# Editor feedback

## Editor Comment 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.

|  |
| --- |
| We thank the reviewers for reflecting upon this important issue. We address the issue in three ways.  First, in the *Introduction*, we now describe conceptual challenges with the operationalization of “demand characteristics” before motivating our interest in a specific subset of these cues: Explicit Demand Characteristics (EDCs).  Second, throughout the paper, we now generally describe our results more narrowly in terms of the effects of EDCs.  Third, in the *Conclusion* section, we added an extended discussion of conceptual challenges with operationalizing “demand characteristics” – especially in regards to generalizing to more typical contexts where the cues are implicit. |

## Editor Comment 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.

|  |
| --- |
| In the updated manuscript, we have removed all instances of the phrase “unreliable”. |

## Editor Comment 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.

|  |
| --- |
| After considering all feedback, we agree that it is best to drop the participant study. To help address the “file-drawer problem”, though, we opted to keep the results in the *Supplementary Materials*. |

## Editor Comment 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.

|  |
| --- |
| Thank you for this suggestion. We made extensive cuts to the updated manuscript, decreasing the word count from approximately 9,300 to approximately 6,500. |

## Editor Comment 5

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

|  |
| --- |
| As now reported on p. 13, 67% and 33% of coded effect sizes were based on main effects and interactions respectively. (As an aside, effect size magnitudes were similar in both cases.) |

## Editor Comment 6

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

|  |
| --- |
| We used the Cohen’s standardizer for between-subject designs and for within-subject designs. This is discussed on p. 14 in the updated manuscript. |

## Editor Comment 7

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 deemed eligible were published in English (p. 10).  We did not pre-register an interest in culture effects, of course. However, we do descriptively report which countries were studied. (As an aside, exploratory analyses did not yield evidence of moderation by country, but non-US sample sizes are small.) |

## Editor Comment 8

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 estimates are reported in the updated manuscript (p. 22). |

## Editor Comment 9

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.

|  |
| --- |
| Thank you for making this point, which we integrated into the updated *Discussion* section (p. 29-30).  To summarize our edits: we acknowledge that assumptions about the normality of the distribution of heterogeneity would lead one to expect counteracquiesence effects (what you refer to above as “negative effects”). However, we also note that this assumption could be incorrect – particularly given that significant counteracquiesence effects were only detected in 2 / 252 tests of demand effects (0 if you aggregate dependent effect sizes). |

## Editor Comment 10

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

|  |
| --- |
| Done. |

## Editor Comment 11

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

|  |
| --- |
| Yes. Thanks for catching that error! |

## Editor Comment 12

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 |

## Editor Comment 13

p. 33. Is this PET or PEESE?

|  |
| --- |
| These analyses are PET – not PEESE. We clarify this matter on p. 18.  (We are open to also reporting PEESE. However, we typically only see this done when the intercept of the PET model is significant.) |

## Editor Comment 14

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

|  |
| --- |
| We do believe 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).  We attempt to more clearly communicate this point in the updated manuscript by stating the issue in more practical terms – i.e., “that it sometimes makes an effect appear bigger than it actually is, and other times smaller” (p. 4). |

## Editor Comment 15

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

|  |
| --- |
| Done. |

# Reviewer 1 feedback

## R1 Comment 1

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

|  |
| --- |
| We meet again, Dr. Corneille. :) |

## R1 Comment 2

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 for summarizing this context.  We too would like the commend the reviewers – both past and present – who have provided such useful feedback on this challenging but interesting paper. |

## R1 Comment 3a

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 have enjoyed discussing these issues with through peer-review. In fact, one of the main reasons we chose to transfer the submission to *Collabra* is because we felt that our conversations with peer-reviewers were just as important as what we more formally discuss in the paper.  Below, we’ve divided up your comment so that we can better indicate how it shaped our revision. |

## R1 Comment 3b

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).”

|  |
| --- |
| In the updated manuscript, we now discuss this conceptual issue in a section titled “Experiments on demand characteristics” To summarize, we:   * Introduce the notion of Explicit Demand Cues (EDCs), described as a subset of the broader demand characteristics construct. * Review face valid concerns that EDC’s are non-representative of typical experiments – and that some have argued that this will cause them to be “less effective” at producing acquiescence (Orne, 2009) or have a “paradoxical” effect of producing counteracquiesence (Orne, 2009). * We use these concerns to contextualize the limitations of researchers’ tendency to use EDC’s to study demand characteristics. This, ironically, includes Orne – who, for example, reported that participants were more likely to report sensory deprivation side-effects (e.g., hallucinations and cognitive impairments) when told that “…Such experiences are not unusual under the conditions to which you are to be subjected” (Orne & Scheibe, 1964).   We revisit these concerns in the Discussion section – where we, for instance, note that our results are inconsistent with concerns that EDC’s typically produce non-acquiescence and/or counter-acquiescence. |

## R1 Comment 3c

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.

|  |
| --- |
| We agree and more explicitly incorporated this thinking into the *Introduction* and *Discussion* section (as noted above). |

## R1 Comment 3d

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.

|  |
| --- |
| This was a point that we now engage with in both the *Introduction* (described above) and *Limitations* section (described below).   * EDC’s have long been recognized as an unusual subset of the demand characteristics conceptual space. They also have long served as a tool for studying demand characteristics. * Contrary to concerns that EDC’s are ineffective are producing acquiescence (Orne, 1976), results suggest that this is actually the most commonly observed outcome. Similarly, contrary to concerns that EDC’s create “paradoxical” counter-acquiescence (Orne, 2009), results suggest that such outcomes (if real) are quite rare. * Yet, the simple question of whether ECD’s act in a qualitatively different manner than other demand cues is an open question.   Regarding the vignette study: we note in our response to Editor Comment 3 that we will remove the study from the manuscript, but briefly add that the unpublished investigation is described in the *Supplementary Materials*. |

## R1 Comment 3e

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.

|  |
| --- |
| We appreciate you pushing us towards improving the conceptual clarity of our writing on this topic.  In the updated *Introduction*, we now explicitly introduce the following distinctions:   * Demand characteristics: *any* cue that may impact participants’ beliefs about the purpose of the study, including instructions, rumors, and experimenter behavior. * Explicit demand characteristics (EDCs): cues that directly communicate an experimental hypothesis to a participant.   This phrasing also appears in the updated title:  *A meta-analysis of the impact and heterogeneity of explicit demand characteristics.* |

## R1 Comment 3f

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.

|  |
| --- |
| Thank you for elaborating. We believe that our responses highlight that this has indeed helped improve the clarity of the updated manuscript. |

## R1 Comment 4

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

|  |
| --- |
| We note in our response to Editor Comment 2 that we have removed this phrase from the updated manuscript. |

## R1 Comment 5

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?

|  |
| --- |
| We have removed this sentence from the updated manuscript. |

## R1 Comment 6

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.

|  |
| --- |
| Corneille, thank you for pushing the ideas, acknowledging the effort, advocating for the importance of the work, and engaging in a long and productive dialogue about this phenomenon. |

# Reviewer 2 feedback

## R2 Comment 1

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.

|  |
| --- |
| Thank you for engaging with the work. Reviewer 1 shared this concern, and we updated the title to:  *A meta-analysis of the impact and heterogeneity of explicit demand characteristics.* |

## R2 Comment 2

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).

|  |
| --- |
| In the resubmission, we have updated the reference style for multi-author papers. We also reviewed and updated all cases where we used the word “significant”. |

## R2 Comment 3

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

|  |
| --- |
| We removed the reference entirely, as it was not essential for the claim we sought to make. |

## R2 Comment 4

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

|  |
| --- |
| Fixed! |

## R2 Comment 5

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

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

## R2 Comment 6

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

|  |
| --- |
| This is now added. |

## R2 Comment 7

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 engaging with our argument. We think you will also be pleased to see that we dedicate more attention to this issue in updated manuscript. (We summarize these changes when responding to Reviewer 1, Comment 3.) |

## R2 Comment 8

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.

|  |
| --- |
| Thank you for pointing out this important historical context. We briefly revisit this context in the *Limitations* section of the manuscript. |

## R2 Comment 9

Page 9. 1840?

|  |
| --- |
| We updated this to be less confusing.  We were attempting to mention the time frame that was searched, in accordance with APA Meta‐Analysis Reporting Standards. We used “1840” to refer to the date of some of the oldest records in the database we searched. However, in the updated manuscript, we say:  “Our search did not have language or date restrictions”. |

## R2 Comment 10

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)?

|  |
| --- |
| Your understanding of Figure 1 is correct – although we hope to challenge your hunch about what participants might reasonably infer.  We agree that there are many scenarios where participants could reasonably infer a researcher’s expectations when they say that an independent variable will “impact mood”. However, we created this exclusion criteria because there were scenarios where we were not confident in this assumption.  For example, consider a study that has people force smiles and tells them it will “impact mood”. Some participants may infer that this means mood is expected to increase (e.g., if they suspect a ‘fake it to you make it’ mechanism) but other participants may infer that mood is expected to decrease (e.g., if they suspect a ‘emotion labor’ mechanism). |

## R2 Comment 11

So 38 studies and 52 effect sizes, correct?

|  |
| --- |
| We are not sure what part of the manuscript you are specifically referring to. To clarify, though, *s* refers to studies and *k* refers to effect sizes. |

## R2 Comment 12

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

|  |
| --- |
| We’re not quite sure about what statement you’re referring to. However, we did ensure that we proof-read and edited this page carefully. |

## R2 Comment 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?

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

## R2 Comment 14

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?

|  |
| --- |
| You raise two important points here.  We believe your first point is about the operationalization of “demand characteristics”. In the Introduction and Discussion section, we spend considerable more time discussing this issue – highlighting the broadness of the construct and defining the particular operationalization examined in the present work. More information about these changes is in our responses to Reviewer 1 Comment 3.  The idea that participants know we often lie to them adds another layer to this puzzle. Put simply, though, we suggested that can be considered an issue that connects to Rosnow and Rosenthal’s receptivity moderator:  “…Even if the infant possessed the astonishing ability to read, it’s possible they would misunderstand the cues (Corneille & Lush, 2023) – which could be considered another form of non-receptivity.” |

## R2 Comment 15

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

|  |
| --- |
| Correct. We offer further clarification in our response to your next comment. |

## R2 Comment 16

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?

|  |
| --- |
| Your understanding of the group comparison moderator is correct.  This moderator analysis does connect to our effect size coding scheme. More specifically, the analysis strategy for this moderator assumes that effect sizes are coded so that positive values always represent changes that are *consistent* with the communicated hypothesis.  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 always acquiesce – but that they are more likely to exaggerate than attenuate a true effect when exposed to explicit 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 *amount of hypothesis-consistent* 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  To summarize, this moderator analysis is connected to the effect size coding scheme – only in the sense, though, that the effect size coding scheme helped set up the analyses. |

## R2 Comment 17

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 explicitly 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 a “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”). |

## R2 Comment 18

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.

|  |
| --- |
| As detailed in our response to Editor Comment 3, we have removed this study from the main manuscript (and moved it to the *Supplementary Materials*). |

## R2 Comment 19

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. |

## R2 Comment 20

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

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

## R2 Comment 21

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

|  |
| --- |
| Done! |

## R2 Comment 22

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

|  |
| --- |
| Correct! |

## R2 Comment 23

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 two vs. three levels are 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 now also decompose variability at multiple levels (see response to Editor Comment 8). |

## R2 Comment 23

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”.

|  |
| --- |
| This typo is now fixed. |

## R2 Comment 24

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.

|  |
| --- |
| Interesting question.  We don’t have the data to speak to this, of course. But we would be reluctant to say that this would be the case in \*every meta-analysis\*. E.g., a meta-analysis of extremely large studies that all examine the same large non-heterogeneous effect would (in theory) produce a distribution that is non-overlapping with zero. |

## R2 Comment 25

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 immense amount of [mostly unexplained] heterogeneity we’ve observed. We point this out this issue in the *Limitations* section. |

## R2 Comment 26

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. |

## R2 Comment 27

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 some analyses (e.g., weight-function modeling with aggregated dependencies) but not others (e.g., PET) |

## R2 Comment 28

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.

|  |
| --- |
| The term “unreliable” is now removed from our title. |

## R2 Comment 29

Page 38. No page number for the quote here.

|  |
| --- |
| Added |

## R2 Comment 30

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. However, we removed it to avoid unnecessary confusion. |

## R2 Comment 31

Glad the authors do not outright reject broader definitions of demand characteristics given my definition.

|  |
| --- |
| Glad to hear it. And we provide additional context in the updated manuscript. |

## R2 Comment 32

Page 40. Footnote here is in an odd place, perhaps should come in the method section.

|  |
| --- |
| Agreed. We’ve now moved it to the methods section. |

## R2 Comment 33

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.

|  |
| --- |
| We removed all instances of the term “unreliable” in the updated manuscript. |