Experimenter Demand Effects[1]

Jonathan de Quidt
Lise Vesterlund
Alistair J. Wilson

February 2018

Abstract: A study's internal and external validity is threatened by experimenter demand effects. This threat is taken seriously by experimental economists, who have developed a number of best practices to suppress or eliminate the potential role of such effects. We outline these best practices and review the literature to show that they are followed in the vast majority of published work. This adherence to best practice likely contributes to the limited evidence of experimenter demand effects uncovered in the literature. Specifically, we are not aware of examples where demand effects have been shown to influence the qualitative inference from a study. While good design goes a long way towards reducing the potential for experimenter demand effects, a complementary option, presented in our final section, is to derive bounds on the effect.

Keywords: Experimenter Demand Effect, Qualitative Results, Best-Practice Design, Mitigating Demand Effects, Measuring Demand Effects

[1] We thank Felipe Augusto de Araujo for excellent research assistance.

1 Introduction

In conducting experiments we hope to uncover causal evidence of behavioral regularities that can inform theory, and effects that are predictive both of behavior in similar experimental studies and in comparable settings outside the laboratory. Claims of *experimenter demand effects* are therefore a serious accusation, threatening both the internal and external validity of the study.[2]

Experimenter demand effects refer to changes in behavior that result from study participants wanting to help the experimenter confirm her underlying hypothesis.[3] With participants deviating from the choice they would select absent the 'experimenter'---henceforth the 'true' preferred choice---the study will produce biased results. Internal validity is compromised because more than the independent variable of interest is changing between treatments, and external validity is threatened because a central feature of the study (the presence of the experimenter) influences behavior, but is generally not a factor in the environments we are trying to model.[4]

Not surprisingly, a strong claim that experimenter demand effects are driving the results is typically a deathblow to an experimental study. Arguing against the potential role for experimenter demand, it is often noted that it is unlikely that participants can guess the experimenter's preferred outcome; that often the experimenter does not have a preferred outcome; and that it is unlikely that the participants will deviate from their true preferred choices to benefit the experimenter.

While an eagerness to confirm the experimenter's hypothesis is the primary bias that comes to mind, note that the concern extends to all experimenter-induced deviations from true preferred choice. Biased results also arise when participants falsely infer a hypothesis or want to contradict the experimenter's inferred hypothesis. Concerns for experimenter demand are thus best addressed by designs where hypothesis speculation is minimized and incentives are salient.[5]

The potential for drawing incorrect conclusions due to experimenter demand depends on the type of study conducted. As noted by Kessler and Vesterlund (2015) the vast majority of experimental studies aim to uncover *qualitative* results. That is, the aim is to identify the direction of an effect, the sign of a particular comparative static. The concern for experimenter demand in such a context is primarily one of internal validity. Did the directional changes in Y result from X, or from other factors that changed alongside the experimental variation of X? Fortunately, we are

---

[2] Rephrasing Guala (2002, p.262), an experimental result is *internally valid*, if the experimenter attributes the production of an effect Y to the factor X, and X really is the cause of Y in the experimental set-up E. The experimental result is *externally valid* if X causes Y not only in E, but also in a set of other circumstances of interest F, G, H, etc.

[3] Similar to Camerer (2011, p.260) we will consider experimenter demand effects as those resulting specifically from 'the experimenter's demand.' Experimenter demand is therefore one of many factors that can reduce the external validity of a study. While other behavioral responses that result from being assessed in the laboratory, or from the context in which decisions are embedded may also threaten external validity, they need not result from experimenter demand. See Zizzo (2010) for a more nuanced discussion and for a parallel discussion of potential demand biases.

[4] The experimenter is an integrated part of studies in which the experimenter role corresponds to agent roles in the environment of interest, for example, the auctioneer, the fundraiser, etc.

[5] Ignoring the fact that deception is not accepted in economic experiments (Ortmann 2002), falsely convincing participants of an underlying hypothesis does not eliminate experimenter demand concerns.

unaware of any evidence for experimenter demand either generating or reversing the directional results of a best-practice experiment. In fact, there is broad consensus that qualitative inference in the laboratory is both internally and externally valid (see Kessler and Vesterlund, 2015).

The potential role for demand biases is greater when assessing a *quantitative* response within a study. Indeed, evidence of behavioral shifts in response to experimental framing is often seen as indicative of such a quantitative response. Recent studies have provided bounds on such effects (de Quidt et al, 2017), though it is not possible to directly assess the magnitude of the actual bias. Recognizing that experimenter demand calls into question the external validity of quantitative measures secured in the laboratory, it is important to note that the effect is only one of many other factors that affect the external validity of an elicited level effect. Irrespective of demand effects, there are few measures elicited in the lab that are expected to be predictive of the levels outside of the laboratory.[6]

Although there is no evidence of experimenter demand effects generating a qualitative inference in best-practice experiments, this does not imply that the profession is not concerned about such effects. Rather, the norms and best practices surrounding experimental design are strongly influenced by the desire to mitigate the potential role for experimenter demand. Moreover, we know that participants do respond (sometimes substantially) to experimental instructions that explicitly spell out the experimental objective.[7] While there is certainly the potential for demand biases if participants can form strong conjectures about the experimenter's objectives, such results point to the value of good design that guards against such intrusions in the first place.

This chapter will review steps that help mitigate and potentially control for experimenter demand effects. Indeed, concerns for experimenter demand effects have played a central role in shaping the norms and best practices for experimental design. In Sections 2 and 3 we review the techniques that constitute best practice for mitigating experimenter demand.

The aim of these procedures is to make it difficult for participants to guess the experimental hypothesis; mitigate emphasis on other potential hypotheses; and reduce participants' responsiveness to such potential speculation. By making hypotheses less salient and more difficult to guess, the experimenter hopes to reduce not only the correlation between participants' guesses

---

[6] The external validity of concern to most experiments is *qualitative* in the sense that a relationship between two variables hold across similar environments, for example, Guala (2002). As numerous factors change between study and non-study environments (incentives, participants, setting or general rules surrounding the decision, etc.), there is rarely a claim that the *quantitative* relationship between two variables is externally valid. Kessler and Vesterlund (2015) note that "Few experimental economists would argue that the magnitude of the difference between two laboratory treatments is indicative of the magnitude one would expect to see in the field or even in other laboratory studies in which important characteristics of the environment have changed. For example, the revenue difference between an English auction and a first-price sealed bid auction in the laboratory is not thought to be indicative of the quantitative difference one would find between any other set of English and first-price sealed bid auctions." Quantitative effect sizes become more important in structural exercises when we wish to compare effect sizes between treatments or experiments.

[7] While de Quidt et al. (2017) explicitly manipulate this channel, responses are also found in experiments that manipulate calls to authority (see for example, Silverman et al. 2014; Karakostos and Zizzo 2016).

and the actual experimental treatments, but also to reduce the distortion between 'true' and observed choices.

As evidence that these are generally accepted norms, in Section 4 we review recently published research in experimental economics, and show that these techniques *are* broadly accepted and applied. Examining experimental work published in the last five years---in both top-general interest journals and the top-field outlet for experimental work---we document the widespread adoption of the specified best practices. Consonant to the absence of evidence for qualitative demand effects, there is also little conclusive *evidence* that these procedures do reduce experimenter demand. However, good design can certainly be a useful as a talisman to ward off critiques that a result might be driven by experimenter demand.

Finally, in Section 5 we discuss ways of evaluating whether the results of a study are driven by experimenter demand. The approach commonly taken is that of robustness checks and replication by others, however recent work points to techniques that help bound the potential for bias in a particular study.

2 General Design and Procedures

The features of interest when reviewing a study are whether deviations from true choice is costly, that is, whether choice is properly incentivized; whether instructions and elicitation material provide cues on preferred behavior; and whether the interaction between the experimenter and participants is controlled and minimized to reduce undue social pressure or undue guidance on preferred choice.

2.1 Incentives

There are strong norms in experimental economics that choices should be incentivized. Incentives are used to induce preferences, and are believed to increase attentiveness, reduce noise in decision-making (by ensuring preferences are "strict") and improve ecological validity. Hypothetical decisions are often treated with skepticism within the profession.

An important advantage of incentivized choice is that it makes it costly to deviate from 'true' preferred choices, and thus should limit experimenter demand effects. For instance, a participant in a binary choice task whose preference is for option 1, but who believes the experimenter wants them to choose option 2, is presumably less likely to go with the experimenter's wishes if she gives up $10 to do so, as opposed to in a hypothetical choice.

Crucial to this argument is that incentives are sufficiently elastic to choices, making deviations from the true choice costly enough. This is a particular concern in designs where optimal choices are driven by weak marginal incentives (either because the monetary incentives themselves are

flat, as in some belief elicitations, see Danz et al. 2017, or because the intrinsic incentives generated are flat, see for example, Araujo et al. 2016 on the slider task).[8]

The literature on incentives documents a partial (but inconsistent) response to incentives. For example, Camerer (1999) reviews 74 experiments with varying incentives, where he finds evidence that stakes improve performance and decrease noise in some tasks. However, he also notes that incentives often do not affect mean behavior, and argues that their importance should not be over-emphasized. In a similar vein, Amir (2012) replicates a number of classic results online at low stakes using Amazon Mechanical Turk. Camerer (2011) argues that insensitivity to stakes suggests that concerns about demand effects are overblown. Ariely et al. (2009) however finds changes in behavior when presented with large incentives, and in reviewing the literature Gneezy et al. (2011) demonstrate that behavior sometimes is very sensitive to the magnitude of the incentives. The research on sensitivity to incentives however does not demonstrate, nor imply, that the response results from experimenter demand. The only direct evidence on such interaction is seen in a recent study de Quidt et al. (2017) (discussed further in Section 5), which finds similar sensitivity to explicit "demand treatments" in both hypothetical and incentivized choices, though the monetary incentives here are somewhat small.

2.2 Neutral Instructions

Participants in economic experiments are generally presented with very abstract decision environments, with much naturalistic context deliberately stripped away. For example, risky decision-making is studied using choices between abstract lotteries, attitudes toward delay and patience through allocation of monetary payments over time, strategic reasoning and behavior through arbitrarily labeled actions and payoff tables.

The reasoning for such abstract frames is a combination of a theory-grounded pursuit for deep preferences, phenomena and mechanisms, and a desire to study domain-general behaviors so that lessons learned can be applied across contexts. There is certainly evidence that behavior can be sensitive to framing in ways that might be orthogonal to the experimental question of interest. Moreover, in strategic settings, frames can serve to change beliefs about other participants' choices and shift the resulting equilibria (Ellingsen 2012).

Abstract framing helps avoid anchoring on a particular environment, mitigates conjectures about what behavior the experimenter anticipates, and focus attention on the information and payoffs provided in the experiment. In our view, these features make abstract frames the best practice for mitigating experimenter demand.

That said, we are not aware of direct evidence that framing, neutral or otherwise, influences demand biases. While there is evidence of response to framing, this need not be evidence of

[8] Araujo et al. (2016) show essentially a null result in a real-effort task as they vary the piece-rate incentives from a half-cent per piece to eight-cents per piece (a 1,500 percent increase), which they attribute to subjects' self-driven incentives to perform in the task and the lack of an outside-option activity.[9] A set of dictator-game experiments by Dreber (2012) find very little sensitivity of giving to different neutral and non-neutral frames.

experimenter demand. For example, changes in beliefs about others' behavior influence cooperation rates when labeling the prisoners dilemma a "community game" rather than a "Wall Street game" or "stock market game" (Kay & Ross, 2003; Liberman et al., 2004; Ellingsen et al., 2012).[9]

2.3 Controlling the experimenter-participant interaction

To reduce experimenter demand effects studies usually aim to control and minimize participant-experimenter interaction. This may be in the form of instruction delivery, anonymous decision making, or by avoiding certain participant pools.

Instruction delivery: Controlling the role and presentation by the experimenter helps limit claims of experimenter demand. Perceived social pressure from the experimenter can be reduced by making the experimenter less salient in the study, and potential inference on experimenter demands can be reduced by following similar scripts for all treatments.

The classic and most popular method of delivering instructions is to distribute a written copy of the instructions to each participant, and then having an experimenter read these aloud. A substantial advantage of this procedure is that it establishes common information and makes clear that other study participants are making similar decisions. However, it has been suggested that, facial expressions, gestures, pitch or tone of voice might subconsciously convey desired behavior to participants (for example, Ortmann, 2005). Options with less potential bias involve video-recorded instructions (which can be included for review); or having the individuals conducting the experiment be unfamiliar with the purpose of the study, the hypothesis, or the treatment status of participants.[10]

Increasingly study participants receive instructions by reading these on their own, either on paper or on the computer screen. While eliminating the possibility of biased gestures, this does have the important drawback that instructions are no longer common information, and there may be uncertainty on whether all participants are in the same study.

Whether instructions are delivered by a human following a script, read onscreen or pre-recorded, participants always have an opportunity to ask questions. To maintain control and homogeneity such questions can be answered in private (though this differs between types of experiments, for example, privately screened and then publicly announced questions are often used in experiments on games where common knowledge is important).

As evidence on the potential role of the experimenter, there is little evidence that gestures and pitch can shift behavior. Bischoff (2011) incentivize actors to try to induce treatment effects (without over-acting) in a solidarity game (Selten 1998) and find little responsiveness. Nonetheless

[9] A set of dictator-game experiments by Dreber (2012) find very little sensitivity of giving to different neutral and non-neutral frames.

[10] This is referred to as "double blind" in medical research. Please note that this terminology is very different from that of economics, where double blind refers to neither the participants nor the experimenter being able to link specific decisions to the individual making them.

it is possible that the experimenter's mere presence can influence behavior. Cilliers (2014) find that the presence in the laboratory of a "silent white foreigner" in a lab-in-field dictator game experiment in Sierra Leone distorted participant choices in the direction of more generosity.

Anonymous decisions: Anonymity helps minimize the interaction between the experimenter and the participant and thus reduces temptation to deviate from the true preferred choice to please the experimenter. Anonymity is addressed both in the elicitation of decisions and in the experimental procedures.

First, recorded responses can be (and almost always are) anonymized, recorded only by an anonymous participant identifier. It is usually a requirement of institutional review boards (IRB) that data will be stored and shared in this way. Participants are prominently informed that their responses are anonymized, both in the process of obtaining informed consent and sometimes in the experimental instructions themselves. Studies may even ensure that decisions are double blind in the sense that no individual can identify who made what choice.

Second, typical laboratory setups are designed with anonymity in mind. Participants are seated in screened booths, make choices via computer interfaces and are paid in private. This serves to alleviate concerns about observation by the experimenter or by other participants. In interactive strategic experiments, participants typically interact through a computer interface and are identified only by a player number that is not linked to their location in the room.

Online experiments on platforms such as *Amazon Mechanical Turk* can achieve high degrees of anonymity, since usually it is impossible for participants to be personally identified by the experimenter or other participants. For sensitive outcomes, researchers have gone further, deliberately inserting noise into measurement to make it impossible to identify a given participant's response with certainty. See for example, Karlan (2012), Fischbacher (2013), or List (2014).

There is some evidence that anonymity may influence behavior in a manner that is consistent with experimenter demand. An influential study by Hoffman (1994) conducted "double blind" dictator game experiments in which the experimenter could not identify how much a participant gave, and found that giving decreased. Barmettler et al. (2012), exploiting more subtle anonymity treatments (and controlling for other design differences), find no effect of anonymity on dictator, ultimatum or trust game behavior. Loewenstein (1999) suggests that by emphasizing that choices are anonymous Hoffman (1994)'s experiments might have implied to participants that selfishness was expected. Key in drawing inference on the potential role of experimenter demand is of course whether we are interested in understanding other-regarding behavior that arises in a completely anonymous setting, or whether a setting with greater observability is the environment of interest.

Participants: Finally, some participants may be more susceptible to experimenter demand than others, and it may be advisable to exclude certain groups for this reason. For example, experimenters might avoid recruiting from among their own students or colleagues (who may have

more information on the objective of the study, and may be more concerned about pleasing the experimenter).

While many early experiments were conducted with students in classrooms—or with staff at the RAND cooperation offices in some of the very earliest experiments—purpose-built laboratories are now where most experiments are conducted. Though most laboratory studies draw participants from the overall student body, it is advisable to avoid participants with prior experience of the particular experiment, those who may be more susceptible to hypothesis guessing (for example, psychology students who are used to debriefing at the end of a study), or those with topic specific knowledge (for example, advanced economics students). Finally, to reduce hypothesis guessing one may wish to mask the names of the researchers conducting the study, and thereby their potential research interests. While there is evidence that the response to treatment differs by participant characteristics, there is no evidence to suggest that such differences result from experimenter demand varying across populations.[11]

2.4 Documentation

Transparent and precise documentation of all participant-facing steps in a study permit assessment of the potential role of experimenter demand effects. Clear documentation helps readers, referees and editors assess whether cues in the instructions or interface may be driving the results. Importantly, documentation also allows for replication and robustness checks, and thus of assessment of potential experimenter demand.

A full account of the interaction in the experiment should include all material participants see (instructions, computer screen shots, survey measures, etc.); a description of procedures; and a script detailing all statements made in and across treatments of the study.

The minimal requirement for review of a manuscript is that "representative" treatment language be provided for the instructions—however more extensive documentation is preferable. Information on all treatments should ideally be included with treatment-specific changes presented side by side (instructions and screen shots). Transparent presentation of treatment changes allow readers to assess for themselves whether experimenter demand might drive the observed comparative static.[12]

---

[11] For online studies, de Quidt et al. (2017) find very similar responses to explicit experimenter demand across experienced (Amazon MTurk workers) and less-experienced populations (respondents to an online political panel survey).

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

3 Masking the hypothesis

As noted in our introduction to the chapter, most economic experiments are designed to test hypotheses about qualitative, causal effects on economic behavior. That is, the aim of the experiment is to identify the ensuing effects from an experimentally controlled variable X on a choice behavior (or outcome) Y. Constraining participants' ability to infer this experimental hypothesis involves limiting their ability to understand that X is the independent variable, that Y is the dependent variable, or that X is predicted to affect Y.[13]

Below we outline ways in which experimental design choices may reduce hypothesis guessing.

3.1 Masking the independent variable

How the design identifies the counterfactual effects of the independent variable is typically mentioned at the start of nearly every paper's experimental design section. Though combinations are possible, there are two main choices for identification:

In a **within-subject design**, identification is achieved by asking the same participant to make distinct choices y_A and y_B as the independent variable X is changed from x_A and x_B, under the assumption that choice behavior is stable across the different decisions. This allows the direct treatment effect to be observed for each subject, $\Delta y^i := y_B^i - y_A^i$. From this point population averages are then used to diagnose the average treatment effect, $\Delta y_{Within} := \frac{1}{N}\sum_{i=1}^{n} \Delta y^i$, and the sign of said effect.[14]

Alternatively, in a **between-subject design** each subject is exposed to a single treatment (either x_A and x_B, with observation of each participant's choice within only one of the treatments). Given N_A subjects under the controlled variable x_A and N_B under x_A, the identification relies on random-assignment to identify the average treatment effect, $\Delta y_{Between} := \sum_{i=1}^{n} \frac{y_B^i}{N_B} - \frac{y_A^i}{N_A}$, and its sign.

Charness et al. (2012) provides an in-depth discussion of the tradeoffs involved between within and between designs, across multiple dimensions, and concludes that each type of design has its merits. The choice between the two is affected by the lack of order effects and the reduced potential for demand effects in a between-subject designs against the greater statistical power in a within-subject design.

With respect to experimenter demand, the advantage of a between-subject design is that each participant is only exposed to one treatment environment. As such, the participants are not provided with clues on the independent variable, and are not able to infer the treatment they are

[13] Of course, participants do not need to know the other treatments to come up with conjectures about what might be expected in their treatment. To bias the experimental results, those conjectures would need to interact with treatment.
[14] More sophisticated econometric approaches can be used to control for other observables and make inference. Our simpler presentation is designed to make clear the key differences in identification.

assigned to.[15] In contrast, in a within-subject design, each subject is exposed to multiple treatments, where the independent variable is revealed through the experimenter's explicit manipulation of it.

Without direct knowledge of the treatment variable, between-subject designs make it difficult for participants to guess what, if any, parts of the experimental environment are being varied between treatments. For example, is it the structure or scale of incentives? The framing? The number of other game participants? And in what "direction" does the variation occur? Are they in the high-stakes or low-stakes treatment?

The concern for experimenter demand has led to between-subject designs becoming the preferred identification strategy. However, there is still little in the way of direct evidence that demand-effects in within-subject designs are prevalent—perhaps because additional experimental techniques are used to mitigate demand effects in within-subject designs.

Three different techniques are used to mitigate and control for experimenter demand in within-subject designs. The first is to use "progressive revelation" of treatment information as the experiment proceeds. Participants are given as little information as possible (subject to avoiding deception) about the treatments they will encounter in later parts of the experiment. For example, participants are initially told that they will face a "series of decisions," and how their compensation will be determined, but information about changes to the environment are provided only as the experiment progresses.[16]

Second, it is common in within-participant experiments to change the order of treatments between participants. Where one subject is given treatment A followed by B, another is given B first and then A. In this way, order effects—caused by experience, wealth effects, experimenter demand effects etc.—from the within-subject design can be ruled out as driving the difference between the A and B treatments.

When the task ordering is varied across subjects and treatment information is revealed progressively, the design effectively becomes a combination of a within and between. That is, the data restricted to just the first task/decision corresponds to that from an equivalent between design, while the data from the subsequent task provides within-subject variation. Similarity in the measures across the between and within components is typically interpreted as evidence for the absence of demand effects. In the event that the design does not permit reversal of order, it is advisable to order elicitations such that it minimizes bias. For example, the initial elicitation should be of choices that are thought to be of greatest importance and potentially most sensitive to experimenter demand, while the choices elicited last should be those that potentially drive

[15] In contrast to the standard between-subject design, Levati (2011) examine behavior in either trust or dictator games (with between-subject assignment) where subjects are informed on the alternative game that other participants will play, in what they call a "hybrid design." They find substantially different behavior from standard results in the literature. While they argue that the provision of information on other treatments in a between-subject design helps subjects interpret the environment and question relatively (thereby increasing the validity of the *relative* results), this reasoning seems confounded with experimenter demand.

[16] By way of evidence that this can have an effect, Burks 2003 find that trusting behavior is lower in trust games when participants know in advance that they will play both game roles during the experiment.

experimenter demand for other elicitations or that are particularly robust to such effects, for example, questions about gender should be held until the end of a study.[17]

A third and less frequent technique for reducing experimenter demand in within subject designs is to insert time gaps or filler/control questions to make the within-study comparison between treatments less salient. Of course, such control questions may introduce new biases if participants perceive them as informative about the hypothesis. Roux (2014) study this question in the context of a Cournot oligopoly experiment. They find no discernible influence of control questions on choices, and importantly this holds whether or not participants were explicitly told that the control questions were randomly generated (and therefore presumably less informative about the hypothesis).

While experimenter demand effects are thought to be smaller in between-subject than within-subject designs, the existing evidence suggests that participants generally fail at predicting the true hypothesis in both designs. Lambdin (2009) replicate three classic experiments (the child custody experiment of Shafir 1993, the Asian Disease experiment of Tversky 1981 and the marbles lottery of Tversky 1989) using both between and within designs. At the end of the sessions they ask participants to guess the experimental hypothesis. Choices were very similar in the between and within implementations, with most participants failing to guess the hypotheses. Accuracy was 7 percent and 3 percent in the child custody and marbles experiments, and somewhat higher at 32 percent for the Asian Disease experiment.[18]

3.2 Masking the dependent variable

While it is commonplace to mask the independent variable, fewer studies attempt to mask the dependent variable of interest. If the experiment collects multiple outcome variables the participant may be in doubt as to which is the primary variable. Sometimes the experimental design will require this as a matter of course. For example, an experiment on risky choice may involve multiple decisions, some of which contain the treatment manipulation, while others measure background preference information. Of course, increasing the number of decisions participants have to make risks reducing the attention paid to the key choices of interest.

An example of this design strategy comes from Abbink et al. (2009), who introduce the "joy of destruction" game in which participants can destroy each others' endowments. For fear that this game played in isolation might induce destruction by experimenter demand, the choice is

[17] Studies that examine how men and women differ in behavior go to great lengths to remove references to gender in the study (see for example Niederle and Vesterlund, 2007). This becomes particularly tricky when there is a need to reveal the gender of an opposing player, perhaps best masked by showing a photo of the opponent (for example, Babcock et al, 2017) or by presenting a recorded greeting by the opponent (Bordalo et al, 2017).

[18] Readers are also referred to Alcott and Taubinsky (2015) who use a post-survey questionnaire to ask subjects what they thought the intent of the study was. The results from the survey indicate "substantial dispersion in perceived intent", though it should be noted that many subjects did guess the correct hypothesis (see Table A.4). Moreover, they also use the Snyder (1975) "Self-Monitoring Scale" to measure subjects' responsiveness to experimenter demand, where they do not find any correlation between this measure and the treatment effect.

embedded in a real-effort task through which the endowments are earned.[19] While participants may believe the experimenter is focused on the real-effort task, the real hypothesis is related to the ancillary task.

Related to the masking of the dependent variable is how the dependent variable is presented to subjects. Experimental design should take care that the form of elicitation does not signal a hypothesis or appropriate behavior. A particular issue here is the use of strategy methods that condition the dependent variable on other features of the environment. For example, Zizzo (2010) conjectures that use of the strategy method in public good games might increase rates of "conditional cooperation" by providing an explicit channel for participants to condition their choices on others'.

Even in settings where strategy methods are ecologically valid, the potential for experimenter demand through the elicitation can lead to altered designs to mitigate the effect. For example, Echenique et al. (2016) worry enough about demand effects in a centralized market game that they move the environment away from the ecologically valid strategy method (stating a preference) toward a direct method of eliciting a sequence of choices. This being said, there are many cases in which responses to strategy method and direct response elicitations have been shown to be the same (for example, Muller et al., 2008; Brandts and Charness, 2011).

4 Evidence on best practice adoption

In the above we discussed how best practices can mitigate concerns about demand effects, in this section, we show that these best practices are in fact prevalent in the profession. To make this claim we rely on data collected from two populations of experimental papers published in the past five years. Our first sample examines all experimental papers published in top general-interest economics journals. Our second sample examines all papers published in the top field journal for experimental research, *Experimental Economics*.

We assembled a set of 66 papers with an experimental component that were published in the "*Top Five*" journals between 2012 and 2017 (the *American Economic Review (AER)*, *Econometrica (ECMA)*, the *Journal of Political Economy (JPE)*, the *Quarterly Journal of Economics (QJE)*, and the *Review of Economic Studies (ReStud)*). These 66 journal articles were found by examining all published refereed work in the respective journals (from Thompson's *Web of Science*) and reading abstracts and design sections.[20] To this we added a sample of 179 papers with novel experiments published in *Experimental Economics*, the top field journal.[21]

---

[19] Abbink et al. (2009) note we "avoid the experimenter demand effect by embedding the destruction choices into a much more cumbersome task, with which the subjects earn their endowments."

[20] For the period, our sample contained 1,830 papers: 826 in the *AER*, 350 in *ECMA*, 174 in the *JPE*, 211 in the *QJE*, and 269 in *ReStud*. The sample of 66 papers were assembled by the paper's coauthors and break down as: 20 in the *AER* (2.4 percent); 13 in *ECMA* (3.7 percent); 7 in the *JPE* (4.0 percent); 9 in the *QJE* (4.3 percent); and 17 in *ReStud* (6.3 percent).

[21] In total we examined 195 published papers, where the journal averages 36.4 papers per complete volume. We then excluded 16 papers that did not contain original data; primarily surveys, editorial notes, and meta-analyses.

The complete sample of 245 papers were then coded by a research assistant according to their design features, sample population, and the reporting of the instructions. The results from the coding exercise are summarized in Table 1.

Table 1: Design Characteristics of recently published experimental papers

	Top Five		***Exp. Econ.***	
	Proportion	*N*§	Proportion	*N*§
Design:				
Between-subject*	59.3%	59	89.2%	139
Abstract Frame	89.4%	66	96.1%	179
Blind	83.3%		94.4%	
Incentivized	90.9%		99.4%	
All the above	45.8%	59	84.2%	92
Setting:				
Classroom	4.5%	66	4.6%	179
Lab	68.2%		84.4%	
Lab-in-the-Field	16.7%		7.3%	
Online	12.1%		2.2%	
Reporting:				
Instructions Available	86.4%	66	82.1%	179
-for all treatments	65.2%		62.6%	
-with clear language changes	56.1%		62.0%	
Progressive Revelation†	75.0%	24	53.3%	15

Note: §- *N* denotes the relevant number of experimental papers in the reported set of journals where we were able to definitively code each variable.

∗-Excludes papers with both between- and within-subject identification.

†-Conditioned on having some within-subject identification of the main effects.

Providing evidence for the idea that between-subject designs are the "norm in experimental economics" (Camerer 2003, p.41) the first row in Table 1 indicates that a clear majority of published experimental papers use a between-subject design. This choice represents the vast majority of designs for papers published in *Exp. Econ.* (approximately 90 percent), and while greater use is made of within-subject designs in the *Top Five* sample, the majority (60 percent) is between subject.[22]

When within-subject designs are used, progressive revelation of information is common. For the 24 within-subject-design papers in our *Top Five* sample, three-quarters used progressive revelation. For the smaller sample of within-subject papers published in *Exp. Econ*, progressive revelation is also common, with eight of fifteen papers.

Similarly clear is the choice over framing, where abstract frames are the norm. Papers using abstract experimental frames are common in both the *Top Five* and *Exp. Econ* samples—representing 89 and 96 percent of papers, respectively.

[22] A number of papers were coded as having "other" designs. These are typically papers with between-subject designs but where additional within-subject identification is also used.

In controlling the experiment-participant interaction, very few papers use double-blind protocols. Just ten papers from the 179 in the *Exp. Econ.* sample are double blind, and none of the *Top Five* papers are. While not double blind the large majority are single blind. 92 percent of the surveyed papers are at least single-blind, where subject anonymity is again more common in the *Exp. Econ* sample than in the *Top Five* (94 percent compared to 83 percent). In a similar vein, less than five percent of the papers we coded were conducted in a classroom setting, with the majority being laboratory studies.

Turning to incentives, a substantial majority of *Top Five* papers (91 percent) and the overwhelming majority of those in *Exp. Econ* (99 percent) report on experiments in which choices were incentivized. While we recorded whether choice is incentivized, we did not collect information on the exact marginal incentives in our samples.

Echoing the design results there are strong norms for reporting in the profession. More than eight-in-ten papers in our sample post representative instructions in the published paper. Furthermore the majority of these studies provide blanket inclusion.

This rate is similar in both the *Top Five* and *Exp. Econ* samples. Where the two samples begin to differ is over the provision of clear language changes. Almost three-quarters of the *Exp. Econ* sample papers provide clear information on the language changes across treatments, where just over a half of top five papers do.[23]

To summarize then, our examination of the literature makes clear that experimental economics has clear design norms, where many of these conventions help reduce experimenter-demand effects. Our examination of the top-field journal for experimental research indicates that 84 percent of published papers exhibit *all* of the following features: incentivized choices in an abstract frame, at least single-blinding for the subjects and a between-subject design. There is less of a norm for the *Top Five*, where such papers represent only 46 percent of published work. However, this is due to slightly greater use of within-subject designs. Allowing for within-subject designs with progressive revelation alongside between-subject, such papers represent 68 percent of *Top Five* papers.

5 Measuring and Assessing Experimenter Demand Effects

Even when we do adhere to the best design practices, and make it easy for readers to see where language has changed, readers might still posit that the treatment effects result from experimenter demand. Additionally, some research questions will not permit the use of all of the best practices we have recommended. Where a plausible argument exists that the estimated effect could be driven by demand there are still options, albeit costly, to refute this. The first is to rerun the study with the suspect part of the instructions removed or altered. If feasible, such robustness checks are typically successful in refuting the claim of bias. A second and more novel approach is instead to

[23] This is not driven by the greater use of within-subject designs. Looking just at *Top Five* papers with a between-subject design, 52 percent of papers provide treatment specific documentation.

measure and bound the size of the demand effects. In this section we describe recent work that has been done in this vein.

Bischoff and Frank (2011) study a solidarity game (Selten 1998) in which participants are matched in three-person groups and each must throw a die to determine whether the individual wins €5. The price is secured with a roll of 1,2,3, or 4, and no price secured with the roll of 5 or 6. Prior to the die roll participants must decide whether and how much of their potential €5 they will distribute to other group members in the even that they do not win the price. A professional actor delivers the experimental instructions in two different ways, attempting deliberately to induce high contributions or low contributions respectively. While they do not find a significant difference in behavior between the two groups, the procedure can assess sensitivity of the particular study in question.

Tsutsui and Zizzo (2013) construct an individual measure of demand-susceptibility at the end of an experiment on group status and trust. They present participants with a sequence of choices between lotteries, of which some are dominated. The dominated lotteries are labeled with "it would be nice if you were to choose" and a smiley face. Participants who chose more dominated lotteries are considered more susceptible to demand. They do not find a correlation between this variable and behavior in their games.

A general approach is proposed in de Quidt et al. (2017) for measuring and bounding experimenter demand. They argue that plausible bounds on the influence of demand can be constructed by deliberately manipulating participants' beliefs about the experimental objective beliefs using "demand treatments." In the simple case of experiments on ordered actions (for example, effort) this amounts to signaling to the participant that the experimenter wants them to do more or less of the action – sufficiently persuasive signals identify an interval containing the demand-free action. They provide theoretical conditions under which this approach is justified, and evidence from eleven classic experimental tasks. They consider both "strong" and "weak" demand treatments - strong treatments tell them they will "do us a favor if" they do more or less of the action, weak treatments tell participants "we expect that" they will do more or less.[24]

Response to the strong demand treatments is substantial, generating standardized bounds on "true" choices that are 0.6 standard deviations wide. This finding demonstrates the *potential* for large experimenter demand effects in the worst case, as indicated by participant's willingness to change their choices to "do the experimenter a favor."[25] In the weak demand treatments---which still convey a stronger signal of the hypothesis than most experiments---the effect is more muted, at around 0.15 standard deviations. The authors conjecture that the weak treatments are a plausible upper-limit on a magnitude of inferred demand effects suggesting demand biases in typical experiments are probably small.

---

[24] Each participant was presented with a single classic game or preference elicitation (for real or hypothetical $1 stakes), with a demand treatment included in the instructions.

[25] A within-subjects application of the treatments finds that the overwhelming majority of participants respond to these manipulations in the direction demanded. This supports the usual notion of demand effects as being driven by participants' efforts to "help" the experimenter, rather than to oppose them. We note that the best practices that we propose are designed to guard against either motive.

Such approaches allow experimenters to construct *quantitative* bounds on treatment effects (by combining the bounds estimated from different treatment groups), as well as to assess the robustness of *qualitative* effects, for example by asking whether the bounds identified contain zero or allow for effect sign-reversals. In an application to treatment effects, de Quidt et al. (2017) show that the qualitative finding of a positive effort response to incentives is robust to even the strong demand treatments.

6 Conclusion

The vast majority of experimental studies aim to identify a qualitative response to treatment. We are aware of no examples where a qualitative finding in a best-practice experiment has been shown to result from experimenter demand. This lack of evidence is likely a result of the care involved in designing and conducting experimental studies. In reviewing the best practice adopted by the profession, we note how many of the most-common design features in experimental studies are motivated by the desire to mitigate experimenter demand. Pushing the adoption of these features as norms is the aim for internally and externally valid results; the desire to secure results that are easily replicated; and the fear of being accused of inference resulting from experimenter demand.

The logical conclusion of techniques to reduce participants' inferences is to make them unaware that there is an experiment at all through a natural field experiment (Harrison 2004). There is mixed evidence that this can lead to differences in choices (Winking 2013, cf. Fessler 2009), and it is difficult to know whether the change in behavior is driven by demand or by a change in the meaning of the treatment or setting. Importantly many experimental designs do not have a natural field experiment equivalent, and the causal inference permitted by the laboratory is more difficult to secure in field studies. Both laboratory and field studies are needed to help advance our understanding of human behavior. Importantly, the potential role of the experimenter in many cases mirrors that of observers in the field environment of interest, be it the auctioneer in an auction, the fundraiser soliciting funds, or the knowing bystanders of most of our human interactions.

The evidence for bias in designs that purposefully induce demand effects clearly indicates why such effects are taken seriously by the profession. These results point to the importance of good design, where the potential for experimenter demand can be best addressed by thoughtful choices at the initial stage of the project.

References

Abbink, Klaus and Abdolkarim Sadrieh, "The pleasure of being nasty," Economics Letters, Dec 2009, 105 (3), 306–308.

Amir, Ofra, David G. Rand, and Ya'akov Kobi Gal, "Economic Games on the Internet: The Effect of $1 Stakes," PLoS ONE, Feb 2012, 7 (2), e31461.

Babcock, Linda, Maria P Recalde, Lise Vesterlund, and Laurie Weingart, "Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability," American Economic Review, 2017, 107 (3), 714–47.

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.

Bischoff, Ivo and Bjorn Frank, "Good news for experimenters: Subjects are hard to influence by instructors' cues," Economics Bulletin, 2011, 31 (4), 3221–3225.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Memory, Attention, and Choice," Technical Report, National Bureau of Economic Research 2017.

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

Brandts, Jordi and Gary Charness, "The strategy versus the direct-response method: a first survey of experimental comparisons," Experimental Economics, 2011, 14 (3), 375–398.

Burks, Stephen V., Jeffrey P. Carpenter, and Eric Verhoogen, "Playing both roles in the trust game," Journal of Economic Behavior & Organization, Jun 2003, 51 (2), 195–216.

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

--- "The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List," in Guillaume R. Frechette and Andrew Schotter, eds., Handbook of Experimental Economic Methodology, Oxford University Press, 2015.

--- and Robin M. Hogarth, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," Journal of Risk and Uncertainty, 1999, 19 (1/3), 7–42.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn, "Experimental methods: Between-subject and within-subject design," Journal of Economic Behavior & Organization, Jan 2012, 81 (1), 1–8.

Cilliers, Jacobus, Oeindrila Dube, and Bilal Siddiqi, "The white-man effect: How foreigner presence affects behavior in experiments," Journal of Economic Behavior & Organization, Oct 2015, 118, 397–414.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth, "Measuring and Bounding Experimenter Demand," June 2017, working paper.

Dreber, Anna, Tore Ellingsen, Magnus Johannesson, and David G. Rand, "Do people care about social context? Framing effects in dictator games," Experimental Economics, Sep 2012, 16 (3), 349–371.

Ellingsen, Tore, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar, "Social framing effects: Preferences or beliefs?," Games and Economic Behavior, Sep 2012, 76 (1), 117–130.

Fessler, Daniel M.T., "Return of the lost letter," Journal of Economic Behavior & Organization, Aug 2009, 71 (2), 575–578.

Fischbacher, Urs and Franziska Follmi-Heusi, "Lies in Disguise-an Experimental Study on Cheating," Journal of the European Economic Association, Jun 2013, 11 (3), 525–547.
Guala, Francesco, "On the scope of experiments in economics: comments on Siakantaris," Cambridge Journal of economics, 2002, 26 (2), 261–267.
Harrison, Glenn W and John A List, "Field Experiments," Journal of Economic Literature, Nov 2004, 42 (4), 1009–1055.
Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith, "Preferences, Property Rights, and Anonymity in Bargaining Games," Games and Economic Behavior, Nov 1994, 7 (3), 346–380.
Karakostas, Alexandros, and Daniel John Zizzo. "Compliance and the power of authority." *Journal of Economic Behavior & Organization* 124 (2016): 67-80.
Karlan, Dean S. and Jonathan Zinman, "List randomization for sensitive behavior: An application for measuring use of loan proceeds," Journal of Development Economics, May 2012, 98 (1), 71–75.
Kay, Aaron C and Lee Ross, "The perceptual push: The interplay of implicit cues and explicit situational construals on behavioral intentions in the Prisoners Dilemma," Journal of Experimental Social Psychology, 2003, 39 (6), 634–643.
Kessler, Judd and Lise Vesterlund, "The external validity of laboratory experiments: The misleading emphasis on quantitative effects," in Guillaume R. Frechette and Andrew Schotter, eds., Handbook of Experimental Economic Methodology, Oxford University Press, 2015.
Lambdin, Charles and Victoria A. Shaffer, "Are within-subjects designs transparent?," Judgment and Decision Making, 2009, 4 (7), 554–566.
Levati, Maria Vittoria, Topi Miettinen, and Birendra Rai, "Context and interpretation in laboratory experiments: The case of reciprocity," Journal of Economic Psychology, Oct 2011, 32 (5), 846–856.
Liberman, Varda, Steven M Samuels, and Lee Ross, "The name of the game: Predictive power of reputations versus situational labels in determining prisoners dilemma game moves," Personality and social psychology bulletin, 2004, 30 (9), 1175–1185.
List, John A, Robert P Berrens, Alok K Bohara, and Joe Kerkvliet, "Examining the Role of Social Isolation on Stated Preferences," American Economic Review, 2004, 94 (3), 741–752.
Loewenstein, G. (1999). Experimental Economics From the Vantage-point of Behavioural Economics. The Economic Journal 109(453), 25–34.
Muller, Laurent, Martin Sefton, Richard Steinberg, and Lise Vesterlund, "Strategic behavior and learning in repeated voluntary contribution experiments," Journal of Economic Behavior & Organization, 2008, 67 (3), 782–793.
Niederle, Muriel and Lise Vesterlund, "Do women shy away from competition? Do men compete too much?," Quarterly Journal of Economics, 2007, 122 (3), 1067–1101.
Ortmann, Andreas and Ralph Hertwig, "The Costs of Deception: Evidence from Psychology," Experimental Economics, 2002, 5 (2), 111–131.
Roux, Catherine and Christian Thoni, "Do control questions influence behavior in experiments?," Experimental Economics, Mar 2014, 18 (2), 185–194.
Selten, Reinhard and Axel Ockenfels, "An experimental solidarity game," Journal of Economic Behavior & Organization, Mar 1998, 34 (4), 517–539.
Shafir, Eldar, "Choosing versus rejecting: Why some options are both better and worse than others," Memory & Cognition, Jul 1993, 21 (4), 546–556.

Silverman, Dan, Joel Slemrod, and Neslihan Uler. "Distinguishing the role of authority “in” and authority “to”." *Journal of Public Economics* 113 (2014): 32-42.

Tsutsui, Kei and Daniel John Zizzo, “Group status, minorities and trust,” Experimental Economics, May 2013, 17 (2), 215–244.

Tversky, A and D Kahneman, “The framing of decisions and the psychology of choice,” Science, Jan 1981, 211 (4481), 453–458.

Tversky, Amos and Daniel Kahneman, “Rational Choice and the Framing of Decisions,” Multiple Criteria Decision Making and Risk Analysis Using Microcomputers, 1989, pp. 81–126.

Winking, Jeffrey and Nicholas Mizer, “Natural-field dictator game shows no altruistic giving,” Evolution and Human Behavior, Jul 2013, 34 (4), 288–293.

Zizzo, Daniel John, “Experimenter demand effects in economic experiments,” Experimental Economics, Oct 2010, 13 (1), 75–98.