# Reviewer 1

This manuscript presents three studies. Study 1 is a meta-analysis of demand characteristic experiments. Study 2 analyses ratings of motivation, opportunity and belief in the experimental hypothesis for participants provided with descriptions of experiments. Study 3 reports an experiment testing the effects of demand characteristic manipulations on facial feedback. Together, these studies are argued to support the claim that demand characteristics are not limited to response bias but can lead to placebo effects resulting from participant beliefs. Relationships between belief ratings (how much a participant believes the described hypothesis) and demand characteristic effects are interpreted as support for this claim (by contrasting with non-significant measures of motivation etc).

I welcome attention to demand characteristics and I think the meta-analysis of experimental manipulations of demand characteristics is potentially useful. However, there are substantial problems with the second two studies and with the manuscript as a whole.

First, the manuscript is not well situated in the literature. Beyond passing references, there is no consideration of existing literature on placebo effects (nor of literature on placebo in psychological experiments) nor of previous research on the imaginative suggestion effects of demand characteristics on experience, and literature is cited incorrectly.

Second, the central claim that evidence of a role for participant beliefs contradicts historical understanding of demand characteristic effects is incorrect (and contradicted in the manuscript).

Third, in each study, central claims are based on invalid statistical inferences (the interpretation of non-significant p values as evidence for the null hypothesis and the assertion of differences between parameters in the absence of statistical tests).

My review will focus on these three issues. The first might be addressed by substantial rewriting to include consideration of relevant theoretical and empirical background. I do not think it is possible to address the second claim without undermining the central claim of the manuscript; evidence of effects of belief on outcome measures do not contradict existing claims about demand characteristics. I am unaware of any theory of demand characteristics for which an effect of beliefs on outcome measures would be inconsistent. For the third issue, inspection of 95% Cis suggests there will be a lack of evidence for central claims once addressed with appropriate statistical techniques.

|  |
| --- |
| More on all of this below. |

\* p.42. Evidence that "demand effects are at least partly driven by participants' beliefs about the study they are participating in… challenges historical distinctions made between placebo effects and demand characteristics"  
  
This central claim is incorrect. That demand characteristics support beliefs about the study they are participating in has been central to the concept since its initial development (as has the idea that existing beliefs form part of the demand characteristics of the situation). It is not clear to me why the authors believe that evidence that demand effects are "at least partly driven by participants' beliefs about the study" could challenge historical conceptions of demand characteristics". Demand characteristics are defined as cues which "convey an experimental hypothesis to the subject" Orne (1962). In other words, they influence what participants believe about the experimental situation. That demand characteristics impact beliefs by definition is even made explicit on p.11 of the manuscript:  
  
"Demand characteristics are sometimes defined as any cue that may impact participants' beliefs about the purpose of the study, including instructions, rumors, and experimenter behavior (Orne, 1962)."  
  
This directly contradicts the claim later in the paper that demonstrating an effect of beliefs on demand characteristic effects challenges historical conceptions of demand characteristics. Perhaps the authors have different definitions of belief in mind for these two conflicting passages? For example, maybe the second claim is intended to be restricted to participant beliefs about whether or not they will experience a particular effect? In placebo and demand characteristics research, these are generally referred to as "expectancies". It seems from the description of belief measures this might be the intended meaning (Study 2: how much they "believed the experimenter's hypothesis"; Study 3: "whether they believed in facial feedback effects), but what is meant here by "beliefs" is not addressed in detail. In any case, the claim fares no better if "expectancies" are the "beliefs" under discussion.

|  |
| --- |
| Some quibbling here. But the general point about clarifying this as an expectancy effect is fair.  I think it’s true that expectancy effects *have* been discussed in this literature. But just because it has been discussed doesn’t mean that it’s a central theme. If one were to perform a systematic conceptual review of the literature, I bet that the majority of researchers are conceptualizing demand effects as being about deliberative changes that people make (i.e., response bias). |

Participant beliefs (or expectancies) are central to the concept of demand characteristics. Placebo effects arising from demand characteristics are controlled for routinely in tests of clinical psychological interventions. See, for example Boot et al., (2013), who argue that placebo effects are pervasive in psychological interventions and call for greater attention to expectancies in controlling for placebo effects arising from demand characteristics in tests of psychological interventions. The authors reference Corneille & Lush (2023) as an earlier example of the claim that placebo effects occur in psychology. Corneille & Lush (2023) draw on the theory (and related evidence) that demand characteristic effects can include the kind of striking changes in experience seen in response to a "hypnotic suggestion" (for example, hallucinations or delusions). The degree to which such changes are driven by placebo or imaginative suggestion mechanisms (and the degree to which these mechanisms differ) requires further unpacking (see later comment).

|  |
| --- |
| I think it’s reasonable that the reviewer concluded that we miscited Cornelle and Lush here because there is a lot of conceptual blurriness.  I understand why those authors distinguish between imaginative suggestion and placebo effects, but I don’t think their distinction (a) has been widely accepted, and (b) is actually empirically testable. There isn’t widespread agreement about the mechanisms that underlie placebo effects, so I’m not expecting there to be agreement about whether/how it differs from imaginative suggestion.  Nonetheless, this should be addressed in a revision. We will have to pay more attention to debates in the placebo literature about mechanism, introducing Cornelle and Lush as yet another way that this definition is causing conceptual blurriness. We can then turn attention back to the fact that we are assessing something very specific: whether reporting that you believe a stated effect is real moderates the extent to which demand effects arise. |

\* p.5: "Traditionally, theorists have conceptualized the effects of demand characteristics as response biases mediated by relatively deliberate changes that participants make to their responses (Orne, 1962; Rosnow & Aiken, 1973; Strohmetz, 2008). In doing so, these theorists distinguished their ideas from conceptually similar work on placebo effects: changes mediated by the relatively automatic activation of beliefs and/or conditioned responses (Zion & Crum, 2018)."  
  
Both of these claims are contradicted by the literature cited to support them. For example, the first citation used to support the argument that demand characteristics effects have historically been conceptualized as deliberate actually makes the opposite claim (Orne, 1962):  
  
"It became clear from extensive interviews with subjects that response to the demand characteristics is not merely conscious compliance. When we speak of "playing the role of a good experimental subject," we use the concept analogously to the way in which Sarbin (1950) describes role playing in hypnosis: namely, largely on a nonconscious level. The demand characteristics of the situation help define the role of "good experimental subject," and the responses of the subject are a function of the role that is created."  
  
While it is arguably true that, historically, demand characteristics have often been interpreted as driving bias in psychology experiments, this was not Orne's position: Orne & Scheibe (1964)  
  
"The demonstrated effectiveness of demand characteristics in this or any experiment is not taken to indicate that subjects openly and willfully cooperate with the experimenter. Rather, it is likely that social cues can determine the subject's actual experience in the situation."  
  
Orne also did not, as claimed, distinguish between demand characteristic effects and placebo effects on the basis that demand characteristics are deliberate and placebo effects are not (I could also not find support for this claim in the other papers cited in support of this argument, either). Orne (1969):  
  
"The totality of these non-drug effects which are a function of the patient's expectations and beliefs in interaction with the medical procedures that are carried out, the doctor's expectations, and the manner in which he is treated have been conceptualized as placebo effect. This is, of course, analogous to the demand characteristic components in psychological studies; the major difference is that the concept of placebo component directly derives from methodological control procedures used to evaluate it."

|  |
| --- |
| I think this once again boils down to the fact that there is no conceptual clarity in this literature. We simplified things in an attempt to bring some order to a chaotic literature, but perhaps we oversimplified. We mostly focused on motivation vs. expectancy-based mechanisms, but I agree that these can be crossed with questions about whether the mechanisms are deliberative vs. non-deliberative (and how those mechanisms operate when you further break down the causal chain).  I am wary of diving too far into the conceptual messiness because it becomes unwieldy very quickly. For me, trying to interpret Orne is like trying to interpret Freud. The ideas are not organized, there is a lot of self-contradiction over the years, and some of the ideas that are introduced make the entire enterprise unfalsifiable and untestable. (E.g., Orne’s claims that there are ‘pacts of ignorance’ suggest that demand characteristics—here defined as expectations that the experimenter wants them to play naïve—are the reason why you can’t get honest participant feedback about the effects of other kinds of demand characteristics (e.g., expectations about the effect the experimenter wanted to see) |

\* P.9 "However, demand characteristics may produce both response biases and placebo effects - the latter which are theorized to be mediated by the relatively automatic activation of beliefs and/or conditioned responses (Coles, Gaertner, et al., 2022; Corneille & Lush, 2022)."  
  
Corneille & Lush do not claim that experience in response to demand characteristics is "relatively automatic". Their model is based on previous work arguing for effects of demand characteristics on experience which are not automatic, but strategic and goal-directed; these effects are only experienced as involuntary, not actually involuntary. According to Dienes & Lush, (2023), changes in experience in response to demand characteristics are driven by the same mechanisms as response to imaginative suggestion (e.g. in hypnosis). Empirical support comes from relationships between trait response to imaginative suggestion and effects in psychology (for example, mirror touch synaesthesia or ASMR; (Lush et al., 2020, 2022). Although this literature is discussed in detail in Corneille & Lush, (2023), which is cited several times, it has been entirely overlooked, as has an earlier similar claim (Michael et al., 2012).

|  |
| --- |
| Good point. I tried hard not to get into all of this conceptual messiness—but perhaps it’s unavoidable. Whether imaginative suggestion is goal directed and distinct from the also conceptually blurry definition of a placebo is outside of what I wanted to discuss, but perhaps we can make cuts to other parts of the manuscript to make room for this. |

\* p.10 "Coles, Gaertner, et al. (2022) found that participants' beliefs about facial feedback effects did not always match the hypothesis communicated; furthermore, both the communicated hypothesis and measures of participants' beliefs moderated the effects of posed expressions on emotion. This finding provides preliminary evidence that demand characteristics can produce both response biases and placebo effects"  
  
Only the authors' own research has been cited as evidence that demand characteristics can drive change in experience. This is described as "preliminary evidence", which might be taken to imply that this is the only evidence available. A reader new to the topic could come away from this manuscript with the impression that this claim is novel (apart from reference to a previous review. It is a live issue in psychology. For example, there is ongoing debate over the claim that demand characteristics can account for changes in experience in body illusions (Ehrsson et al., 2021; Forster et al., 2022; Lush, 2020; Lush et al., 2020; Reader, 2022; Roseboom & Lush, 2022; Seth et al., 2021; Slater & Ehrsson, 2022).

|  |
| --- |
| I don’t think the inferred meaning of the sentence is what we were trying to communicate. The goal wasn’t to say that this is the first paper to demonstrate that demand characteristics impact experience. The goal was to say that this paper provided some evidence that the impact of demand characteristics might be driven by *both* motivation- and expectancy-based mechanisms. |

\* The manuscript provides no indication of whether or not the authors' position differs from existing positions. For example, how does the cited model presented by Corneille & Lush (2020) differ from the model presented in the manuscript? An obvious difference is that the effects are described as placebo rather than imaginative suggestion ("phenomenological control") effects. As far as I can tell, the authors consider placebo and imaginative suggestion to be interchangeable terms. However, whether or not they are depends on which theories of placebo and suggestion effects are under consideration. On some theories (most prominently, response expectancy theory) response to imaginative suggestion and placebo operate by the same mechanisms. Switching between "placebo" and "suggestion" may be relatively unproblematic on such theories. For most theories, different mechanisms underlie these phenomena and so the terms are not interchangeable. Attention to theories of placebo and suggestion would be required to support any claims relating to these issues. It would also necessary to interpret empirical evidence which supports the claim that imaginative suggestion effects can arise from demand characteristics.

|  |
| --- |
| Yeah, I’m not really sure if there is anything we can do here to make everyone happy. The model is based off of the Coles et al. paper, where the main goal was to distinguish between motivation- (response bias) and expectancy-based (“placebo”) mechanisms. Phenomenological control was not discussed because it’s a relatively new proposed mechanism, but it’s closer to what Coles et al. would describe as a response bias.  Not yet sure the best way of addressing this issue. In some ways, it seems safest to just avoid conversations about mechanism altogether given that it’s ambitious to comprehensively map out different mechanistic explanations for how these effects could work. |

\* Throughout the manuscript, key claims are based on informal comparison of significant and non-significant tests.  
E.g., in the abstract:  
  
"Contra prior frameworks, we did not find evidence that demand effects were driven by participants' motivation or opportunity to adjust their responses. We did, however, find robust evidence that such effects are driven by participants' beliefs, as in placebo effects."  
  
This central argument (and any others based on asserting differences between slopes) are not valid. A non-significant p value does not, in isolation, provide evidence for the null hypothesis. In order to support claims of null effects, appropriate tests must be employed (e.g., Bayes factors or equivalence tests Dienes, 2014; Lakens et al., 2020). This issue is present throughout the studies. Further, any claims of a difference (e.g., in slopes - for example p 35-36 and Fig 8) must be backed by a statistical test of that difference, not merely asserted. It is apparent from the 95% CIs that for many cases here the tests are unlikely to provide such support, and that larger samples will be required for many analyses to distinguish between the null hypothesis and no evidence either way.

|  |
| --- |
| There’s a mismatch between what we tried to communicate here and what was interpreted. We tried to say “we did not find significant evidence of…” (i.e., we failed to reject the null), but it was not interpreted that way.  The idea of testing the difference in slopes is interesting. It’s obviously not the way I initially conceptualized the analysis strategy, but it doesn’t seem unreasonable. |

Other comments  
  
\* There are no power analyses for studies 2 and 3. A cut-off rule is reported for Study 2 (convenience sample) but not for Study 3.

|  |
| --- |
| There was no power analysis, and Studies 2 and 3 had the same cut-off rule because they were run at the same time. Let’s make that more clear |

\* CIs are reported for Study 2, but not Study 3.

|  |
| --- |
| We can fix this. |

\* Standard deviation and standard error are not reported in the text (SD/SE).

|  |
| --- |
| We can check this. |

\* Error bars are not identified in figure legends.

|  |
| --- |
| We can fix this. |

\* There are incorrect statistical claims e.g., p.36, lines 5-6 - a p value of .172 is reported as significant.

|  |
| --- |
| Strange. I’ll look into this. Could be a coding issue. |

\* Non-significant p-values are interpreted (with caveats) as weak evidence for an effect or as continuous measures of evidence. This is not a valid interpretation, even if followed by caveats. E.g., p.20, p.40, (also p.27):  
  
"The effects of demand characteristics also tended to be slightly more positive for in-person ( = 0.31, 95% CI [0.18, 0.45], < .001) vs. online ( = 0.09, 95% CI [-0.10, 0.28], = .373) studies; however, this difference was not significant, (1, 189) = 3.61, = .059.  
  
p.40 "Results indicated that the effect of facial poses on happiness tended to be slightly larger among participants who reported being more motivated to confirm the hypothesis, = 0.04. However, the estimation of this moderating relationship was not significant, t(472.40) = 1.86, p = .063. Furthermore, the estimation of this moderating relationship was less robust when including participants who did not correctly identify the communicated hypothesis, = 0.03, t(585.46) = 1.57, p = .117.

|  |
| --- |
| Will have to think about this more. It’s really an attempt to describe both the magnitude and significance of the estimated effect, but this reviewer clearly didn’t like that. (In my opinion, only focusing on significance is problematic, but it’s a difference in philosophy) |

# 

# Reviewer 2

Reviewer #2: This paper seeks to conduct a systematic review of studies measuring demand effects in experimental psychology. Demand effects occur in theory when research subjects infer the hypothesis being tested and behave artificially so as to help confirm that hypothesis. In the present paper, 195 effect sizes from 40 studies are examined in a meta analysis.  
  
The goals of this paper are important. Most studies seeking to empirically evaluate the presence/magnitude of demand effects use only a handful of tests. Thus the scale of this paper is the basis for the contribution.  
  
However, I have a few concerns about the quality of the meta analysis and the reliability of the conclusions drawn. In general, I read these results as showing even weaker evidence for demand effects than the manuscript suggests. I have several suggestions for how to investigate this and revise accordingly if necessary.  
  
First, it is unclear which studies are being replicated, and the details of how demand effects were tested in each are obscure. The citations in Figure 2 are illegible. The authors should generate a detailed appendix citing each study and explaining the basis of each test within.

|  |
| --- |
| Very cool idea. We could also provide a few more examples in the paper.  We could also perhaps clarify that these aren’t necessarily direct “replication studies”. They’re studies that are intentionally manipulating the stated hypothesis—often times (but not always) using paradigms that others have used in the literature. |

Second, the authors interpretation of the results is unwarranted at times. The paper states: "demand characteristics most often produce hypothesis-consistent shifts (63%), but sometimes produce negligible shifts (18%) or shifts in the opposite direction of the communicated hypothesis (19%)." However, the vast majority of the effects displayed in Figure 2 are statistically indistinguishable from zero. We should not conclude there are hypothesis-consistent shifts in 63% of cases based on point estimates alone. If the confidence intervals overlap zero, we cannot reject the null hypothesis of zero demand effects. In what percent of cases can an effect actually be detected? Of those, what is the breakdown of directionality?

|  |
| --- |
| Ah. Looks like we didn’t clearly communicate what that comes from. It’s from the estimated distribution of effects, not the actual observed effects. |

The authors find that on average, there is a modest hypothesis-consistent demand effect of 0.2 standard deviations. To place this in context, the authors should conduct an additional analysis that tests whether published conclusions would substantively change if demand effects of this size were imposed. Specifically, would the sign of the published result change, and would statistical significance be affected (assuming constant confidence intervals)? In which cases? It is possible very little would change about these papers in terms of the conclusions drawn. If so, that should lead the authors to temper their claims even further about the severity of demand effects.

|  |
| --- |
| I think we’re struggling to correctly convey ways in which our conclusions are being informed by estimated distributions of effects.  The argument isn’t about whether a typical empirical finding would be overturned by the average demand effect; it’s more about the \*distribution of true demand effects\* given that we are talking about non-fixed effects. |