# Editor feedback

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.

|  |
| --- |
| Part of major edit. |

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.

|  |
| --- |
| We will replace all instances of “unreliable” with “heterogeneous” |

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.

|  |
| --- |
| Will remove participant study |

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.

|  |
| --- |
| Will take a pass to |

p. 12. How many effects were based on main effects and how many on interactions?

|  |
| --- |
| Will have to manually code this and then report |

p. 13. What standardizer was used when computing d/g for within subjects studies?

|  |
| --- |
| Great question. We used the Cohen’s standardizer for between-subject designs and for within-subject designs. This is discussed on p. 13 in the updated manuscript. |

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?

|  |
| --- |
| Although language was not a basis for exclusion, we note in the updated manuscript that all records that were deemed eligible for inclusion were published in English.  In the updated manuscript, we also note how many observations were collected outside of the United States (spoiler alert: very few). |

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> ).

|  |
| --- |
| Thank you for this helpful suggestion. I-squared is reported in the updated manuscript. |

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.

|  |
| --- |
| Excellent point. Mention this in the limitations section perhaps? |

p. 27. Suggest not using the word “drastically”. It’s a bit subjective and probably an overstatement.

|  |
| --- |
| Done. |

p. 27. Should the prediction interval be -.46 to .89?

|  |
| --- |
| Yes, thanks for catching that coding error! |

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.

|  |
| --- |
| Yes, those *d*’s should indeed be *g*’s. This is fixed in the updated manuscript |

p. 33. Is this PET or PEESE?

|  |
| --- |
| These analyses are PET – not PEESE. We clarify this matter on p. 33.  We are open to also running and reporting PEESE, but we only generally see this done with the intercept of the PET model is significant. |

p. 36. First paragraph. Is there a difference between (a) and (b)?

|  |
| --- |
| We argue that these two issues – error rate and estimation bias – are related but distinct methodological issues.  E.g., consider a scenario where there is a real effect, but participants exhibit that effect more strongly when they are aware of the researcher’s hypothesis. In this scenario, demand characteristics are not leading to Type 1 or Type 2 error (point A) – but they are leading to biased estimates (point B). |

p. 36. Suggest replacing “mid 1900s” with mid 20th Century.

|  |
| --- |
| Done. |

# Editor feedback

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

|  |
| --- |
| We meet again, Olivier. :) |

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.

|  |
| --- |
| Thank you! And we would like to commend the reviewers – both past and present – who have provided such useful feedback on this challenging but interesting paper! |

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.

|  |
| --- |
| Thank you for that note, Olivier. We’ve enjoyed discussing these issues with you over peer-review. In fact, one of the main reasons we chose this journal is because we felt that the conversations with peer-reviewers was just as informative as what is actually in the paper. |

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).”

|  |
| --- |
| We do agree that the use of explicit demand cues represents a departure from the regular experimental setting in psychology – wherein we are usually worried about researchers inferring a non-communicated hypothesis.  However, we are less in agreement about how to characterize “classic conceptualizations of demand characteristics”. This largely stems from inconsistencies we’ve observed in Orne’s own work on demand characteristics.  For example, you mention that Orne (1962) argued that “If…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” (p. 779). Yet, two years later, Orne and Scheibe (1964) use obvious manipulations of the stated effects of sensory deprivation in order to conclude that the effects are not fully driven by demand characteristics. Similarly, expectations were clearly articulated to hypnosis participants in Orne and Evans (1965). Gustafson and Orne (1965) also explicitly articulated their expectations to participants.  We elaborate on this issue further in the updated manuscript. |

[Comment continues] 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.

|  |
| --- |
| [Response continues]  We are in agreement that this should not considered a ‘minor’ departure from the modal instantiation of demand effects. We note this in the updated manuscript.  In the updated manuscript, though, we also highlight a methodological challenge: explicitly manipulating cues is one of the most common ways we’ve seen researchers attempt to document the effects of demand characteristics. There appears to be a Streetlight Effect in this literature – one that you rightfully point out drastically constrains the generalizability of research on this topic. |

[Comment continues]

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.

|  |
| --- |
| [Response continues]  This is not the position we seek to advocate in the manuscript.  Our more honest impression is that (a) we don’t know if this difference matters, (b) we did not find any tests of this idea in our literature review (although it may have been missed if it didn’t meet other inclusion criteria), and (c) investigating the issue would present a variety of conceptual issues. As an example of the latter point, we point out in the updated Discussion that:  “At its broadest, demand characteristics are defined as almost any cue that may impact participants’ understanding of the purpose of the study, including instructions, rumors, and experimenter behavior (Orne, 1962). However, such a definition arguably creates a boundless conceptual space where any systematic change in a research design or setting might be considered a threat to scientific inferences.” |

[Comment continues]

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.

|  |
| --- |
| [Response continues]  We were not intending to be ambiguous in manuscript. Our goal was to (1) precisely define our operationalization of “demand characteristics” (as we do on p. 7 and 8), and then (2) use that phrase to refer to that operationalization (as opposed to always typing out “effects of explicit demand characteristics”).  This is the approach we prefer, particularly given that we remind the reader about our precise definition at several points in the manuscript (e.g., in the Title, Abstract, Methods, and Limitations section). |

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.

|  |
| --- |
| We agree with this final point and plan to make this ultra-explicit in the updated manuscript |

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

|  |
| --- |
| We changed this to ‘heterogenous’ after the editor indicated that they also prefer this terminology.  To clarify the original thinking, though, we used the phrase “unreliability” in the past manuscript because the heterogeneity is currently *unexplained*. If one cannot predict whether demand will produce a negative, nil, or positive effect, we felt that it was fair to characterize the effects as “unreliable”.  However, this isn’t a strong semantic preference – so we use “heterogenous” in the updated manuscript. |

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?

|  |
| --- |
| Good question. We discuss “experimental psychology” because that was the literature we primarily focused on. We essentially argue through a ficticious metaphor that an epistemically unsatisfying set of conditions applies to our current understanding of demand characteristics. |

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.

|  |
| --- |
| Fantastic review. |

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.

|  |
| --- |
| We’re probably going to change the title |

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).

|  |
| --- |
| Fixed (manually at end) |

Page 5. I don’t think there is an Orne 1958. I assume the authors mean here Orne 1959.

|  |
| --- |
| Fixed |

Page 6. “Corneille and Lush (2023), THESE CHANGES could…”

|  |
| --- |
| Fixed |

Introduction is an accessible summary of demand characteristics and its merging with the Rosnow model. Well done.

|  |
| --- |
| Thank you! |

Page 7. I’m not familiar with the PICO framework so maybe a very brief explanation here would help.

|  |
| --- |
| Added. |

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.

|  |
| --- |
| Thank you for being open. |

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.

|  |
| --- |
| Note earlier on that there are conceptual issues we tackled.  This is why we highlight this on p. 8.  We also dedicate a section to it in the Discussion  We now bring this issue to the forefront in the introduction though. We’ll introduce more precise language.   * So how do we go about understanding demand characteristics? We experimentally manipulate them   Way to keep the Rosnow and Rosenthal piece:   * Say that they predict heterogeneity. Sometimes positive, sometimes negative, sometimes null. We observe evidence that is consistent with this prediction. * But we don’t see any evidence that these studies, which presents a promising approach to understanding demand characteristics, are testing these mechanisms. * So the patterns [largely] conform to the Rosnow and Rosenthal predictions. But there aren’t direct tests. We provide one in the supplementary materials – but note that these analyses were consistently judged too methodologically contrived to be of any use.   But as a compromise, we are going to ask Corneille to allow us to generalize our conclusion.   * A is a [potentially minor] subset of B * A   But in the updated manuscript, |

Page 9. 1840?

|  |
| --- |
| Yes, that was the first record in our dataset (obviously irrelevant though) |

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)?

|  |
| --- |
| Correct.  We agree that there are setting where participants can reasonably infer what the researcher expects (even in the absence of explicit information about direction). But we excluded it because (a) we saw settings where it wasn’t so obvious, and (b) we sought methodologically clean tests. It’s part of the limited construct we looked at |

So 38 studies and 52 effect sizes, correct?

|  |
| --- |
| Correct. |

Page 12. Statement about effect size index might be moved to the top of page 13.

|  |
| --- |
| Not sure what statement you’re referring to |

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?

|  |
| --- |
| Correct. This helps ensure that positive values always indicate acquiescence to demand characteristics (and negative values always indicate counteracquiescence). |

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?

|  |
| --- |
| For the scope of this investigation, we are focusing on a very specific (but common) operationalization of demand characteristics, which does indeed just focus on what experimenters tell participants about their hypothesis.  Deception continues to be popular in social psychology – in part because of concerns about demand characteristics. However, we don’t want to speculate about the extent to which subjects in the studies we examined were aware of this. Studies were run at different universities, different departments, and different historical moments. |

In other words, the top dozen studies in Figure 1 (four of them by Coles et al.) produced the opposite effect than intended?

|  |
| --- |
| Correct. |

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?

|  |
| --- |
| Correct.  It is related to the coding of effect sizes because we have to standardize the direction in order to conduct analyses.  For example, imagine you are studying the effects of music on happiness – and assume that the effect is usually positive and small, *d =* .20. One possibility is that participants are more likely to exaggerate than attenuate a true effect in the presence of demand characteristics. I.e., when told that the researcher expects improvements in mood (positive demand), music now has a stronger effect of *d* = .50*.* When told the researcher expects *diminishments* in mood (negative demand), music now has a weaker effect of *d* = .10.  In this scenario, the *absolute* change produced by positive demand is larger than what is produced by negative demand (Δ*d* = .10).  Positive demand: |Δ*d*| = |.50 - .20| = .30  Negative demand: |Δ*d*| = |.10 - .20| = .10  In other words, this pattern would suggest that participants’ responses were more affected by positive demand than negative demand. |

And on page 15, there is a control group comparison for positive demand, negative demand or nil demand. These are different things?

|  |
| --- |
| Yes, we use “control group comparison” to mean a condition wherein researchers did not articulate a link between the IV and DV (i.e., a condition without explicit demand characteristics).  This is an important methodological detail to point out because many experiments on demand characteristics do not contain “control group”. Indeed, as shown in Table 1, it was quite common for researchers to compare the effects of one demand characteristics manipulation (e.g., “this music will make you *more* happy”) to the effects of another demand characteristics manipulation (“this music will make you *less* happy”). |

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.

|  |
| --- |
| We have moved this study to the Supplement. You, Reviewer 1, the Editor, and multiple reviewers at Psychological Bulletin (where the work was previously reviewed) have consistently been unconvinced of its value.  We respond to your point below – mostly in an attempt to convince you the work is still valuable in the Supplement.  First, I want to note that Orne did indeed discuss and use the “simulating subjects” paradigm quite extensively. E.g., see X, Y, and Z.  We’re not sure why you’re skeptical of the Law of Large Numbers argument. Perhaps you do not think the ratings are valid at all – or perhaps you believe that the observed variability is systematic? In that case, we agree that the argument should be rejected. However, we suspect the ratings are at least somewhat valid and that the observed variability is unsystematic. In those conditions, our understanding is that the Law of Large Numbers holds.  We make this more explicit in the Supplement so that interested readers can judge themselves. However, this work will play little-to-no-role in the updated manuscript. |

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.

|  |
| --- |
| We appreciate your point about distinguishing between ‘levels’ and ‘random effects’.  We have the distinction straight in our own heads. However, there was indeed some prose that we believe may have made things unnecessarily confusing:  “To separate variability in these true effects from sampling error, 3LMA models three sources of variability: sampling error of individual studies (level 1), variability within studies (level 2), and variability between studies (level 3; often referred to as “**random effects**”).”  What we meant to acknowledge here is that readers are probably already familiar with the idea a “random effects meta-analysis”, which is particularly focused on modeling variability at Level 3 (in addition to Level 1, of course). But we think this ironically made things more confusing and have dropped the phrase from the updated manuscript. |

Page 23. Anonymity is surrendered here by the reference to Cole et al.

|  |
| --- |
| Good point. Removed the reference. |

Page 24. Again, I would like to see discussion of the meta-analysis. Period. It gets muddled here with the participant study.

|  |
| --- |
| Done! |

Page 25. So the overall effect is statistically significant and positive.

|  |
| --- |
| Correct! |

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?

|  |
| --- |
| We did report variability at Levels 2 and 3, although we’ve additionally included the estimate of variability at Level 1.  The *Q* statistic can be interpreted similarly regardless of whether 2 vs. 3 levels of modeled. It continues to represent the weighted sum of squared differences between observed effect sizes and the fitted values under a fixed-effects-only model (i.e., no random effects). In the context of 3LMA, it is best interpreted as a measure of “a global test of heterogeneity”. We describe it as so in the updated manuscript, while also providing the estimated variance decomposition. |

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”.

|  |
| --- |
| Yes, it was a typo in the code that creates our computationally reproducible manuscript |

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.

|  |
| --- |
| We wouldn’t be so bold to say that this would be the case in \*every meta-analysis\*. E.g., a meta-analysis on extremely large studies that all examine the same non-heterogeneous effect could (in theory) produce a distribution that is non-overlapping with zero.  The interpretation of 63% is very subjective. |

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).

|  |
| --- |
| We don’t think so. Not with the amount of unexplained heterogeneity and methodological heterogeneity that we have observed. |

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.

|  |
| --- |
| You’re understanding is correct. We are glad the footnote helped. |

Page 33. So no evidence of publication bias, but there are outliers? Is that right?

|  |
| --- |
| We prefer to say that “publication bias analyses were inconclusive” because significant evidence of publication bias was observed in two analyses (PET with 3LMA; weight-function modeling with aggregated dependencies) but not others (PET with aggregated dependencies; comparison of published vs. unpublished) |

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.

|  |
| --- |
| We agree that “heterogeneous” does not equate to “unreliability” – if the heterogeneity is explained.  In this context, it is not. Sometimes you will see a positive effect, sometimes you will see a negative effect, and sometimes you will see no effect. In the absence of an explanation for this variability, we believe it is fair to characterize such an effect as “unreliable”. |

Page 38. No page number for the quote here.

|  |
| --- |
| Added |

Page 39. I don’t think Flake and Fried spoke to demand characteristics. They spoke to the need for better measures in psychology. Period.

|  |
| --- |
| True. We meant to use this citation as a nod to the broader issue of poor measurement in psychology. Hopefully this is more clear:  “and perhaps they need better measures (Flake & Fried, 2020) of the psychological mechanisms that may underlie demand effects.” |

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.

|  |
| --- |
| Agreed and done. |

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.

|  |
| --- |
| Others have also disagreed with that phrasing. We defend it, but acquiesce. |