Experimenter Demand Effects[1]

Jonathan de Quidt
Queen Mary University of London and Institute for International Economic Studies

31 January, 2024

1. Introduction

Experimenter demand effects are biases that arise due to participants a) forming a belief about what the researcher expects or desires them to behave or respond, and b) acting according to that belief. They are problematic because they obscure the naturalistic behavior which is typically what the researcher would truly like to observe. Experimenter demand effects are a concern in both experimental and survey-based research.

To give a motivating example, consider the classic *dictator game,* used by social scientists to measure participants' altruism.[2] In the basic design, the participant is given an endowment of money, say, $10, and asked how much they would like to give to another person (e.g., an anonymous second participant in the study, a charitable cause, or some other worthy recipient). People commonly give one-third to one-half of the endowment away.

While the most straightforward interpretation is that this reveals (considerable) altruism, one is concerned that perhaps the participant guesses she is expected to give something (why else would the researcher be asking her the question?) and does so. Had she encountered the same decision in a more naturalistic setting, she would have given zero, meaning that we overestimated her "natural" altruism, perhaps by quite a lot.

Now suppose the researcher adds a second treatment that strongly emphasizes to participants that their giving decisions are <u>anonymous</u> and will never be traceable back to them. Giving goes down. But is that because the anonymity manipulation made them more willing to act according to their true selfish preferences? Perhaps they just inferred that the researcher *expected* them to give less, and are responding to that perceived demand? It is difficult to know, but in the latter case we would again fail to identify the "natural" effect of the change in framing.[3]

Turning to an example from surveys, suppose a political pollster affiliated with an organization with known political slant is surveying voters about their voting intentions. We might be concerned that voters are more likely to report what they think the surveyor wants to hear, namely that they support that organization's preferred candidate. Once again, we fail to measure their "natural" voting behavior.

---

[1] This article draws heavily on my earlier work on the topic: de Quidt et al., (2018, 2019), as well as a keynote lecture that I gave at the 14th Nordic Conference on Behavioral and Experimental Economics. See also Mummolo & Peterson, (2019) and Zizzo (2009).
[2] The basic game structure described here was first introduced by Forsythe et al., (1994).
[3] See Barmettler et al., (2012), Hoffman et al., (1994), Loewenstein, (1999).

Experimenter demand effects relate to the broader term "demand characteristics," used by psychologists (Orne, 1962). They are related to, but distinct from, social desirability bias (Edwards, 1953), which concerns cases where participants believe some actions, behaviors, or responses, to be socially desirable (i.e., irrespective of what the researcher wants them to do). They are also related to, but distinct from, so-called "Hawthorne effects" which concern the simple effect of being observed in a study (see, e.g., Adair, 1984). Some but not all of the remedies discussed in this article may be relevant to these other types of bias.

2. Theory

de Quidt et al., (2018) provide a theoretical framework which is useful to guide the discussion. ζ denotes the "environment" in which a decision, or "action" takes place. ζ includes everything relevant to the decision: both features that the experimenter wishes the participant to condition their action on (rules of the game, incentives, participants' identity), and unintended features that that make the participant suspect a particular action is expected or desired. Assume the action a is a real number. Higher values correspond to "more" action.[4]

If the researcher is just trying to measure a given behavior, or a survey response, we call this the *natural action*, given ζ, written as $a(\zeta)$. If they want to measure a *treatment effect*, which is the change in natural action when the environment changes from ζ_0 to ζ_1, we write that as $a(\zeta_1) - a(\zeta_0)$.

Absent demand effects, measuring these objects would be straightforward. But the participants may make inference from their environment about what the experimenter wishes them to do. Then their action becomes $a^L(\zeta)$ where L denotes "latent" (unobservable to the researcher) demand. Bias arises when $a^L(\zeta) \neq a(\zeta)$. For natural actions and treatment effects we have:

$$\text{Bias for natural action} = a^L(\zeta) - a(\zeta)$$

$$\begin{aligned}\text{Bias for treatment effect} &= \left(a^L(\zeta_1) - a^L(\zeta_0)\right) - \left(a(\zeta_1) - a(\zeta_0)\right)\\ &= \left(a^L(\zeta_1) - a(\zeta_1)\right) - \left(a^L(\zeta_0) - a(\zeta_0)\right)\end{aligned}$$

de Quidt et al., (2018) develop a formal decision-making model that provides a foundation for the biases above. The bias depends on three things, each of which depends in turn on ζ.

First, the participant's belief about the *direction* of the experimenter's hypothesis or desired behavior. Does the experimenter want them to do "more" or "less" a than they naturally would? For example, in our first dictator game example the the participant was surprised to be given money to split, and suspects she is supposed to give more than she naturally would. When the experimenter decided to strongly emphasize her anonymity, she instead inferred that the experimenter expects her to act selfishly.

[4] E.g. effort on a task, giving in a dictator game, intensity of support for a candidate, life satisfaction on a Likert scale. Ordinal or categorical outcomes can often be expressed on a numeric scale, e.g. the ranking of a candidate or the probability of choosing a particular option.

Second, the *strength* of that belief, i.e. how confident is the participant in their conjecture about the experimenter's wishes? If completely uncertain (they think it is equally likely the experimenter wants more or less a), we would expect no bias. Different environments send stronger or weaker signals which will affect this degree of confidence, similarly different individuals might draw stronger or weaker conclusions.

Finally, the participant has a "preference" for pleasing the researcher. This preference can depend on the environment ζ in several ways. First, different cues might amplify or dampen their desire to please the researcher (e.g., if they feel the researcher has been kind to them, they might want to reciprocate). Second, different design features can matter, a good example is financial incentives as discussed below. Third, some participants might be naturally more altruistic toward the researcher than others.

Thus, in general, biases will be largest when (1) participants perceive a clear direction to the researcher's expectations for their behavior, (2) they perceive this signal to be strong and clear, and (3) they are motivated to follow this expectation.

3. Mitigating demand effects

It is standard to take lengths to minimize demand biases. de Quidt et al., (2019) explain how researchers can design their study with this goal in mind, and show that these techniques have been widely adopted. This section provides a high-level discussion, I refer the reader to the chapter for full details.

The theoretical framework shows us what will be successful strategies for mitigating demand biases: (1) obfuscating the direction of the researcher's hypothesis, (2) reducing participants' confidence in their inferences, and (3) dampening motives to act according to the researcher's wishes.

Doing (1) is useful because even if each participant still has a belief about the hypothesis, they are more likely to contradict and offset one another (and could potentially just become part of the error term in the analysis). When measuring treatment effects, obfuscation will tend to reduce the correlation between participants' beliefs and their treatment status, reducing the likelihood spurious findings. (2) is helpful because the magnitude $|a^L(\zeta) - a(\zeta)|$ will tend to become smaller the less confident the participant is in their conjecture (the less information they have about what they are expected to do, the less they will deviate from the natural action. And (3) is helpful for obvious reasons – if the participant does not desire to please the experimenter, their beliefs about the hypothesis become irrelevant.

Each of the recommendations in de Quidt et al., (2019) does one or more of these things. I highlight a couple of examples.

First, researchers commonly use "between-subject" experimental designs, meaning that participants only see one environment (ζ_0 or ζ_1). Take our dictator game example and let ζ_1 denote the anonymity treatment. If the participant sees both of those treatments (in a within-subjects design) she may find it easy to figure out in which one she is expected to be selfish.

If she sees only one, even if she guesses the experiment is about generosity, she will find it harder to guess whether she is in the "high-generosity" or the "low-generosity" treatment arm. This addresses both points (1) and (2).

Second, researchers often try to avoid including lots of contextual information that could inadvertently signal the hypothesis, again targeting (1) and (2). For instance, the researcher might choose call a dictator game an "allocation decision" rather than a "generosity game."

Third, researchers often incentivize choices (this is the norm in experimental economics). There are multiple motives behind using incentives. For our purposes, they can help by making it more costly for the participant to choose something different from their naturally preferred action, dampening the preference motive for pleasing the experimenter (3). For example, it is costless to promise a high donation in a hypothetical-choice dictator game, but it is more painful to do so when the game is played for real money.

4. Bounding demand effects

It is difficult to know for sure if mitigation has been successful. de Quidt et al., (2018) therefore propose an alternative strategy, which is to use "demand treatments" to *amplify* the magnitude of experimenter demand effects, and estimate bounds on the bias due to experimenter demand, and use those to assess robustness of the main conclusions. Intuitively, even if participants are guessing at the researcher's hypothesis, any resulting bias must surely be smaller than if we *explicitly tell them* the hypothesis!

Demand treatments constitute additions to the experimental instructions that deliberately push participants toward a higher or lower action. de Quidt et al., (2018) propose two portable phrasings. For the dictator game these are:

Weak: "We expect that participants who are shown these instructions will give more/less to the other participant than they normally would."

Strong: "You will do us a favor if you give more/less to the other participant than you normally would."

In the theoretical framework from earlier, a positive/negative demand treatment causes participants to take a new action, denoted $a^{+}(\zeta)$ or $a^{-}(\zeta)$ respectively. Under reasonable assumptions, we would expect that the natural action $a(\zeta)$ lies somewhere in between these bounds. Similarly, for treatment effects, the natural action lies in between $a^{-}(\zeta_1) - a^{+}(\zeta_0)$ (lower bound), and $a^{+}(\zeta_1) - a^{-}(\zeta_0)$ (upper bound).

This bounding approach can be used by researchers in their own studies. de Quidt et al., (2018) provide extensive guidance on how to do. They argue that phrasings like "weak" are normally sufficient for realistic bounds, given that this is already a strong signal relative to the typical research demand. Researchers should choose or adapt the appropriate demand treatment for their setting.

A common concern with this approach is that adding additional demand treatment arms is costly. I want to highlight two pieces of advice, and one pitfall.

First, it is often possible to implement demand treatments in a cheap within-subjects way without recruiting additional participants. This involves adding a surprise repeat of the main experimental task at the end of the study, this time with a demand treatment included. For example, a study on dictator game giving could add a surprise extra dictator game at the very end, this time with a message encouraging the participant to give more or less than they normally would. de Quidt et al., (2018) show that this works very well in practice.

Second, if bounding a treatment effect, it is often sufficient to just estimate a lower bound, or just an upper bound (depending on whether we are concerned about over- or underestimating the true effect). For a lower bound on the effect of dictator-game anonymity, we would tell participants in the control group (ζ_0) to give more than normal ($a^+(\zeta_0)$), and those in the anonymity group (ζ_1) to give less ($a^-(\zeta_1)$).

The pitfall is that it is usually *not* sufficient to just use a single demand treatment to try to measure how serious demand biases are. Consider a researcher that thinks people are giving differently than they naturally would ($a^L(\zeta) \neq a(\zeta)$). He adds a positive demand treatment and finds that the change in behavior is small: $a^+(\zeta) \approx a^L(\zeta)$. Does that mean that demand biases are small in his experiment? Not necessarily – it could simply be that participants are already so convinced that they should give more, that an additional signal in that direction makes no difference! That is, it could be that $a^+(\zeta) \approx a^L(\zeta) \gg a(\zeta)$. If, however, he had also measured $a^-(\zeta)$ and found that $a^+(\zeta) - a^-(\zeta)$ was small, he would be on firmer ground.

de Quidt et al., (2018) call the object $a^+(\zeta) - a^-(\zeta)$ "sensitivity." They argue that sensitivity is good indication of the scope for large demand effects in a given setting. They apply this methodology to a wide range of experimental tasks. They find that "weak" manipulations typically have modest effects: sensitivity is small. This is good news because it suggests that indeed in games similar to those tested, one probably does not need to be too concerned about large demand biases.

de Quidt et al., (2018) focused on classic tasks from experimental economics, mostly bounded actions and not treatment effects, and used online samples. In contemporary work, Mummolo & Peterson, (2019) conducted a similar study using a range of tasks from political science, again with online samples. Again, they found small effects of demand-treatment-like manipulations. Danz et al., (2023) recently implemented demand treatments in a set of laboratory and online experiments, testing additional classic treatment effects or comparative statics from experimental economics. Once again, sensitivity to demand treatments was low and did not reverse any of the qualitative findings in those tasks.

We therefore have evidence from a fairly wide range of tasks and settings suggesting that demand biases are often small. That does not mean that we can ignore them altogether, especially in tasks that differ significantly from those tested, in studies with more context-laden designs or action spaces. Demonstrating the need for continued caution, the "strong" treatments in de Quidt et al., (2018) do lead to large changes in behavior, which implies that demand effects *can* be large and we need to be on our guard.

5. Conclusion

In this article, I provided a definition of demand effects and sketched a theoretical framework that has been used to understand and study them in the past. I highlighted two alternative approaches to dealing with demand effects: mitigation, and bounding. I briefly described the evidence from implementing bounding approaches, which gives cause for cautious optimism. It does not appear that experimental social science is run through with large and unpredictable demand effects in general. Still, we need to be on our guard, and take the measures appropriate to our own study settings and designs.

6. References

Adair, J. G. (1984). The Hawthorne Effect: A Reconsideration of the Methodological Artifact. *Journal of Applied Psychology*, *69*(2), 334–345.

Barmettler, F., Fehr, E., & Zehnder, C. (2012). Big experimenter is watching you! Anonymity and prosocial behavior in the laboratory. *Games and Economic Behavior*, *75*(1), 17–34. https://doi.org/10.1016/j.geb.2011.09.003

Danz, D., Lepper, M., Lezama, G., Mustafi, P., Vesterlund, L., Wilson, A. J., & Winichakul, K. P. (2023). Simon Doesn't Say: Minimal Qualitative Distortions from Experimenter Demand. *Economics: Faculty Publications, Smith College, Northampton, MA*.

de Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and Bounding Experimenter Demand. *American Economic Review*, *108*(11), 3266–3302. https://doi.org/10.1257/aer.20171330

de Quidt, J., Vesterlund, L., & Wilson, A. J. (2019). Experimenter demand effects. In A. Schram & A. Ule, *Handbook of Research Methods and Applications in Experimental Economics* (pp. 384–400). Edward Elgar Publishing. https://doi.org/10.4337/9781788110563.00031

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, *37*(2), 90–93. https://doi.org/10.1037/h0058073

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior*, *6*(3), 347–369. https://doi.org/10.1006/game.1994.1021

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, Property Rights, and Anonymity in Bargaining Games. *Games and Economic Behavior*, *7*(3), 346–380. https://doi.org/10.1006/game.1994.1056

Loewenstein, G. (1999). Experimental Economics From the Vantage-point of Behavioural Economics. *The Economic Journal*, *109*(453), 25–34. https://doi.org/10.1111/1468-0297.00400

Mummolo, J., & Peterson, E. (2019). Demand Effects in Survey Experiments: An Empirical Assessment. *American Political Science Review*, *113*(2), 517–529. https://doi.org/10.1017/S0003055418000837

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783. https://doi.org/10.1037/h0043424

Zizzo, D. J. (2009). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98. https://doi.org/10.1007/s10683-009-9230-z