that some of the variation across groups is driven by differences in beliefs, and hence monotone sensitivity may not hold.

Finally, we find that females respond more to our strong demand treatments than males, but no significant heterogeneous sensitivity to demand by whether choices are incentivized, education, prior experience or the participant pool. We find some evidence that more attentive participants conform more to our strong demand treatments.

Future work might employ similar treatments to study how to decrease demand in experiments. Researchers could examine what features of the environment, the experimenter, and the mode of data collection (e.g. online vs. laboratory) influence demand effects.

References

Almås, Ingvild, Alexander Cappelen, and Bertil Tungodden, "Cutthroat Capitalism versus Cuddly Socialism: Are Americans more Meritocratic and Efficiency-seeking than Scandinavians?," *Working Paper*, 2016.

Altonji, Joseph G, Todd E Elder, and Christopher R Taber, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

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

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

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

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

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

Berinsky, Adam J, Michele F Margolis, and Michael W Sances, "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys," *American Journal of Political Science*, 2014, *58* (3), 739–753.

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

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

Camerer, Colin F, Robin M Hogarth, David V Budescu, and Catherine Eckel, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-production framework," *Journal of Risk and Uncertainty*, 1999, pp. 7–48.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci, "Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers," *Behavior Research Methods*, 2014, *46* (1), 112–130.

Chassang, Sylvain, Gerard Padró i Miquel, and Erik Snowberg, "Selective trials: A principal-agent Approach to Randomized Controlled Experiments," *The American Economic Review*, 2012, *102* (4), 1279–1309.

Conley, Timothy G, Christian B Hansen, and Peter E Rossi, "Plausibly Exogenous," *Review of Economics and Statistics*, 2012, *94* (1), 260–272.

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

DellaVigna, Stefano and Devin Pope, "What Motivates Effort? Evidence and Expert Forecasts," *National Bureau of Economic Research*, 2016.

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

_ , _ , _ , and Gautam Rao, "Voting to Tell Others," *Review of Economic Studies, forthcoming*, 2017.

Deutsch, Morton, Yakov Epstein, Donnah Canavan, and Peter Gumpert, "Strategies of Inducing Cooperation: An Experimental Study," *Journal of Conflict Resolution*, 1967, *11* (3), 345–360.

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

Gerber, Alan S, Donald P Green, and Christopher W Larimer, "Social Pressure and Voter Turnout: Evidence from a Large-scale Field Experiment," *American Political Science Review*, 2008, *102* (01), 33–48.

Gneezy, Uri and Jan Potters, "An Experiment on Risk Taking and Evaluation Periods," *The Quarterly Journal of Economics*, 1997, *112* (2), 631–645.

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

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

Hauser, David J and Norbert Schwarz, "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks Than do Subject Pool Participants," *Behavior research methods*, 2016, *48* (1), 400–7.

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

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

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

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

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

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

_ , "On the Interpretation of Giving in Dictator Games," *Journal of Political Economy*, 2007, *115* (3), 482–493.

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

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

Nevo, Aviv and Adam M Rosen, "Identification with Imperfect Instruments," *Review of Economics and Statistics*, 2012, *94* (3), 659–671.

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

Prelec, D., "A Bayesian Truth Serum for Subjective Data," *Science*, Oct 2004, *306* (5695), 462–466.

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

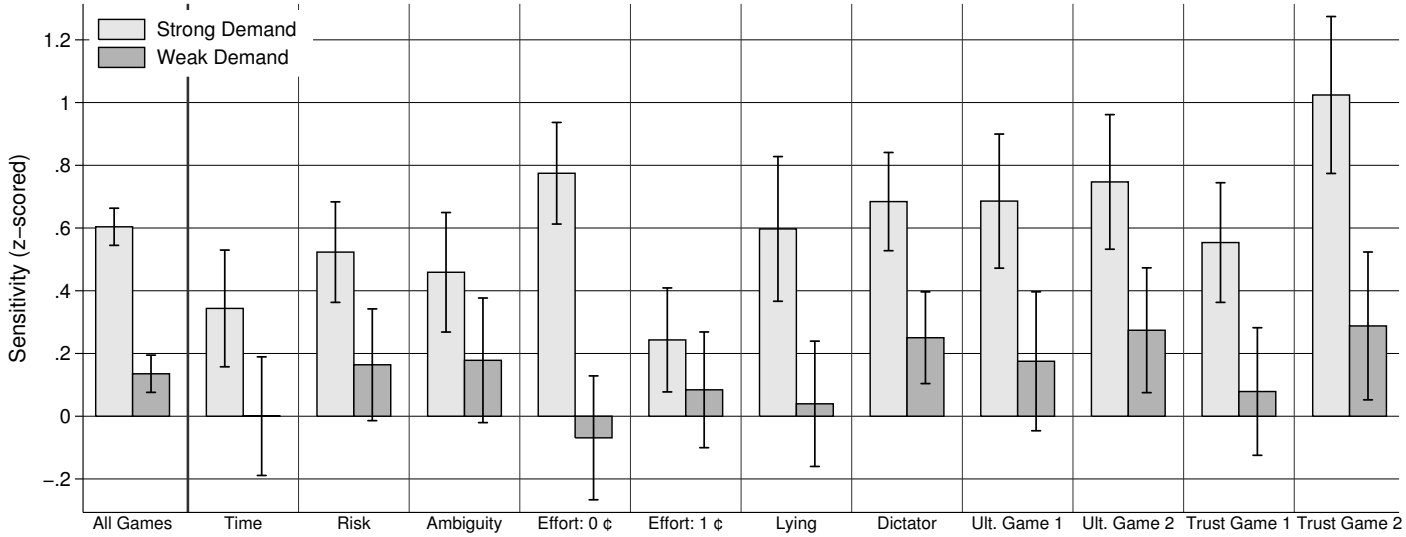
Shmaya, Eran and Leeat Yariv, "Experiments on Decisions Under Uncertainty: A Theoretical Framework," *American Economic Review*, 2016, *106* (7), 1775–1801.

Thaler, Richard H, "Anomalies: The Ultimatum Game," *The Journal of Economic Perspectives*, 1988, *2* (4), 195–206.

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

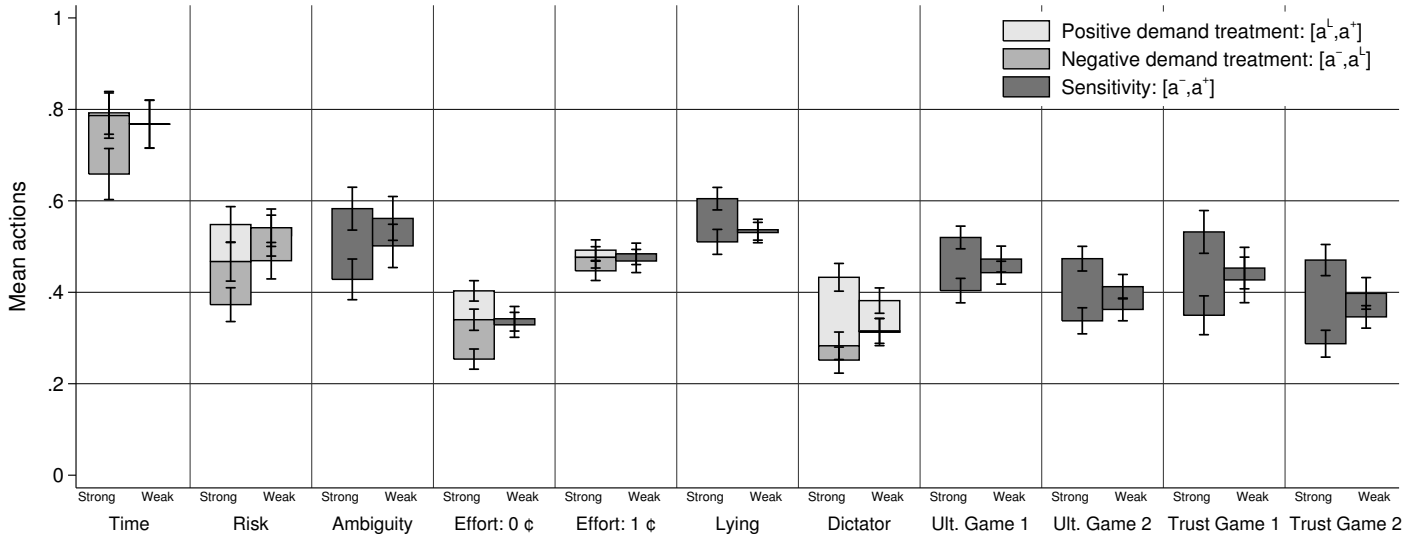
7 Main Figures and Tables

Figure 1: Sensitivity to demand treatments, z-scored



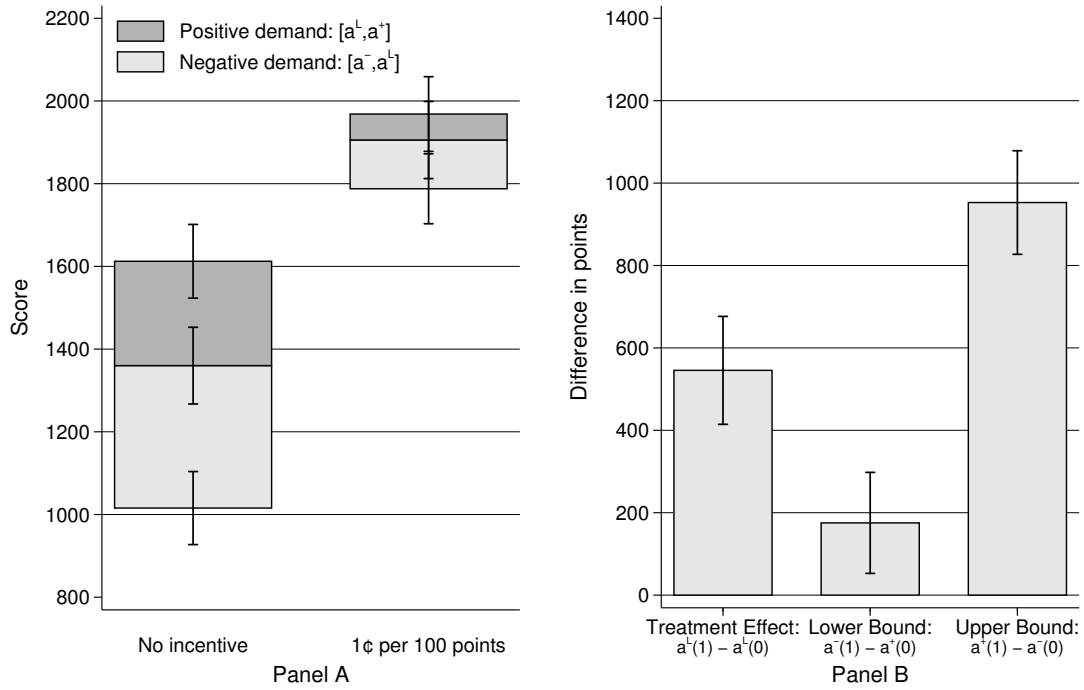
Notes: This figure uses data from all MTurk experiments with strong and weak demand treatments using real stakes. We present the z-scored sensitivity of behavior to our demand treatments, i.e. the normalized difference in behavior in the positive and negative demand condition. In our strong demand treatments we reveal the experimental objective to our respondents, while in the weak demand treatment we reveal the experimental hypothesis.

Figure 2: Bounding natural actions



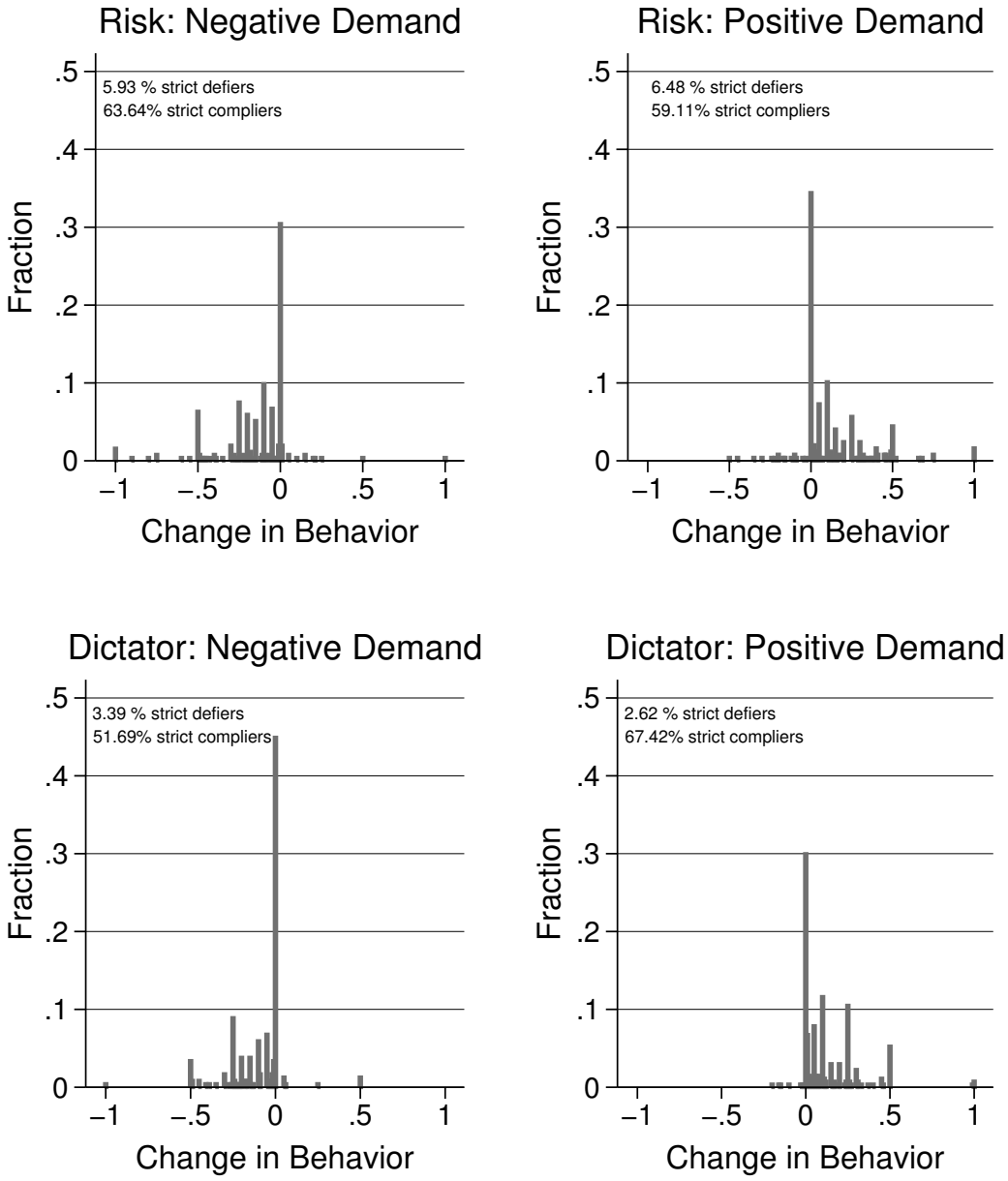
Notes: This figure uses data from all MTurk experiments with strong and weak demand treatments using real stakes. It displays the response to our strong and weak demand treatments across 11 standard preference measures. Shown are the means and the 95 percent confidence intervals for the average behavior in the positive demand treatment arm, the negative demand treatment arm, and the "no-demand" treatment arm. In the strong demand treatments we reveal the experimental hypothesis to our respondents, while in the weak demand treatment we reveal an experimental hypothesis to our respondents.

Figure 3: Bounding treatment effects



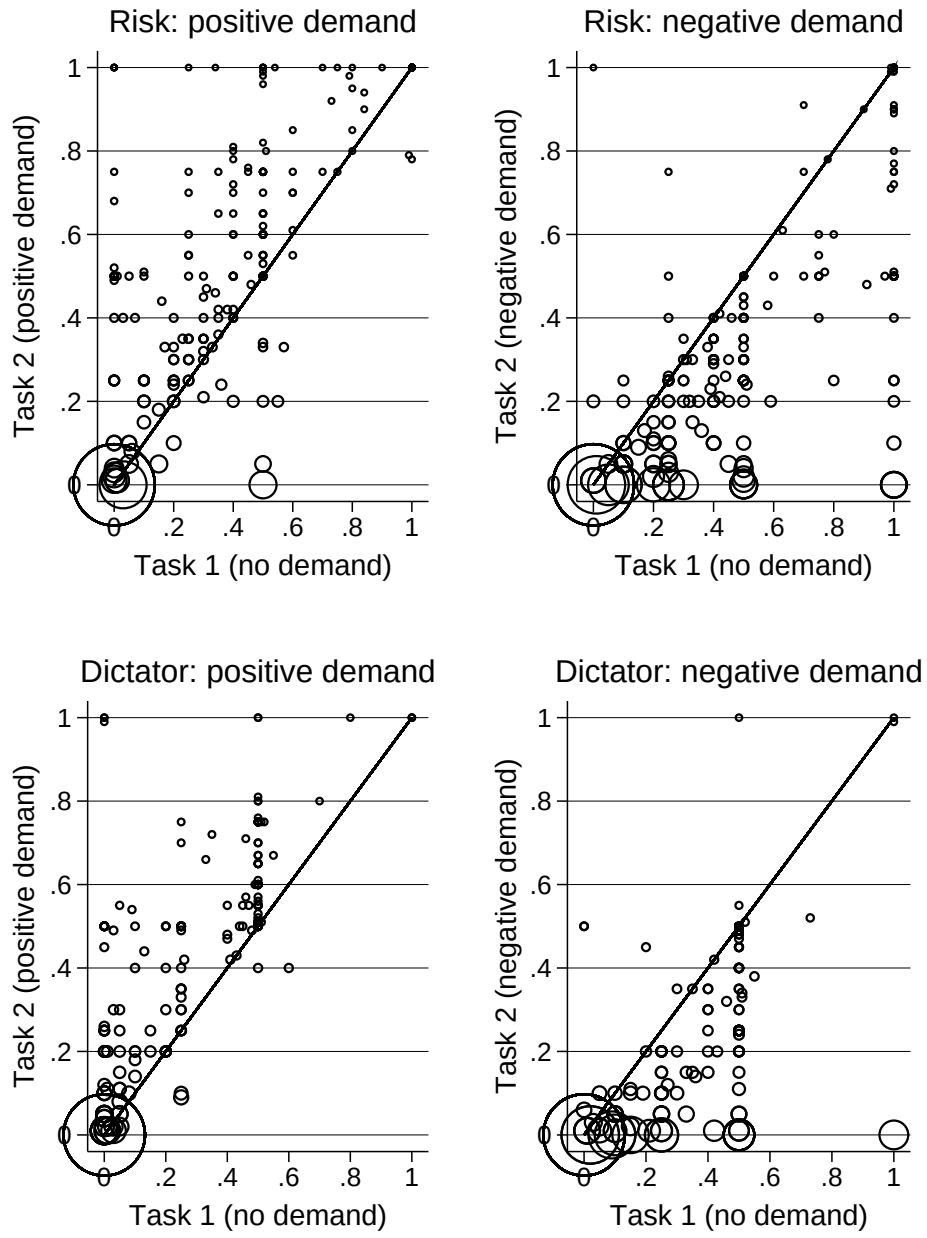
Notes: This figure uses MTurk data from the effort experiment with the strong demand treatments (experiment 3). Panel A displays mean behavior in the different demand treatment arms, disaggregated by the incentive condition, and its the 95 percent confidence interval. Panel B displays the upper and lower bounds of treatment effect estimates and their confidence intervals. In these demand treatments we reveal the experimental objective to our respondents.

Figure 4: Distribution of Response: Results from the Within Design



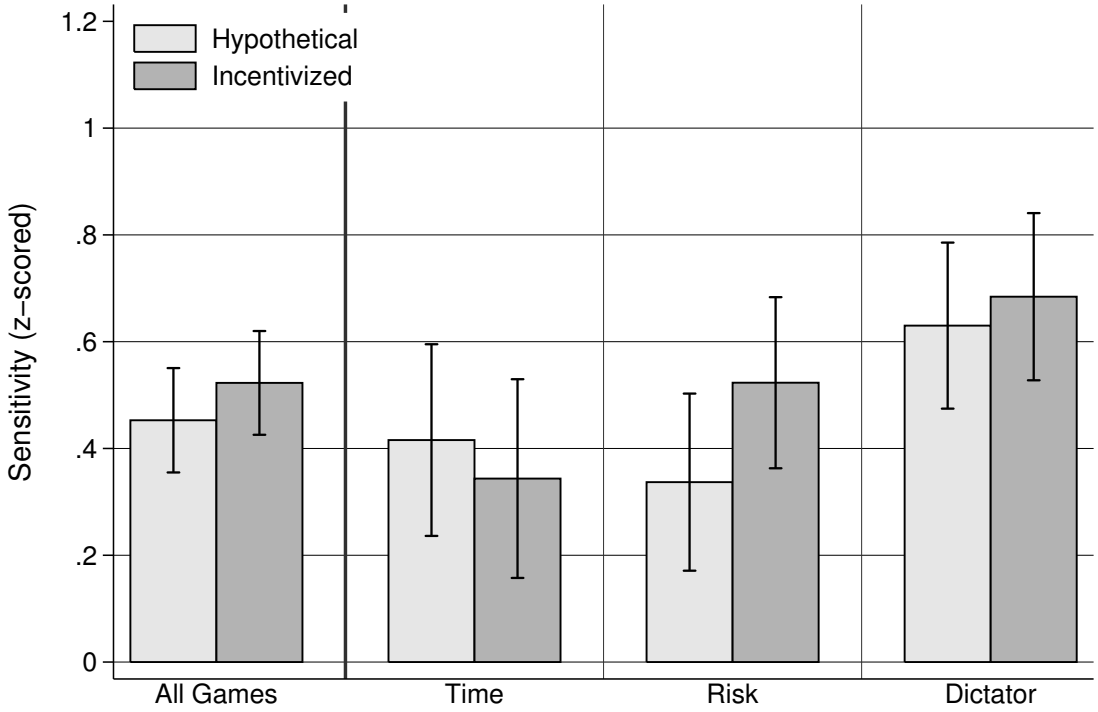
Notes: This figure uses MTurk data from experiment 7 and displays the distribution of changes in behavior (in task 2 compared to task 1) to our strong demand treatments. In these demand treatments we reveal the experimental objective to our respondents.

Figure 5: Scatterplot of Responses: Results from the Within Design



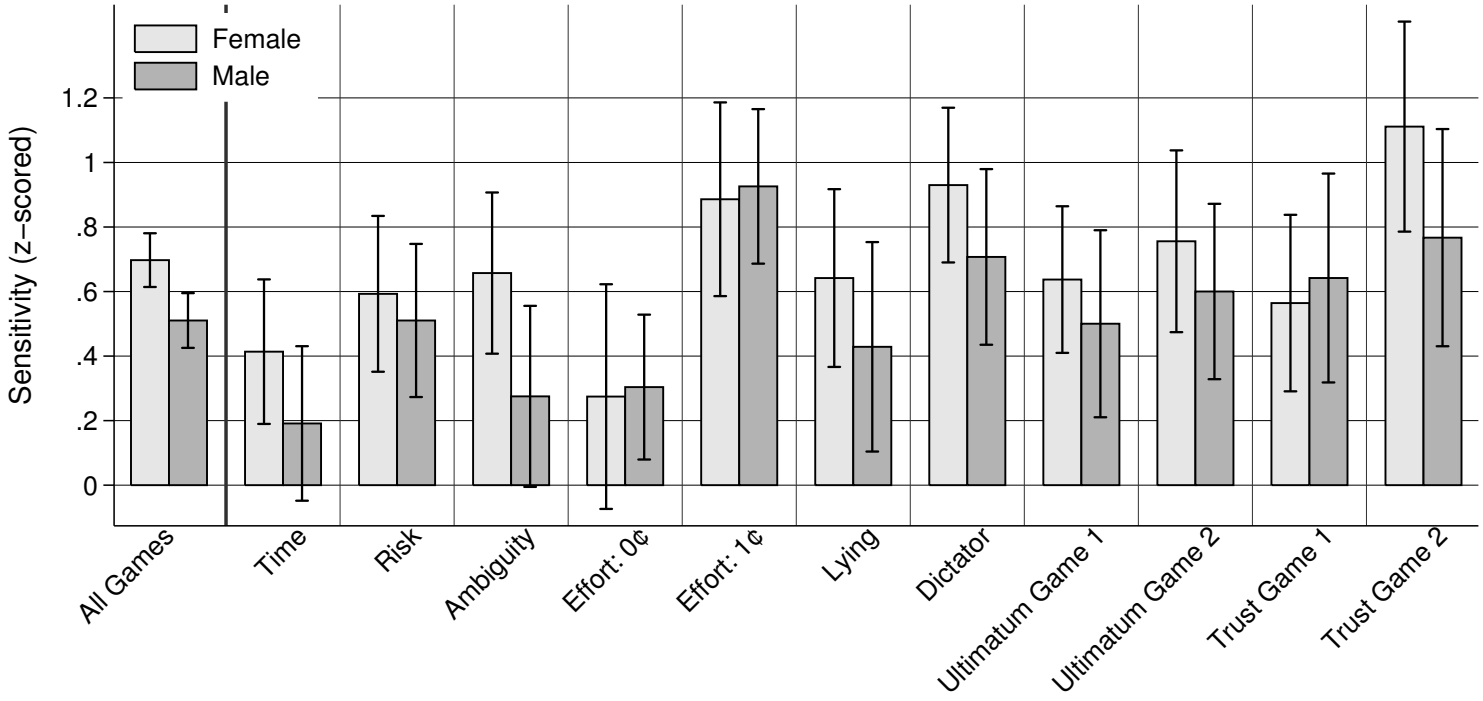
Notes: This figure uses MTurk data from experiment 7 and displays the scatterplot of responses in task 1 (neutral condition) and task 2 (demand condition). In these demand treatments we reveal the experimental objective to our respondents. The size of the rings is proportional to the frequency of outcomes.

Figure 6: Response to strong demand treatments by incentives



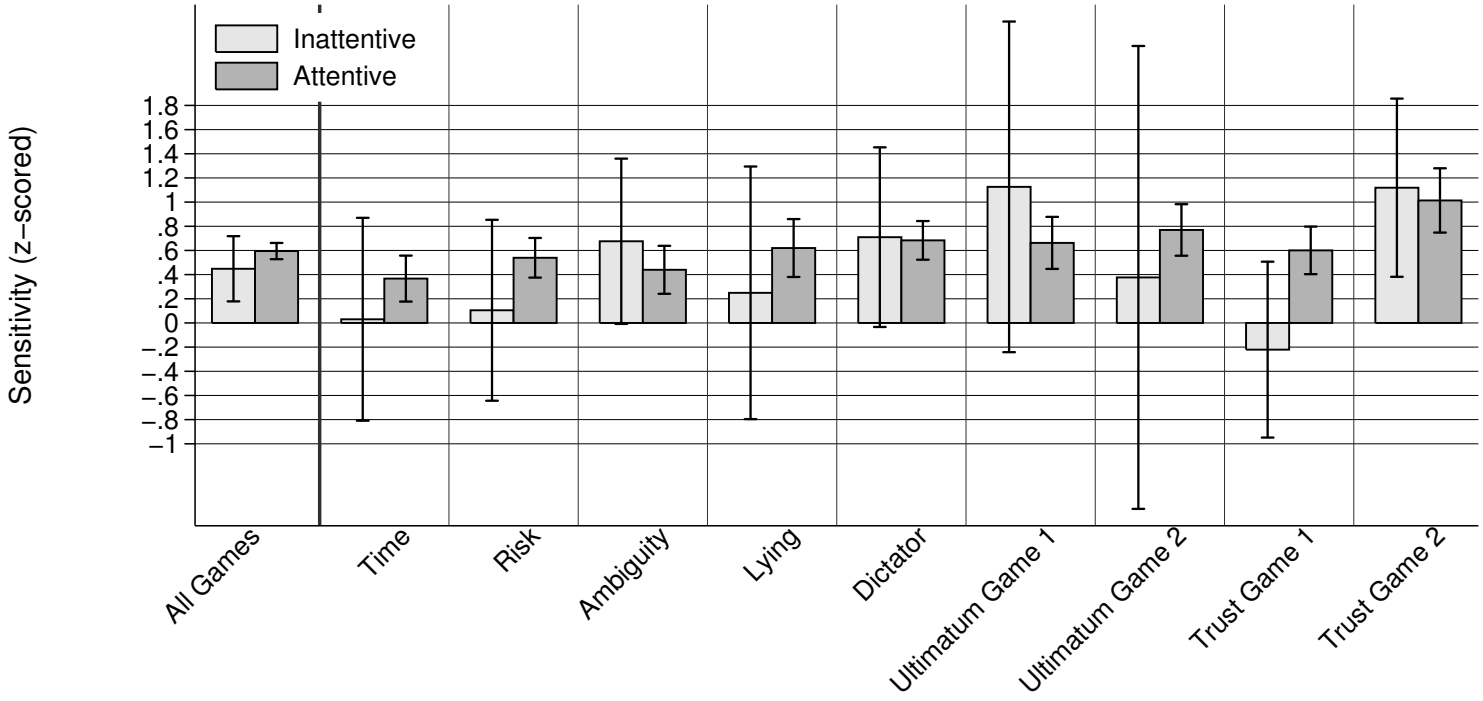
Notes: This figure uses MTurk data from experiment 1 and displays the sensitivity of behavior to our strong demand treatments by whether choices are incentivized or hypothetical. In these demand treatments we reveal the experimental objective to our respondents. The behavior in these treatment arms is z-scored at the game-level using the mean and standard deviation in the negative demand condition.

Figure 7: Gender differences in response to strong demand treatments



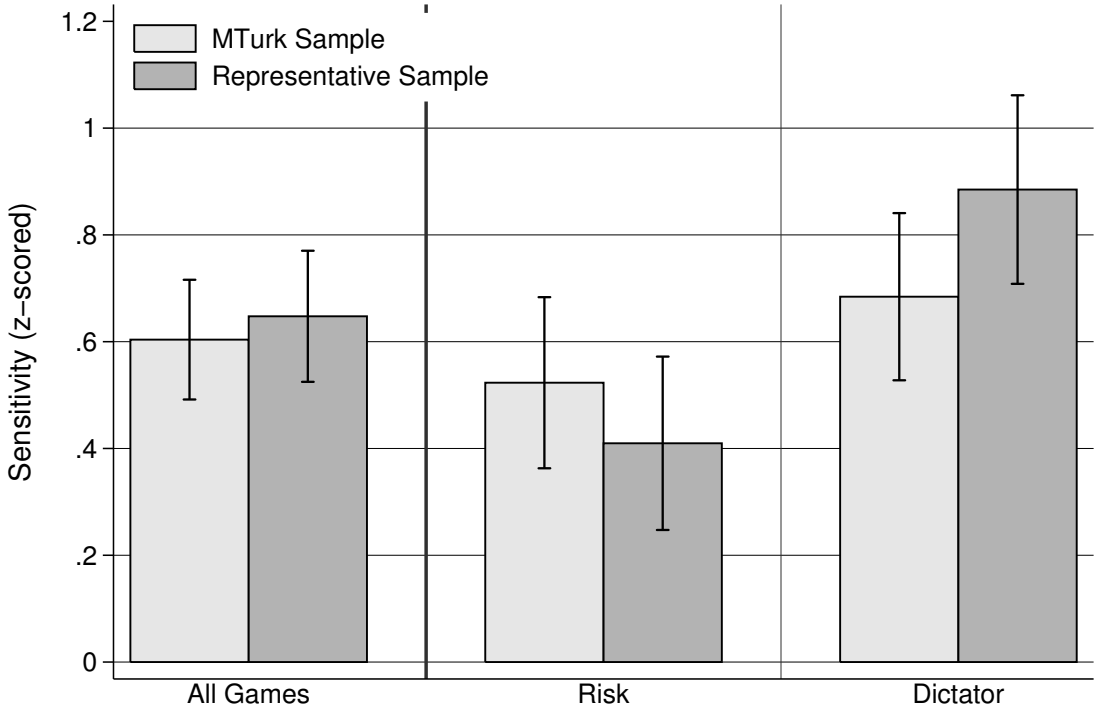
Notes: This figure uses data from all MTurk experiments with strong demand treatments using real stakes. This figure displays the sensitivity of behavior to our strong demand treatments for males and females separately across 11 standard experimental paradigms. The behavior in these treatment arms is z-scored at the game-level using the mean and standard deviation in the negative demand condition. In the figure we display the average sensitivity at the game level along with the 95 percent confidence interval. In these demand treatments we reveal the experimental objective to our respondents.

Figure 8: Response to strong demand treatments by attention



Notes: This figure uses data from all MTurk experiments with strong demand treatments using real stakes. This figure displays the response to our strong demand treatments by our respondents' level of attention. The behavior in these treatment arms is z-scored at the game-level using the mean and standard deviation in the negative demand condition. In the figure we display the average sensitivity at the game level along with the 95 percent confidence interval of the sensitivity. In these demand treatments we reveal the experimental objective to our respondents.

Figure 9: Response to strong demand treatments by population



Notes: This figure uses data from experiment 1 on MTurk using real stakes as well as data from experiment 4 with the representative online panel. This figure displays the response to our strong demand treatments separately for the MTurk sample and the representative online sample. In the figure we display the average sensitivity at the game level along with the 95 percent confidence interval of the sensitivity. In these demand treatments we reveal the experimental objective to our respondents.

Table 1: Response to strong demand treatments, all incentivized 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: Unconditional Means											
Positive demand	0.792	0.548	0.583	0.403	0.492	0.605	0.433	0.520	0.474	0.532	0.470
	(0.024)	(0.020)	(0.024)	(0.011)	(0.011)	(0.012)	(0.015)	(0.013)	(0.014)	(0.024)	(0.017)
No demand	0.786	0.467		0.340	0.476		0.283				
	(0.025)	(0.022)		(0.012)	(0.012)		(0.015)				
Negative demand	0.659	0.373	0.428	0.254	0.447	0.510	0.252	0.404	0.338	0.350	0.288
	(0.028)	(0.019)	(0.023)	(0.011)	(0.011)	(0.014)	(0.014)	(0.014)	(0.014)	(0.022)	(0.015)
Panel B: Sensitivity (positive - negative)											
Raw data	0.134***	0.175***	0.155***	0.149***	0.045***	0.095***	0.181***	0.116***	0.136***	0.182***	0.183***
	(0.037)	(0.027)	(0.033)	(0.016)	(0.016)	(0.019)	(0.021)	(0.018)	(0.020)	(0.032)	(0.023)
Z-score	0.344***	0.523***	0.459***	0.775***	0.243***	0.597***	0.684***	0.686***	0.747***	0.554***	1.024***
	(0.095)	(0.082)	(0.097)	(0.083)	(0.085)	(0.118)	(0.080)	(0.109)	(0.109)	(0.097)	(0.128)
	[0.001]	[0.001]		[0.001]	[0.012]		[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)	0.015	0.242***		0.328***	0.085		0.566***				
	(0.089)	(0.088)		(0.085)	(0.089)		(0.082)				
	[0.404]	[0.002]		[0.001]	[0.128]		[0.001]				
Negative - Neutral (z-score)	-0.328***	-0.281***		-0.447***	-0.159*		-0.118				
	(0.097)	(0.086)		(0.084)	(0.086)		(0.079)				
	[0.001]	[0.001]		[0.001]	[0.070]		[0.047]				
Observations	730	730	404	735	717	366	773	409	425	383	373

Notes: This table uses data from all MTurk experiments with strong demand treatments using real stakes. In Panel A we display the unconditional means and standard errors of those means in the positive, negative and no-demand treatment arms respectively. In Panel B we present the raw and z-scored sensitivity of behavior to our demand treatments. In Panel C we display the sensitivity of behavior in the positive and negative demand condition compared to the no-demand condition. In the demand treatments we reveal the experimental objective to our respondents. Robust standard errors are in parentheses. False-discovery reate adjusted p-values are in brackets. The p-value of an F-test which tests for differences in response to demand across all games is 0.000. * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level.

Table 2: Bounding treatment effects

	(1) Score	(2) Score (z-scored)
Panel A: Sensitivity		
Positive demand **[1]**	252.342***	0.357***
	(65.307)	(0.092)
Negative demand **[2]**	-344.466***	-0.488***
	(65.030)	(0.092)
1-cent bonus **[3]**	545.549***	0.773***
	(66.799)	(0.095)
Positive demand × 1-cent-bonus **[4]**	-189.547**	-0.268**
	(92.800)	(0.131)
Negative demand × 1-cent-bonus **[5]**	226.574**	0.321**
	(91.188)	(0.129)
Panel B: Bounding		
Conventional treatment effect: **[3]**	545.549***	0.773***
	(66.799)	(0.095)
Lower bound: **[2]** + **[3]** + **[5]** - **[1]**	175.315***	0.248***
	(62.366)	(0.088)
Upper bound: **[1]** + **[3]** + **[4]** - **[2]**	952.811***	1.349***
	(64.136)	(0.091)

Notes: In this table we use data the real effort experiment using strong demand treatments (experiment 3). In Panel A we present the sensitivity of effort to monetary incentives, our demand treatments and interactions of the demand treatments and monetary incentives. In Panel B we display the conventional treatment effects, the lower as well as the upper bound of treatment effects. In column 1 we present the results in terms of raw real-effort, while in column 2 we display z-scored real effort. Robust standard errors in parentheses. * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level.

Table 3: Response to weak demand treatments, all incentivized 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: Unconditional Means											
Positive demand	0.768	0.524	0.562	0.329	0.484	0.537	0.382	0.473	0.412	0.453	0.398
	(0.027)	(0.023)	(0.024)	(0.014)	(0.012)	(0.012)	(0.014)	(0.014)	(0.013)	(0.023)	(0.017)
No demand		0.541					0.313				
		(0.021)					(0.015)				
Negative demand	0.768	0.469	0.501	0.342	0.468	0.531	0.316	0.443	0.362	0.427	0.346
	(0.026)	(0.020)	(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	0.000	0.055*	0.060*	-0.013	0.016	0.006	0.066***	0.030	0.050***	0.026	0.051**
	(0.038)	(0.030)	(0.034)	(0.019)	(0.017)	(0.016)	(0.020)	(0.019)	(0.018)	(0.034)	(0.021)
Z-score	0.000	0.164*	0.178*	-0.069	0.084	0.040	0.250***	0.175	0.274***	0.079	0.288**
	(0.096)	(0.091)	(0.101)	(0.101)	(0.094)	(0.102)	(0.075)	(0.113)	(0.101)	(0.104)	(0.120)
		[0.077]					[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)		-0.051					0.260***				
		(0.092)					(0.078)				
		[0.237]					[0.001]				
Negative - Neutral (z-score)		-0.215**					0.010				
		(0.087)					(0.077)				
		[0.041]					[0.426]				
Observations	426	743	393	392	383	413	761	361	413	355	347

Notes: This table uses data from all MTurk experiments with weak demand treatments using real stakes. In Panel A we display the unconditional means and standard errors of those means in the positive, negative and no-demand treatment arms respectively. In Panel B we present the raw and z-scored sensitivity of behavior to our demand treatments. In Panel C we display the sensitivity of behavior in the positive and negative demand condition compared to the no-demand condition. In the demand treatments we reveal the experimental hypothesis to our respondents. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets. The p-value of an F-test which tests for differences in response to demand across all games is 0.063. * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level.

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		0.183**	0.259***		0.213***	0.306***
		(0.091)	(0.091)		(0.078)	(0.064)
$h^L(0)p^L(0)$		-0.720***	-0.488		-0.506***	-0.162
		(0.174)	(0.298)		(0.193)	(0.247)
$h^L(>0)p^L(>0)$		-0.404			0.958	
		(2.072)			(1.732)	
$h^L(1)p^L(1)$			-0.372			0.197
			(1.033)			(0.665)
$h^L(4)p^L(4)$			-6.412**			-6.516***
			(3.082)			(1.864)
s	0.033	0.188**	0.288**	0.031	0.242**	0.529**
	(0.049)	(0.095)	(0.125)	(0.045)	(0.097)	(0.221)
k	6.0e-26	4.2e-23	3.9e-16	4.6e-08	3.0e-06	2.3e-04
	(3.9e-25)	(1.5e-22)	(1.7e-15)	(1.9e-07)	(4.9e-06)	(3.7e-04)
γ	7.228***	6.354***	4.196***	6.5e-03***	4.5e-03***	2.2e-03***
	(2.188)	(1.215)	(1.543)	(2.1e-03)	(8.2e-04)	(7.7e-04)
Observations	729	1699	1699	729	1699	1699
R-squared	0.125	0.167	0.168	0.169	0.205	0.207

Notes: This table uses data from the the real effort experiments on MTurk with strong demand treatments. Coefficients s and ϕ are measured in cents. s measures the respondents intrinsic motivation. ϕ measures the monetary equivalent of acting according to the experimental objective for a worker who is certain about the experimenter's objective. γ is the estimate of the cost curvature (inverse of the elasticity of effort) and k is the scaling parameter. $h^L(0)p^L(0)$ measures latent demand in the no-incentive condition. $h^L(>0)p^L(>0)$ measures latent demand in the 1-cent and 4-cent incentive conditions. $h^L(1)p^L(1)$ measures latent demand in the 1-cent incentive condition. $h^L(4)p^L(4)$ measures latent demand in the 4-cent incentive condition. Robust standard errors in parentheses. * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level.

Table 5: Moderators of response to 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: Design Characteristics												
Sensitivity × Incentive	0.058	-0.072	0.183					0.058				
	(0.072)	(0.132)	(0.116)					(0.121)				
Observations	3000	998	1000					1002				
Panel B: Respondent Characteristics												
Sensitivity × Male	-0.143**	-0.222	-0.128	-0.382**	0.040	0.029	-0.213	-0.238	-0.137	-0.156	0.078	-0.344
	(0.057)	(0.167)	(0.125)	(0.192)	(0.196)	(0.211)	(0.217)	(0.151)	(0.188)	(0.200)	(0.216)	(0.239)
Observations	6013	494	1071	404	495	475	366	1118	409	425	383	373
Sensitivity × Attention	0.135	0.311	0.461	-0.276			0.249	-0.043	-0.272	0.228	0.908**	-0.084
	(0.141)	(0.393)	(0.398)	(0.414)			(0.358)	(0.610)	(0.394)	(0.538)	(0.409)	(0.310)
Observations	5043	494	1071	404			366	1118	409	425	383	373
Sensitivity × Representative sample	0.054		-0.127					0.236*				
	(0.098)		(0.130)					(0.141)				
Observations	2189		1071					1118				

Notes: The outcome variables are normalized at the game-level. In Panel A we display heterogeneous treatment effects of the strong demand treatments by design characteristics, i.e. whether our respondents' choices are incentivized or hypothetical. In Panel B we display heterogeneous treatment effects by respondent characteristics, namely by gender, attention and whether our respondents come from MTurk or a representative sample. The variable male takes value one if our respondent is male and zero otherwise, attention takes value one if our respondent correctly completed the screener and zero otherwise. The variable representative sample takes value if our respondent comes from a representative sample and zero when they come from the MTurk sample. In the strong demand treatments we reveal the experimental objective to our respondents. * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level.

Table 6: Moderators of response to weak 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: Design Characteristics												
Sensitivity × Incentive	0.079		0.140					0.017				
	(0.085)		(0.125)					(0.115)				
Observations	1976		978					998				
Panel B: Respondent Characteristics												
Sensitivity × Male	-0.001	-0.020	-0.070	0.059	0.349	0.035	0.064	-0.137	-0.113	0.015	0.282	-0.218
	(0.058)	(0.173)	(0.133)	(0.203)	(0.238)	(0.232)	(0.188)	(0.148)	(0.193)	(0.185)	(0.229)	(0.229)
Observations	5618	426	1046	393	392	383	413	1089	361	413	355	347
Sensitivity × Attention	0.147	-0.413	-0.067	0.440			0.224	0.692*	0.105	0.364	0.119	-0.389
	(0.124)	(0.395)	(0.305)	(0.504)			(0.305)	(0.377)	(0.296)	(0.230)	(0.362)	(0.301)
Observations	4843	426	1046	393			413	1089	361	413	355	347
Sensitivity × Representative sample	0.026		0.028					0.026				
	(0.100)		(0.140)					(0.139)				
Observations	2135		1046					1089				

Notes: The outcome variables are normalized at the game-level. In Panel A we display heterogeneous treatment effects of the strong demand treatments by a design characteristics, i.e. whether our respondents' choices are incentivized or hypothetical. In Panel B we display heterogeneous treatment effects by respondent characteristics, namely by gender, attention and whether our respondents come from MTurk or a representative sample. The variable male takes value one if our respondent is male and zero otherwise, attention takes value one if our respondent correctly completed the screener and zero otherwise. The variable representative sample takes value if our respondent come from a representative sample and zero when they come from the MTurk sample. In the weak demand treatments we reveal the experimental hypothesis to our respondents. * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level.

Table 7: Overview of Experiments

Experiment	**Sample**	**Games**	**Demand Treatments**	**Real or Hypothetical**
Experiment 1	MTurk (N=4495)	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	MTurk (N=2964)	Dictator Game and Investment Game	Weak positive demand, weak negative demand and no-demand treatment	Both real stakes and hypothetical choices
Experiment 3	MTurk (N=1452)	Effort experiment with 1 cent bonus and Effort experiment with no bonus	Strong positive, strong negative and no-demand treatment	Real stakes (real effort experiment)
Experiment 4	Representative online Panel (N=2941)	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	MTurk (N=5068)	Trust game (first and second mover), Ultimatum game (first and second mover), Coinflip game, Investment Game with uncertainty (ambiguity aversion) and Convex Time Budgets	Strong positive demand, strong negative demand, weak positive demand and weak negative demand	Real stakes
Experiment 6	MTurk (N=775)	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	MTurk (N=1002)	Dictator Game and Investment Game	Within design: Task 1: no demand treatment; task 2: strong positive demand or strong negative demand	Real stakes

Notes: This table summarizes the key design features of each of the experiments. In experiment 5 for convex time budgets we only employ the weak demand treatments, while for all other games we employ both strong and weak demand treatments. In experiment 2 we also have an additional incentive treatment arm in which respondents receive four cents per 100 tasks in which we do not induce any additional demand.

Table 8: Design by Games: Strong Demand Experiments

Game	**Description**	**Sample**	**Show-up fee**	**Choice set**	**Demand Instructions**
Dictator Game	Choose to split money between yourself and another participant.	MTurk (N=1,508) Representative Sample (N=899)	$.25	action ∈ [0,1]	"You will do us a favor if you give more (less) to the other participant than you normally would."
Investment Game	Choose to how much to invest in a risky project.	MTurk (N=1,499) Representative Sample (N=902)	$.25	action ∈ [0,1]	"You will do us a favor if you invest more (less) than you normally would."
Investment Game with ambiguous returns	Choose to how much to invest in a project with uncertain returns.	MTurk (N= 404)	$.25	action ∈ [0,1]	"You will do us a favor if you invest more (less) than you normally would."
Convex Time Budgets	Choose between receiving money today vs. money in seven days.	MTurk (N=1,488); Rep. Sample (N=899)	$.25	action ∈ [0,1.2]	"You will do us a favor if you choose to receive more (less) in seven days than you normally would."
Effort: No bonus	Alternately press the a and b button without receiving any bonus.	MTurk (N=735)	$1	action ∈ [0,4000]	"You will do us a favor if you work harder (less hard) than you normally would."
Effort: 1-cent bonus	Alternately press the a and b button while receiving 1 cent per 100 scores.	MTurk (N=717)	$1	action ∈ [0,4000]	"You will do us a favor if you work harder (less hard) than you normally would."
Trust Game 1st mover	Choose to send an amount of money to the other player.	MTurk (N=383)	$.25	action ∈ [0,.2,.4,.6,.8,1]	"You will do us a favor if you send more (less) to the other participant than you normally would."
Trust Game 2nd mover	Choose to send back some money to the other player. (Strategy method)	MTurk (N=373)	$.25	action ∈ [0,1.2]	"You will do us a favor if you send back more (less) to the other participant than you normally would."
Ultimatum Game 1st mover	Offer a split to the other player.	MTurk (N=409)	$.25	action ∈ [0,1]	"You will do us a favor if you offer more (less) to the other participant than you normally would."
Ultimatum Game 2nd mover	Choose to send back some money to the other player.	MTurk (N=425)	$.25	action ∈ [0,1]	"You will do us a favor if you require a higher (lower) minimum amount than you normally would."
Lying	Report the number of "Heads" after 10 coin-flips.	MTurk (N=366)	$.25	action ∈ [0,1,2,3,4,5,6,7,8,9,10]	"You will do us a favor if you report more (fewer) heads than you normally would."

Notes: This table summarizes the key design features of each of the games.

Table 9: Design by Games: Weak Demand Experiments

Game	**Description**	**Sample**	**Show-up fee**	**Choice set**	**Demand Instructions**
Dictator Game	Choose to split money between yourself and another participant.	MTurk (N=1,482) Representative Sample (N=865)	$.25	action ∈ [0,1]	"We expect that participants who are shown these instructions will give more (less) to the other participant than they normally would."
Investment Game	Choose to how much to invest in a risky project.	MTurk (N=1,482) Representative Sample (N=883)	$.25	action ∈ [0,1]	"We expect that participants who are shown these instructions will invest more (less) than they normally would."
Investment Game with ambiguous returns	Choose to how much to invest in a project with uncertain returns.	MTurk (N= 393)	$.25	action ∈ [0,1]	"We expect that participants who are shown these instructions will invest more (less) than they normally would."
Convex Time Budgets	Choose between receiving money today vs. money in seven days.	MTurk (N=426)	$.25	action ∈ [0,1.2]	"We expect that participants who are shown these instructions will 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.	MTurk (N=392)	$1	action ∈ [0,4000]	"We expect that participants who are shown these instructions will work harder (less hard) than they normally would."
Effort: 1-cent bonus	Alternately press the a and b button while receiving 1 cent per 100 scores.	MTurk (N=383)	$1	action ∈ [0,4000]	"We expect that participants who are shown these instructions will work harder (less hard) than they normally would."
Trust Game 1st mover	Choose to send an amount of money to the other player.	MTurk (N=355)	$.25	action ∈ [0,.2,.4,.6,.8,1]	"We expect that participants who are shown these instructions will 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)	MTurk (N=347)	$.25	action ∈ [0,1.2]	"We expect that participants who are shown these instructions will send back more (less) to the other participant than they normally would."
Ultimatum Game 1st mover	Offer a split to the other player.	MTurk (N=361)	$.25	action ∈ [0,1]	"We expect that participants who are shown these instructions will offer more (less) to the other participant than they normally would."
Ultimatum Game 2nd mover	Choose to send back some money to the other player.	MTurk (N=413)	$.25	action ∈ [0,1]	"We expect that participants who are shown these instructions will require a higher (lower) minimum amount than they normally would."
Lying	Report the number of "Heads" after 10 coin-flips.	MTurk (N=413)	$.25	action ∈ [0,1,2,3,4,5,6,7,8,9,10]	"We expect that participants who are shown these instructions will report more (fewer) heads than they normally would."

Notes: This table summarizes the key design features of each of the games.

Table 10: Results from the Within Design

	Dictator			Risk		
	Within	Between	Difference	Within	Between	Difference
Panel A: Unconditional Means						
Positive demand	0.383	0.433	-0.050**	0.560	0.548	0.012
	(0.017)	(0.015)	(0.023)	(0.021)	(0.020)	(0.029)
No demand	0.271	0.283	-0.012	0.448	0.467	-0.020
	(0.011)	(0.015)	(0.019)	(0.015)	(0.022)	(0.026)
Negative demand	0.193	0.252	-0.058***	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***	0.181***	0.008	0.242***	0.175***	0.067*
	(0.022)	(0.021)	(0.031)	(0.029)	(0.027)	(0.040)
Z-score	0.794***	0.736***	0.058	0.709***	0.514***	0.195*
	(0.093)	(0.086)	(0.127)	(0.084)	(0.080)	(0.116)
Panel C: Monotonicity						
Positive - Neutral (z-score)	0.512***	0.609***	-0.097	0.377***	0.238***	0.139
	(0.044)	(0.088)	(0.129)	(0.041)	(0.087)	(0.124)
Negative - Neutral (z-score)	-0.376***	-0.128	-0.249**	-0.427***	-0.277***	-0.151
	(0.045)	(0.086)	(0.122)	(0.042)	(0.084)	(0.119)
Observations	502	773	1275	500	730	1230

Notes: This table uses data from the within design (experiment 7) and inventivized choices from the dictator game and the investment game in experiment 1. These experiments employ strong demand treatments in which the experimental objective is revealed to participants.