Dear Dr. Coles,

I have now received two very thorough reviews of your manuscript, “Meta-analysis suggests that the effects of demand characteristics can be consequential, unreliable, and difficult to explain”, from experts in the field. Reviewer 1 identifies himself as Olivier Corneille. I also independently read the manuscript before consulting these reviews. Based on these reviews we believe that the work is very promising and has the potential to make an important contribution to the literature on demand characteristics, but some revisions are necessary before it is suitable for publication. I detail the main points below.

1. Both reviewers believe you need to constrain your inferences to be more closely tied to the specific demand characteristics you are examining – studies in which the researchers’ hypotheses are explicit. As the reviewers note, however, demand characteristics may have less influence when the hypotheses are obvious and more influence when they are implicit. You do acknowledge this possibility but I think you need to front it more fully and calibrate your conclusions to what you have actually studied.
2. Both reviewers also believe there is not much basis for arguing that DCs are unreliable (as opposed to heterogenous). I agree and think you should modify your title and text accordingly.
3. Reviewer 2 recommends dropping the participant study. It is not all that clear what to conclude from it and it may detract from the central meta-analysis. I am open to your keeping it in the paper if you think you can provide a strong rationale for doing so. However, if you are to keep it, I think you need to present as a second formal study (with more details about the sample and its recuirtment) following the meta-analytic study. As it stands, the description of it breaks up the flow of the meta-analytic presentation and makes the paper hard to follow.
4. Somewhat relatedly, the paper is quite long and discursive (with quite a few footnotes). Please trim it as best you can. Possibly some parts might be moved to the supplement.
5. Your meta-analysis is very well conducted and presented. Nonetheless, I had a number of questions (mostly minor ones). I present these page by page.

p. 12. How many effects were based on main effects and how many on interactions?  
p. 13. What standardizer was used when computing d/g for within subjects studies?  
p. 15. You mention that language was not a basis for exclusion. How many non-English papers were included and were there language/culture effects?  
p. 27. It would be helpful to provide I-squared estimates. For multilevel MA these can be obtained using Harrer & Ebert’s dmetar package ( <https://dmetar.protectlab.org/reference/mlm.variance.distribution> ).  
p. 26. In terms of heterogeneity, yes, there’s quite a bit of it but in general the positive effects are stronger than the negative ones (e.g., none of the negative effects are individually significant whereas quite a few of the positive ones are). Of course, we don’t know for sure, but one reading of that is that there may be a group of studies where DCs have a positive effect and another where they do not. That is, the latter don’t necessarily reflect negative effects but rather no effects randomly distributed around zero.  
p. 27. Suggest not using the word “drastically”. It’s a bit subjective and probably an overstatement.  
p. 27. Should the prediction interval be -.46 to .89?  
p. 28. Should the ds in the text be gs? More generally though, there’s no need to report these stats in the text if they’re already in the table.  
p. 33. Is this PET or PEESE?  
p. 36. First paragraph. Is there a difference between (a) and (b)?  
p. 36. Suggest replacing “mid 1900s” with mid 20th Century.

In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all necessary copyright permissions have been obtained. This may be the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at [editorialoffice@collabra.org](mailto:editorialoffice@collabra.org).

If you choose to revise, we ask that you submit your revision within six weeks.

Sincerely,  
Louis Moses

Review signed: Olivier Corneille (I always sign my reviews).

I previously had the opportunity to review this manuscript when it was submitted to another journal. At that time, I provided many positive and encouraging comments while also raising some concerns. I would like to commend the authors for their careful attention to the feedback they received on that prior submission. Their receptiveness and thoughtful engagement with the reviewers’ comments have significantly strengthened the manuscript.

In the following, I offer one final main comment and two very minor ones for the authors’ consideration. I want to emphasize that I do not intend to impose my perspective. Rather, I hope the authors will find my point and suggestion persuasive. If not, I can accept that.

Main point :

As is now acknowledged at various (but not all) places in the manuscript, this meta-analysis is concerned with a very special type of demand characteristics where the experimenter directly communicates the experimental hypothesis to the participants. This represents a large and I believe very significant departure both from regular experimental studies in psychology, and from the classic conceptualization of “demand characteristics”.  
Admittedly, demand characteristics are cues that may range from “very subtle” to “quite blatant”. However, as I noted in my previous reviews, when the experimental hypothesis is directly communicated to the participants, this fundamentally changes the social contract between the participants and the researchers, as well as inferences participants can draw regarding their role in the testing situation. Experimenters almost never communicate their hypotheses to the participants. If they do, participants may now infer that the experimenter is incompetent, dishonest, or is attempting to test participants’ level of integrity or conformism. This active sense-making point is key to the very concept of “demand characteristics”. For the sake of clarity, I’m borrowing here from a theoretical paper that my co-authors and I are currently writing on demand effects (and in doing so, I’m borrowing from Orne):

“ For instance, Orne (1962) noted that “(…) If, on the other hand, the demand characteristics are so obvious that the subject becomes fully conscious of the expectations of the experimenter, there is a tendency to lean over backwards to be honest.” (Orne, 1962, p. 779). This point was also stressed by Orne (2009) in his late work: “In fact, demand characteristics may be less effective or even have a paradoxical action if they are too obvious. With the constellation of motives that the usual subject brings to a psychological experiment, the ‘‘soft sell’’ works better than the ‘‘hard sell.’’ (p. 116).”

I’m willing to accept that the direct and blatant communication of the experimental hypothesis by the experimenters to the participants is an extreme, yet still relevant, form of “demand characteristics”. However, I honestly have a very difficult time conceiving that directly communicating experimental hypotheses to participants is just a minor departure from usual studies where demand effects may also (yet, differently) operate, and from the usual understanding of the term.

Needless to say, if the departure is major, then what we can learn from this particular and extreme instantiation of “demand characteristics” may or may not help us understand how demand characteristics influence effects in regular studies where experimenters almost never directly communicate their hypothesis to the participants. Likewise, the current results from the vignette studies suggesting that participants’ social influence reports are not quite useful for anticipating the effect of communicating directly the experimental hypothesis to the participants, may or may not be relevant for understanding how participants anticipate effects in the vast majority of studies where such information is not delivered.

How should we address this issue?

One option is to posit that what I (and Orne, and I believe also most demand theorists) see as a major difference does not matter at all. If so, I believe that this should be squarely stated in the manuscript, and should also be supported on conceptual, logical and/or empirical grounds.

Another option, currently applied in the manuscript, is to remain highly ambiguous on this important point, and go back and forth (and also drifting) between “effects of explicit demand characteristics”, “effects of demand characteristics”, and “demand effects”. I don’t believe that this wishy-washy option best contributes to solid scientific advances. Instead, as I’ve argued elsewhere (see references below), I believe it often brings about conceptual and empirical confusions, and slows down scientific progress.

A third option, one that I would see as clearly preferrable, is to systematically refer to what this meta-analysis is really about and that makes it so special: “effects arising from directly communicating the experimental hypothesis to the participants” and then perhaps, for short: “effects of explicit/blatant demand characteristics”. Needless to say, consistency would be important throughout the manuscript, including title and abstract.

My present recommendation may be seen as excessive or unjustified by the authors and the editor. If so, I would respect their judgment and decision. The fact is that I am particularly sensitive to questions of conceptual clarity (<https://doi.org/10.1177/108886832091132>; <https://pubmed.ncbi.nlm.nih.gov/37642084/>) and scientific writing practice (<https://elifesciences.org/articles/88654>). Here, I am just trying to be consistent with recommendations I’ve made elsewhere.

Besides this, I am very sympathetic to the authors’ conclusions. I also believe that most of them would actually apply to regular studies. In my view, however, the present meta-analysis does not speak to demand effects arising in regular experimental studies. To repeat, this is because experimenters do not communicate their hypotheses to the participants in regular experimental studies.

Minor:

Title and elsewhere: I still do not understand what the authors mean by “unreliable” (as opposed to “heterogeneous”) effects.

p. 36: “If one agrees that such a characterization is problematic, we argue they face an uncomfortable observation: our meta-analysis suggests that this characterization also currently applies to experimental psychology. » I did not get what this means. Why do the authors make claims about what applies to “experimental psychology” in this sentence?

To conclude: I think that the authors came up with a very interesting paper. I also realize that it probably required a huge amount of work. I believe that this work is useful and has the potential to attract attention. I’ve stated my main concern above several times by now. Hopefully, I have now communicated it more clearly than I did before. Alternatively, my analysis may be wrong or my concern unjustified – and so this probably should be clarified. In any case, I think that it is wise at this stage to leave this point to the consideration of the authors and editor. I would not want to repeat myself again, nor do I want to impose my views onto others. I wish the authors the very best with this manuscript, and, more generally, with the important and challenging research they’ve engaged in. I won’t accept to provide comments on this or a revised version of this manuscript again (as I feel I have now said all I believe should be said). Therefore, if the authors find it helpful to reach out to me to discuss or clarify any of the above points, I would see no problem if they choose to do so.

I quite enjoyed reading this paper. As someone raised on demand characteristics (and meta-analysis) as a graduate student a few decades ago, I approached this paper with some trepidation given the title. However, the authors did a good job of making most of their case. But I don’t agree with their title. What follows are a few comments, substantive and otherwise.  
Page 4. Stylistic thing but the latest version of APA Style calls for multi-authored papers to be referenced as et al. (e.g., Coles et al., 2022). I also suggest deleting use of the word “significantly” as in statistically significant (or add the word “statistically” before significant).  
Page 5. I don’t think there is an Orne 1958. I assume the authors mean here Orne 1959.  
Page 6. “Corneille and Lush (2023), THESE CHANGES could…”  
Introduction is an accessible summary of demand characteristics and its merging with the Rosnow model. Well done.  
Page 7. I’m not familiar with the PICO framework so maybe a very brief explanation here would help.  
Initially I reacted negatively to limiting the search to papers that make reference to demand characteristics, since I see demand characteristics as “a potentially boundless conceptual space”, as the authors put it, but I was convinced by their argument here.  
As an aside, my view of demand characteristics is this — Experimentation in psychology in the earliest days often followed a behaviorist tradition with the research participants being rats or rabbits (or infants). They don’t know they are in an experiment. They don’t know that experimenters have hypotheses. They don’t know that experimenters sometimes deceive their participants. But human participants know all this and, especially when the research is experimental and social in nature (vs. say a survey of consumer preferences), they are inquisitive. And that inquisitive nature can produce spurious results. Orne in the 70s referred to a two-experiment problem when participants think the experiment and its manipulations are about one thing and experimenters think the experiment and its manipulations are about something different. When those two visions clash, experimenters risk misinterpreting their participants’ behavior.  
Page 9. 1840?  
Page 11, Footnote 2. My hunch is participants might often reasonably infer what their mood should be when a researcher says that an independent variable would ‘impact mood’. Is that the third condition in Figure 1, Records Included (“Explicit demand cues do not specify direction… N = 3)?  
So 38 studies and 52 effect sizes, correct?  
Page 12. Statement about effect size index might be moved to the top of page 13.  
Page 13. The wording here is a bit confusing. If participants were told the intervention should produce mood boosting, mood boosting is coded as positive. If participants were told the invention should produce mood dampening, mood DAMPENING is also coded as positive (given participants followed the experimenter’s instruction), correct?  
And is demand characteristics all about what the experimenter tells them (given we lie to participants all the time)? A colleague and I in a study of debriefing were surprised to see how popular deception remains in social psychology research. Maybe subjects know this?  
In other words, the top dozen studies in Figure 1 (four of them by Coles et al.) produced the opposite effect than intended?  
Page 14. Again referring back to page 13, I’m confused by the group comparison. Is the point here to look for whether participants are more likely to respond differently if told to go positive than to go negative or go nowhere? But that is unrelated to coding of effect sizes, yes?  
And on page 15, there is a control group comparison for positive demand, negative demand or nil demand. These are different things?  
Page 15. I’m struggling with the research participants section, for reasons identified by the authors in the discussion. I think the paper is stronger as a meta-analysis. This is something else. While Orne (1969) talked about research participants helping researchers understand demand effects, I think he meant participants in the actual study (post-experimental questioning), not this which reminds me of the idea of simulating subjects (subjects not in the study but asked to pretend they are). That approach was tried for example in the Milgram study but didn’t work terrible well there. Simply put, I think this is a different study from the meta-analysis. And as noted by the authors, those simulating participants were not very reliable. And I don’t buy the argument about the Law of Large Numbers (page 19). More confused, disinterested participants don’t buy you much in my view no matter how many more of them you stack up. I think their inclusion deflects from the meta-analysis.  
Page 21, Back to the meta-analysis. Are the authors confusing the three levels with random effects? Because their explanation of multiple true effects vs. a single true effect is about random effects, not levels. It is true that multiple levels allows one to model different sources of variability, but that’s a different argument.  
Page 23. Anonymity is surrendered here by the reference to Cole et al.  
Page 24. Again, I would like to see discussion of the meta-analysis. Period. It gets muddled here with the participant study.  
Eleven of the studies…  
Page 25. So the overall effect is statistically significant and positive.  
Page 27. And there is statistically significant variability — at which level? Or is the comparison at the between vs. within levels? It seems to be the latter — the Q of 972.42. In other words, is it correct to conclude that there is variability in the effect sizes? And should there be reported somewhere variability at each level?  
I don’t follow the 95% prediction interval being between -.46 and -.46. That must be a typo. Because that is NO range, not a “wide range”.  
Isn’t it the case in any meta-analysis that many effects are in the appropriate direction but some are “negligible” or in the opposite direction? 63%, almost two thirds, strikes me as laudable.  
Anything to be learned by looking at those dozen studies that went the wrong way? That would be an appropriate addition to the moderator analysis. (Of course, one might also look at outliers in the positive direction).  
Page 29, To be clear, the F tests are testing what? The F for group comparison of 1.89 indicates no difference between the five categories, correct? While the F of 8.35 is a test of whether the effect of g = .16 in positive vs. control is different from zero (hence why the 95% CI does not include zero)? I think that is what the footnote is saying, but I am seeking confirmation.  
Page 33. So no evidence of publication bias, but there are outliers? Is that right?  
Having reading all that, I’m puzzled by the title. The authors are saying demand characteristics are consequential. Agreed. Difficult to explain. Okay, at the study level, yes, based on the moderator analysis. Unreliable? Because they are heterogeneous? Heterogeneous does not equal unreliable.  
Page 38. No page number for the quote here.  
Page 39. I don’t think Flake and Fried spoke to demand characteristics. They spoke to the need for better measures in psychology. Period.  
Glad the authors do not outright reject broader definitions of demand characteristics given my definition.  
Page 40. Footnote here is in an odd place, perhaps should come in the method section.  
Page 41. Twice here the authors restate their title that demand characteristics are consequential, but also unreliable, difficult to predict and challenging. Again, I disagree on reliability. What’s wrong with challenging? We do agree that the situation is pessimistic but there is always hope. Whatever the case, I found this to be an unnecessarily bleak conclusion to their thoughtful investigation.