# Editor feedback

1. Reviewers 2 and 3 noted an important scope limitation of all three studies: that the evidence reviewed is really about the effects of *explicit cues to experimental hypotheses rather than about effects of demand characteristics in general*. It would seem prudent to revise the title, abstract, and general discussion to more accurately characterize this scope.

|  |
| --- |
| Updated title to: What have we learned about demand characteristics? A meta-analytic review of experimental manipulation of explicit hypothesis cues  Will update language throughout. |

2. Reviewers 2 and 3 raised concerns about the "hybrid" conceptual framework on which all three studies rely. I agree with Reviewer 2 that it is essential to clarify if and how the proposed framework differs from existing accounts in the literature, and I would ask that you respond to all reviewers' points regarding characterization of past literature on theoretical mechanisms.

3. Reviewer 1 offered extensive feedback on systematic review methodology, with pointers to many very recent developments. Broadly, I don't think it's reasonable to ask for authors to be deploying every latest advance in AI or statistical methodology for meta-analysis, but please take these suggestions under advisement. I do think providing a PRISMA flowchart is essential. I also shared the reviewer's concerns about the narrow set of search terms used to identify candidate studies. It may be useful to examine frequently terms used in the identified studies to identify synonyms or other useful keywords for augmenting the search.

4. The reviewers were somewhat skeptical of the value of collecting motivation and expectancy ratings based on vignettes (although they also appreciated the creativity of this approach), and I expect that a degree of skepticism would be shared by many readers. To strengthen the case for use of such measures, it would useful and appropriate to provide more information about the psychometrics of the measurements, such as estimates of intra-class correlations and/or other statistics regarding the aggregated scores. At minimum, some understanding of the reliability of the measurements seems required in order to judge the evidentiary value of the finding that motivation ratings are not associated with effect size (considering that randomly generated scores would not be associated either).

5. Based on the examples in Figure 6, I also noted that the vignettes seem to incorporate information about some of the same study characteristics examined in Study 1, which led me to wonder how much of the variance in the ratings might be associated with these characteristics (e.g., student population, monetary compensation or not, in-person or online). As validation evidence of the ratings, it may be useful to provide some information on the extent to which they can be predicted by the study characteristics. As a sharper test of the measures, it seems like it would also make sense to test the ratings in a model that also controls for the specific study characteristics described in the vignettes.

6. Reviewer 1 suggests coding studies for risk of bias. It is journal policy to capture the methodological quality of the included studies as part of your method. Please clarify and if necessary, add this component to your manuscript and/or supplemental files. Consult the most recent [*Bulletin* editorial](https://psycnet.apa.org/fulltext/2021-07795-001.pdf) (page 9) for more information (also Table 2).

7. It is journal policy that meta-analytic reviews published in *Psychological Bulletin* present a description of the literature summarized at the outset of the Results, before the quantitative material you present. This material often sets the stage for all of the analyses by telling readers how your literature looks: Your team did the hard work of coding studies for descriptive and moderator features, yet there is no description of the central tendencies (or extremes) of these variables. How do the studies look? When were they conducted and where? With what populations? How do they look in terms of descriptive and moderator variables? You should not expect that readers will form the same views from a perusal of your list of studies and coded features. Instead, compose a new table that meaningfully summarizes these items, showing central tendencies (e.g., means, medians, and modes; ranges). At times, this information can point up serious issues, such as confounds of important moderators with each other or with descriptive variables, although at present that is difficult to ascertain. Constructing a summary might mean that you ought to do additional moderator testing (e.g., if you see that there is a confound you did not realize exists); of course, in such instances label it as post hoc, found after analyses commenced. It is possible that this information could change your conclusions and it is possible that aspects of this portion of your results merits mention in your Discussion section. You might consult Table 2, Step 5, “Were descriptive statistics presented?” in the most recent [Bulletin editorial](https://psycnet.apa.org/fulltext/2021-07795-001.pdf) as a guide.

Finally, as a general point of guidance, Psychological Bulletin rarely publishes multi-study articles (a format that is more common for journals like Psychological Science), and I would encourage you to give some consideration to how this manuscript might be restructured to better emphasize the systematic review and synthesis that are the central contribution. For instance, Study 2 is essentially an examination of some specific moderators in the database identified in Study 1. It therefore seems like little might be lost by restructuring them as a single study. Study 3 involved new primary data collection and entirely distinct analytic methods, and so it makes sense to present it as its own thing. Still, if I understand correctly, the data from Study 3 is nonetheless relevant and would meet inclusion criteria of the systematic review in Study 1. Should the summary findings from Study 3 therefore be included in the synthesis results too?

# Reviewer 1 Study 1

First of all, I find this study to be very impressive in terms of how much work the authors have put into it, and I appreciate the way the authors stick to the principles of open science and pre-registration. However, I have certain major concerns about the article.

|  |
| --- |
|  |

I think the authors must conduct a more comprehensive search for the review to be relevant for Psychological Bulletin…I would suggest that the authors expand their search to be more comprehensive but then draw upon various automated screening tools…

|  |
| --- |
| Will definitely need to expand the search. Hopefully reviewers will give good ideas about how to do so:  “participant role” (248)  “demand effects” (45)  “good subject effect” (5)  “expectancy effect” (186)  “evaluative apprehension” (6)  "self-fulfilling prophecy" AND "experiment" (32)  *BAD (too many hits)*  Self-fulfilling prophecy (551)  It might be fun to try out an AI tool and report a little bit of validation work.  Advice from editor: It may be useful to examine frequently terms used in the identified studies to identify synonyms or other useful keywords for augmenting the search. |

I find the search to be outdated…I would therefore suggest that the search should be updated…

|  |
| --- |
| That’s fine. Easy enough to do. |

The authors opted to combine effect size estimates from both repeated measures and independent group designs. In my opinion, these designs should be separated since the former most often overestimates the true effect. If not separated the authors must at least follow the guidelines provided by Morris & DeShon (2002) and somehow justify why the two designs can be said to estimate the same treatment effect. Yet I don’t recommend this approach.

|  |
| --- |
| Interesting.  We do mostly follow the Morris and DeShon (2002).  Effect sizes were transformed into a common metric, using drm and design-specific estimates of sampling variance. Sensitivity analyses on the assumed correlation produced virtually no change in our overall effect size estimate, suggesting that we could have essentially ignored the fact that this was a repeated measure design. However, I guess it’s possible that it’s inflating estimates of heterogeneity, which could also be assessed in a sensitivity analysis.  Morris and DeShon’s paper is a slog. (I’ve read it before and am trying to convince myself to read it carefully again.) One thing I noticed while re-skimming is that they argue that people should better aim for commensurability by multiplying drm by the square root of 2(1-p). I’ve never seen that done, but we could try it.  Morris and DeShon (2022) argue (p. 123) that the effects can be combined if no significant difference effects are observed across the two designs. That was the case, although we might not have standardized things the way they would have recommended. (We would have to re-read carefully to know.)  That being said, Morris and DeShon point out that this could be decided on a conceptual level. If you think the designs are estimating fundamentally different effects, you shouldn’t combine. That’s a tough call here. I could certainly make an argument for why within-subject designs are different: they give participants more opportunity to adjust responses to fit the hypothesis.  If we split analyses by design (deviating from pre-reg), the moderator analyses will probably be dramatically underpowered. The most likely conclusion would be that we can’t conclude anything from the meta-analysis, which the reviewers might not actually hate. |

I think it is pivotal that the authors conduct some kind of risk of bias assessment of the studies included in the meta-analysis. I would recommend using the ROB2 tools, see <https://sites.google.com/site/riskofbiastool/welcome>.

|  |
| --- |
| I’m not really a fan of this ROB2 tool for this. It’s too specific to clinical trials. E.g., it talks about random assignment procedures, baseline differences, deviations from protocol, etc. – all which are not usually assessed and reported in psych.  The downs and black method was described I the editorial that the editor linked to. That one seems somewhat tricky but more feasible: <https://cdn-links.lww.com/permalink/jps/a/jps_2020_03_05_heip_15-1405_sdc3.pdf> |

The authors must add a PRISMA (2020) flow chart (Page et al., 2021) to easily depict the search, screening, and exclusion processes.

|  |
| --- |
| We can do this. |

I generally miss certain paragraphs in this study, that is 1) a paragraph that elaborates on previous reviews, 2) a paragraph regarding the specific contribution of the current review and why it is needed, and finally 3) a paragraph concerning deviations from the pre-registered protocol. To develop the latter, see Lakens (2023) for inspiration. This point tends to be a major issue as well, especially if the contribution is minor relative to previous reviews. I don’t know the area well enough to comment further on this.

|  |
| --- |
| We can add this.  There is no meta-analysis; all previous reviews were narrative.  We documented deviations as we went, but we will have a lot more if we respond to these reviwer comments. |

Since many of the calculated effect sizes are based on small sample sizes, I would suggest that the authors calculate Hedges *g* (Hedges, 1981) instead of Cohen *d*.

|  |
| --- |
| That’s fine. |

I appreciate that the authors have made a protocol. Unfortunately, the pre-registration has not been frozen at a specific time point (as far as I could see – if I am wrong please add a link to the frozen version), thus I cannot see when the changes to the protocol were made. The document was last updated 2024-01-23 which is after the study was conducted. Therefore, (I don’t believe this to be the case) HARking cannot be entirely excluded.

|  |
| --- |
| The original pre-registration is time stamped (<https://osf.io/kwpys?mode=&revisionId=&view_only=>), but because we did not use an OSF template (none existed at the time), we were not able to find a way to update it as deviations occurred.  You’ll see that the deviations are correctly described: we switched to multi-level modeling and added plans for a Study 2.  That being said, the revisions we made to this meta-analysis require substantial deviations from the pre-registration analysis plan, which we will have to document in the updated manuscript. |

Due to the complex data structure, including both correlated and hierarchical dependency structures, I would highly recommend using the CHE-RVE models (including the SCE model for subgroup analyses) developed by Pustejovsky and Tipton (2021). This can easily be implemented via the robust function from the metafor package.

|  |
| --- |
| I’m fine with that. |

I will recommend using the CHE-RVE models for publication bias testing as well.

|  |
| --- |
| We could consider this, but I’m not sure how it would work for things like weight function modeling. Maybe we just add it as yet another approach. |

On a similar matter, to ensure valid inferences when testing for relative effects across moderator/subgroup factors, I think the authors should apply at least one of the robust tests developed and described by Joshi et al. (2022) and Tipton and Pustejovsky (2015), respectively.

|  |
| --- |
| I think this is fine. I suspect using robust methods is going to lead us to not be able to conclude much. |

Hereto, I highly recommend that the authors share the script in which they conduct the final analyses. From the shared files, I could only find the effect size calculation script (which I enjoy inspecting).

|  |
| --- |
| We can point the reviewer to all the scripts. They missed the analysis script because it’s also the manuscript script. |

The forest plot must be rescaled to ensure a better resolution. The authors can also consider using forest plots as presented by Fernández-Castilla et al. (2020) or Winter et al. (2022, p. 45, see Figure 2).

|  |
| --- |
| I’m not sure if we can address this. There’s too many studies! Perhaps we could panel it out? |

When looking at the forest plot, I find it quite odd that many within-study effects replicate each other both when it comes to the main effect size estimates and the corresponding confidence intervals. For example, when looking at the data, all 15 effect size estimates for Kenealy (1988) are exactly the same (i.e, . To me, it seems so extreme that I can hardly believe it, and this is not the only case. Is this a mistake? I suspect that something has gone wrong with the effect size calculation.

|  |
| --- |
| Not a mistake.  In the manuscript, we mention that we didn’t throw away an observation if they described significance but didn’t provide any additional information needed to compute an effects size. Instead, we assumed a specific p-value and then tried to derive and effect size from that p-value.  In that case, the entire paper only described significance. We can add a footnote to point this out to the reader, but it will also come into play if we code for quality. |

I think the authors should conduct a moderator analysis investigating differential effects across the type of used control, that is “where either no hypothesis or a different hypothesis was communicated” to the control group.

|  |
| --- |
| We actually already have this analysis and discussed it quite a bit. The reviewer (understandably) missed it. |

In the protocol, the authors write that they would use the publication bias test developed by Mathur and VanderWeele, but I cannot find this sensitivity analysis.

|  |
| --- |
| We could do this or add it as a deviation.  The reason I dropped it is because the method wasn’t developed to handle nested structures. |

“

When estimating publication bias/small study effects, the authors opted to use regression tests with aggregate effect sizes. To my understanding, this approach is not recommended by Rodgers and Pustejovsky (2021). If I am correct (the editor knows better than me), I think these tests could be removed from the study.

|  |
| --- |
| I don’t agree with that interpretation. R and P’s discussion is nuanced, and we follow it pretty closely.  “Alternatively, the 3PSM likelihood ratio test addresses limitations of these modified regression-based tests by more directly testing for selection and by providing increased power…When testing for selective reporting, prac- titioners should consider taking an exploratory approach by report- ing multiple tests, such as the Egger MLMA and Egger sandwich tests, along with exploratory 3PSM likelihood ratio tests after aggregating or sampling effect sizes.” |

In connection with the above comment, it is unclear to me if the authors removed the artificial correlation between the effect size estimates (i.e., standardized mean differences) and their related variance as suggested by [Pustejovsky and Rodgers (2019).](https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1332) If not done, I suggest that the authors do so.

|  |
| --- |
| We can try to do this. I’m just not sure if there is an R package for it |

Just a minor note: Next time the authors want to blind their material, they must make sure that the material is truly blinded. It is not enough just to change the color of the text. It was very easy for me to disclose the identity of the authors. This was also the case because the authors did not blind the document properties of the Word file. Consequently, I could not save the document without disclosing the identity of the main/first author.

|  |
| --- |
| We should try to address this, although blinding is pretty impossible these days. Just remove the embedded data from OSF and you’ll know who is who |

# Reviewer 1 Study 2

To be frank, I am very unsure if I understand the importance of this study. For me, this study represents too indirect evidence to be relevant. I admire the creativity but I hardly see the contribution. I might have misunderstood something here, but I think the authors must make it much clearer why this study is important.

|  |
| --- |
| We might have to drop the study. But I think the justification is that none of the studies we found are trying to explain how this works. |

I am quite skeptical about the choice of a student sample since I find it too convenient, which seriously restricts the generalizability of the study.

|  |
| --- |
| It’s a good point. But I think this is actually a case where a student sample is well justified: we want them to evaluate what a primarily student population of participants would do. |

Since I am skeptical about whether the authors have detected a representative sample of studies, I consider it to be a “risk business” to hinge Studies 2 and 3 so much on Study 1. If the meta-analysis in Study 1 is adequately updated, I am not sure if Studies 2 and 3 are needed for the article to make a profound contribution to the field.

|  |
| --- |
| I’m definitely open to dropping these studies. |

All estimations must be aligned with the suggested model changes mentioned in Study 1.

|  |
| --- |
| I’m definitely open to dropping these studies. |

I was not sure how to interpret Figure 7. A more thorough description is needed.

|  |
| --- |
| Really tough figure, but a necessary one. I hate it. |

# Reviewer 1 Study 3

Generally, I lack a better explanation of why this study is needed and what the contribution is. Since I was able to disclose the identity of the authors, I can see that it is a replication of one of the main author’s previous studies. As with the choice of student sample, this seems to be too convenient of a choice to me, and I think the authors should be honest about the overlap in authorships between Study 3 and the Coles et al. study. Furthermore, the authors must come up with a very good reason why a replication of this study is important relative to other studies from the meta-analysis. Otherwise, I just see this as cheery picking.

|  |
| --- |
| That’s a good point. There definitely is a convenience factor, but this is one of the only pre-registered paradigms and it reliably produced strong demand effects. Because I ran the studies, I trusted the findings enough to try to build off them. |

To be honest, I find the interpretation of the results rather odd. On page 43 the authors interpret an effect of 0.04 to “slightly larger”, an effect of 0.03 to be represent a null effect, and an effect of 0.05 to be a “larger” effect. Unless I misunderstand the used scale (see point 26), this graduation of effects seems strange to me since I would consider such small differences to be negligible.

|  |
| --- |
| Great catch. I agree that this is problematic. |

I am not sure I understand why the authors only want to compare different types of demand characteristics and not to a control where no hypothesis is communicated. I think this might be fleshed out in more detail.

|  |
| --- |
| I’m not really sure what was going through my head at that time. My bet, though, is that I wanted to isolate a stronger demand effect so that we could better study moderators of the effect. Comparing positive demand to nil demand (vs. control) should produce an additive (i.e., larger) demand effect.  But I don’t think this reasoning is flawless. |

I generally found it hard to follow the estimation and effect calculations in this study. I am not sure either that I understand how the comparison group was constructed. The authors could elaborate on this.

|  |
| --- |
| Not sure if I understand where they got confused here, but can keep looking |

Also, see points 19 and 20 above for further concerns about this study.

|  |
| --- |
| I think those concerns are more justified in this study. We could discuss those limitations futher |

# Reviewer 2

I see much value in this manuscript. First of all, this research addressed an important question that is plagued with conceptual and methodological difficulties. This is a courageous endeavor, which I believe should be generally supported. Second, the authors went beyond meta-analytic results and reported one additional vignette study that they related to their meta-analysis (Study 2), as well as an exploratory experimental study (Study 3). Third, the manuscript is generally clear and balanced, and the approach is generally cautious and transparent.

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

1. Demand characteristics versus demand effects: I think that these should not be confused. Demand characteristics (i.e., cues in the situation) elicit expectations about experimental hypotheses, which in turn produce demand effects. Effects of demand characteristics can be quantified. However, I don't think that demand characteristics (i.e., cues) can be quantified. Likewise, I don't think that demand characteristics (i.e., cues) have "underlying mechanisms".

|  |
| --- |
| I agree. We can update. |

Subtle vs Explicit demand characteristics: the present research was concerned with the effects of communicating experimental hypotheses very explicitly to the participants. This is in contrast to demand effects arising in the large majority of experiments where participants draw (correct or incorrect) inferences about experimental hypotheses that remain untold. I believe it is important to systematically stress that the current studies cannot quantify "demand effects in general". Rather, they inform us about how Ps react to experimental hypotheses that are explicitly stated by the experimenter, which rarely applies. In my view, this critical distinction should be highlighted in the manuscript, implying a more cautious and representative title/abstract, etc.

|  |
| --- |
| I agree. We can update.  We could also stress the conceptual problems with studying ‘subtle’ demand characteristics. |

3. The vignette procedure used in Study 2 is a form of instruction-based procedure. These procedures, and their relevance to demand effects have been recently discussed here: <https://doi.org/10.1525/collabra.82234>

|  |
| --- |
| That’s true. I actually considered citing this paper, but decided against it because our method isn’t quite the same as what they’re describing.  It’s similar to an IBR in the sense that participants are provided verbal descriptions of the study. It’s different, though, in the sense that we are not seeing they predict/simulate the same responses as the original study. In this sense, there is no original study. Nonetheless, many of the same limitations apply, including concerns that what is measured in the IBR is somehow different than what happened in the original study. In part, this is why we ran Study 3. |

The authors discuss only one hybrid model, by Coles et al. (2022). However, the model by Corneille & lush (2022; quickly alluded to in Footnote 1) is also a hybrid model. The latter model states that participants form expectations about experimental hypotheses. Then, they form a motivation to comply or not with this perceived hypothesis. Finally, this motivation elicits demand effects through faking, imagination, or phenomenological control. Whereas faking produces fake responses, imagination and phenomenological control produce genuine experiences and responses (in the case of phenomenological control, without awareness of producing these experiences intentionally). As it appears, the model integrates expectations and motivation, and goes beyond by specifying processes through which demand effects produce, intentionally or not, faked and genuine experiences. Of importance too, the mismatching case discussed by the authors when presenting their model (and addressed in Study 2 and 3) corresponds to the two-experiments problem in Orne, 1973 (see also Sharpe & Welton, 2016) \*\*and is covered by the hypothesis-mistaken case in Corneille & Lush (2022)\*\*.  
  
Again, I don't want to impose my model over the one proposed by the authors. However, I think it is fair to ask the authors (1) to provide an accurate description of Corneille & Lush (2022)'s model (see also point 4b below), (2) to avoid discussing uniquely Coles et al. (2022)'s model in the hybrid model case, and (3) to spell out why they think Coles et al. (2022)'s model is competitive (e.g., more comprehensive, more accurate, or more parsimonious).

|  |
| --- |
| I think we should discuss Corneille & Lush (2022) in the intro.  One reason why it was sidestepped initially is that we very intentionally pre-registered this as a test of the motivation-opportunity model by Rosnow vs. a motivation-opportunity-and-belief model discussed by Coles et al. But Coles et al. could have very well been Corneille and Lush, because they both discuss the importance of considering belief.  I think the only difference between Coles et al. and Corneille and Lush is that Corneille and Lush don’t discuss expectancy effects for non-motivated participants. Phenomenological control occurs voluntarily (at least at first), whereas placebo effects can occur involuntarily.  If my understanding is correct, the potential limitations of Study 2 and 3 make it difficult to convincingly arbitrate between the theories. To do so, you would need an experiment where participants believe an effect but aren’t motivated to adjust their responses. |

Coles et al. (2022)'s model relies on a classic motivation + opportunity framework. However, the "opportunity" question is a thorny one, whose complexity I believe was addressed in Corneille & Lush (2022). I find it useful to raise two points here:  
  
5a. Controllability\_a: Uncontrolled effects may arise at the time Ps complete the dependent variables, but these uncontrolled responses may reflect a largely controlled process arising at an earlier stage in the experiment. For instance, participants taking part in an evaluative conditioning study where CSs are paired with USs may intentionality produce (e.g., through imagination) a controlled impression of the CSs that is consistent with their paired US; once this controlled impression has been formed, it may produce uncontrolled responses on the DV at a later stage of the experiment.

|  |
| --- |
| Yeah, that’s super tricky. |

By the way, I would like to invite the authors to avoid relaying widespread uncontrollability misconceptions about the IAT. There is plenty of evidence that the IAT is sensitive to both social desirability and experimental demands (e.g. <https://doi.org/10.1037/0022-3514.81.5.842>), and that it can be faked and controlled (e.g. <https://econtent.hogrefe.com/doi/10.1027/1618-3169.51.3.165>; <http://link.springer.com/10.3758/s13428-015-0568-1>)

|  |
| --- |
| We can remove this just to avoid the hate. But the IAT is clearly *less* controllable than a self-report. |

5b. Controllability\_b: In Footnote 1, the authors suggest that Corneille & Lush (2022) posit that "motivation-based mechanisms can produce demand effects even when participants do not have the opportunity to adjust their responses". Instead, the model proposes that demand effects can produce genuine experiences (in the case of imagination and phenomenological control, as opposed to faking), and that this can be the case even when participants are not aware of their intention of producing it (in the case of phenomenological control).

|  |
| --- |
| Yes, but it seems that a key part of phenomenological control is that participants are at first motivated but then somehow convince themselves that it was all involuntary? Or is the claim that phenomenological control. |

Note that the phenomenological control case (i.e., unconscious intention) severely limits the utility of motivational self-reports, as used in the vignette study (i.e., Study 2), for estimating the contribution of motivational processes to demand effects. As to whether the model posits demand effects when there is a lack of opportunity to control responses, see previous point.

|  |
| --- |
| But aren’t they motivated at first? |

6. Figure 6 reports illustrative vignettes. However, the content of these vignettes does not suggest that the experimenter is expecting a particular directional effect to emerge. For instance, "The researcher informs you that \*they are interested in\* the beneficial effects of listening to Mozart on test-taking capabilities" does not imply that the experimenter is expecting a positive effect - it just informs Ps of which effects is being \*examined/tested\* by the experimenter. I sampled a few vignettes on the OSF page, and it seems that the issue does not apply to those I've checked, though. Hence, I don't know how problematic the issue is. It is possible that only a few vignettes were problematic, or that none of them was but that Figure 6 misreported the vignette content. Or perhaps I got confused in the OSF file. In any case, I believe this point calls for a clarification. If it turns out that the meta-analysis involved studies where the experimenter's expectation was clearly stated, whereas naïve observers in Study 2 had to provide judgments based on vignette studies where no clear experimental expectation was communicated, then I believe this would pose a major threat on the validity of Study 2.

|  |
| --- |
| This is a good point about the limitations of Study 2. We can’t describe the original studies precisely for two reasons (1) they’re not precisely described in the papers we meta-analyzed, and (2) we have to keep the description short.  One interesting point that they’re bringing up, though, is that we actually focused on “relatively explicit” manipulations of demand characteristics. It didn’t have to be a reasearchers saying “I hypothesize X”; explicitly saying “I’m studying the benefits of X” was considered explicit enough. |

7. Demand effects arise in specific individuals forming a specific understanding of a study setting in light of their prior beliefs. This understanding can converge or diverge across participants. Also, if sharing a same interpretation, Ps may decide to comply or react to this understanding. It would be worth discussing the limits of a meta-analytical approach in this context - in particular, to what extent does it make sense to compare "average" effects across study procedures instead of looking at the distribution of effects in individual studies?

|  |
| --- |
| I think we could discuss that.  In the ideal world, we would look at distribution of effects in individual studies – but this would require a level of open data sharing that just doesn’t exist. The within-study variation tells us more about individual differences, whereas the between-study variation tells us more about situational differences. They’re both interesting. |

Minor  
  
8 p. 9: "We then examined the extent to which the effect was moderated by motivation to adjust responses, opportunity to adjust responses, and expectations about the hypothesized effect » -- please be more precise here and refer to the estimation of the relation between meta-analytic effects and self-reported forecasts collected on a separate group of participants based on short vignettes. I also recommend discussing the limitation of self-reported forecasts for making inferences about mechanisms involved in experienced procedures.

|  |
| --- |
| Yeah we could do that if we keep it. |

9. p. 21: I was not sure to understand why/what it means that "participants' responses most strongly shift when researchers communicate that they hypothesize no change in response". How should this shift be interpreted this shift, if compared to a control condition?

|  |
| --- |
| Hah! Yeah, very hard to parse this.  Compared to a condition where no hypothesis is communicated (what I think you mean by control here), this means that the effect being studied becomes attenuated.  So, imagine a scenario where participants receive nil demand, no demand, or positive demand. This is saying that (a) the difference between the nil and no demand condition is bigger than (b) the difference between the positive and no demand condition. |

# Reviewer 3

1) Throughout, key conclusions rest on invalid statistical inferences. For example, the reported analyses do not support the central claim that the results are "contrary to conventional motivation accounts". There is no statistical evidence for a difference between slopes and claims of no effect are based on invalid interpretation of non-significant p values. It is clear from the 95% CIs that tests would be unlikely to support the authors' conclusions. If there is no evidence for a difference between slopes of expectancy and motivation, it cannot be argued that an expectancy account is supported and a motivation account is contradicted (e.g., see Gelman & Stern, 2006, which I was puzzled to see was cited in the manuscript but not applied to interpretation of the results). If there is no statistical evidence for the null (e.g., from equivalence tests or Bayes factors; <https://doi.org/10.1525/collabra.28202>) it cannot be concluded that there is no effect (a non-significant p value alone does not support claims that "motivation accounts" are contradicted). As another example, Study 2 is motivated by the claim that students show different results to workers, but the claimed is not backed by a test. These are not the only examples of invalid statistical inference. For instance, non-significant p values are incorrectly interpreted as evidence of an effect. E.g., on p.43, p = .063 is interpreted as evidence of a "slightly larger" effect and a p value of .117 as an effect which was "less robust".

|  |
| --- |
| I think there is a lot of non-charitable interpretation of the inferences being made here.  We don’t ever argue for the null; we argue that we fail patterns predicted by theories.  We don’t argue that one effect is stronger than the other (i.e., that there is an interaction). We argue we found one and failed to find the other. We could highlight that their patterns don’t seem to substantially differ, but this does not change the fact that if you were specifically looking for one of those effects, you would not find it. |

2) The cited literature does not support central claims. A central theme in the manuscript is a distinction between what the authors refer to as "motivation accounts" and "expectancy accounts" of demand characteristics. The "motivation account" is attributed to researchers who have worked on demand characteristics. The "expectancy account" is attributed primarily to the authors of a review on placebo effects (who do not discuss demand characteristics). The implied claim seems to be that researchers who are cited as favouring "motivation accounts" have historically not considered expectancy effects (or placebo) to be involved in demand characteristics. The authors then claim to have created a "hybrid account" by proposing that both expectancy and motivation may be involved in the effects of demand characteristics. However, response to demand characteristics has been considered to involve both expectation and motivation since the term was coined by Orne in the mid-20th century. In the manuscript, Orne is cited as favouring a "motivation account", but his classic 1962 paper which the authors cite in support of this claim discusses expectation effects arising from demand characteristics in a sham sensory deprivation experiment. Orne (1969) contains an extended discussion of "beliefs and expectations" in demand characteristic effects and describes placebo effects as "analogous to the demand characteristic components in psychological studies". Orne clearly held to what the authors describe here as their "hybrid account" over 60 years before they claim to have introduced it. A single paper (Corneille & Lush, 2022) is cited in support of the claim that there are long-standing disagreements about how participants alter participant responses, but the paper does not describe such disagreements and does not support the claimed distinction between "motivation accounts" and expectancy accounts at all. Other directly relevant work has been overlooked entirely. For example, Kantar et al's (2002) treatment of "therapeutic demand characteristics", which directly links demand characteristics and placebo effects. Even Rosnow & Rosenthal (1997), who focused on motivation as a research target, describe effects of beliefs about the experimental situation as a "kind of placebo effect in psychological research".

|  |
| --- |
| We were not trying to imply this claim, and it is easy to correct.  We never argued that Rosnow and Rosenthal ignored placebo altogether. Just that, as you mentioned, they focused on “motivation as a research target”. The same goes for placebo researchers. These are groups of researchers who are studying a similar thing but typically focusing more of their attention on one mechanism over the other.  We don’t claim that Coles et al. introduced hybrid accounts. Just that they discuss it.  I think the most productive thing we could do to address this would be to mention that, of course, people have considered these mechanisms together. We’re pointing to illustrative examples so that the reader can understand some of the potential mechanisms. |

Problems with both theoretical motivation and statistical inference aside, the claim that "motivation accounts" are contradicted seems to imply the claim that motivation does not play an important role in demand characteristics. If this is indeed the authors' claim, it seems to me implausible, and I do not think the procedure employed could even in principle support it. I am not convinced that extending Orne's quasi-experimental "non-experiment" design (or 'instruction-based replications'; Corneille & Bena, 2023) beyond their original purpose of assessing whether participants can work out an experimental hypothesis is likely to provide meaningful results. If there were evidence for no effect of motivation here, I would want to rule out the possibility that participants provided with short descriptions of experiments fail to accurately predict their motivation to respond to demand characteristics before concluding that motivation is not important in demand characteristic effects.

|  |
| --- |
| That’s a good point. Maybe the participants do indeed fail to predict their motivation. |

I paid little attention to the meta-analysis. In theory, this could stand apart from Study 1 and Study 2 . However, Figure 2 does not seem to support the authors claims. For example, it is claimed that 63% of manipulations produce hypothesis-consistent shifts in response. It appears from the plot that the 95% CIs cross zero for many of the studies. Is this interpretation based on just the point estimates?

|  |
| --- |
| It's based on the estimated – not observed – distribution of effects. |

Also, I may have missed an explanation of this, but why are identical CIs repeated in Figure 2? E.g., two CIs from Kenealy (1988) are repeated 8 times each.

|  |
| --- |
| Addressed in feedback to Reviewer 1. |