# 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.

|  |
| --- |
| We have developed an updated title that we hope you and reviewers agree (a) nicely illustrates our main thesis, (b) accurately reflects the scope of our review, and (c) is not too verbose: “A demanding problem: Meta-analysis suggests that demand characteristics exert effects that can be powerful, unreliable, and difficult to explain”.  We also made the scope of our review – and the limitations of that scope – more explicit in the abstract and throughout the general discussion. |

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.

|  |
| --- |
| We overhauled the manuscript to address this concern.  In the updated manuscript, we now primarily focus on a single influential framework by Rosnow and colleagues. For the sake of comprehensiveness, we initially attempted to dedicate substantial attention to ideas very recently proposed in the literature (e.g., Coles et al., 2023; Corneille and Lush, 2023). However, there were four reasons why we were convinced to narrow our focus:   * Discrepancies between frameworks often involve mechanistic details we were unable to evaluate in our review (e.g., whether demand effects emerge through faking vs. phenomenological control). * Reviewers convinced us that our attempts to test mechanisms through participants ratings had notable limitations. Although we believe these ratings have value, their limitations convinced us to describe them less centrally. * We were able to streamline the manuscript by (a) centrally focusing on a the most influential framework, and (b) making relatively brief remarks about modern extensions. * Reviewer 1 offered encouraging remarks about the value of a manuscript draft that less centrally focused on participant ratings.   We respond to specific reviewer comments in the sections below. |

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.

|  |
| --- |
| The updated manuscript includes a PRISMA flowchart, an updated/expanded search, and several of Reviewer 1’s methodological recommendations.  We respond to specific comments from Reviewer 1 in the sections below. |

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).

|  |
| --- |
| We overhauled parts of the manuscript to address this excellent feedback.  An examination of reviewers’ concerns did indeed reveal that participants provide unreliable ratings about the predicted magnitude and mechanisms underlying demand effects. We believe this is an important insight in itself given that Orne (1969) influentially recommended that participants could provide researchers with insights about demand characteristics.  According to the Law of Large Numbers, participants’ relatively imprecise ratings should converge into relatively precise estimates of the true mean at larger samples. We tried to exploit this statistical tendency by collecting additional ratings from Prolific workers, which substantially reduced the length of the confidence intervals of the estimated values.  These challenges and insights are now described throughout the manuscript. |

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.

|  |
| --- |
| Interesting point.  In our *Methodology,* we now clarify that participants who reviewed vignettes did indeed receive information about (a) whether students vs. non-students were sampled, (b) whether subjects received compensation, and (c) whether the study was conducted online or in-person. This was done to help participants better understand the study the study they were evaluating.  Although not currently described in the manuscript, we explored whether these study features are associated with (a) the probability that participants correctly identified the study hypothesis (a newly described moderator), (b) motivation ratings, (c) opportunity ratings, and (d) belief ratings. To do so, we used linear mixed effect models (with random intercepts for studies) to separately model each variable with the study features entered as effect-coded factors. We then used ANOVA tests to examine if these factors improved model fit.  The probability of identifying the study hypothesis was significantly related to whether the study was described as paid vs. unpaid. Ratings of opportunity to adjust responses were also significantly related to this variable – as well as whether the study was conducted online. We did not find that motivation or belief scores were significantly related to any of the aforementioned study features.  We hesitate to read into and describe this exploratory analysis for three reasons. First, none of the participant ratings were significantly associated with observed demand effects in the updated manuscript. Second (and relatedly), reviewers convinced us to include a far more critical description of the validity and reliability of these participant ratings – an insight we feel is important in itself. Third, the only one of the aforementioned study features that ratings were associated with – payment – was significantly associated with observed demand effects. Thus, we do not have a particularly strong basis for suspecting issues like confounding.  That being said, we are happy to reconsider this decision – especially because we agree the analysis is intrigue. |

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).

|  |
| --- |
| The updated manuscript includes ratings from a modified variant of a checklist described in the editorial: Downs and Black (1998). We found that many checklists contained a large number of items that were not applicable to literature we reviewed (e.g., baseline allocations, deviations from trial protocols, etc.). The same limitation applied to the Downs and Black (1998) checklist, but we felt we were able to assess most of the included items. The three scales we coded (reporting quality, external validity, and internal validity) were not significantly associated with observed effect sizes, so they are only discussed descriptively. |

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.

|  |
| --- |
| At the beginning of the updated *Results* section, we now include a summary description of the literature. This information is also now summarized in a Table.  Beyond your request, we reviewed the linked *Psychological Bulletin* editorial and ensured that our manuscript contains relevant information for each question described in Table 2. |

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?

|  |
| --- |
| Thank you for your guidance on this issue.  Following your recommendation, we combined Studies 1 and 2. Reviewers impressions of Study 3 were not generally positive – so we removed it from the report, included it in the synthesis, and briefly mention its original goals in the Discussion section. |

# 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.

|  |
| --- |
| Thank you for your constructive feedback! |

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…

|  |
| --- |
| The updated manuscript now reports the results of an expanded search based on terms that were conceptually similar to demand characteristics in Rosnow and Rosenthal’s influential book on experimental artifacts: “participant role” OR “demand effects” OR “good subject effect” OR “expectancy effect” OR “evaluative apprehension”.  Although we appreciate automated screening tools, we opted to screen these new records manually. |

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

|  |
| --- |
| In addition to the expanded search described above, we repeated our original search to identify newly published record. |

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.

|  |
| --- |
| Thank you for raising this important point.  Our methodology was inspired by many issues considered by Morris and DeShon’s (2002). For example, as described in the *Methodology*, we transformed effect sizes into a common metric using design-specific estimates of sampling variance. Sensitivity analyses on the assumed within-subject correlation used to make the effect size indexes more comparable produced virtually no change in our overall effect size estimate. And we did not detect significant differences between studies that manipulated demand characteristics within- vs. between-subjects.  That being said, we do agree with Morris and DeShon that researchers should also consider this issue at a conceptual level. In our view, there are reasonable arguments for both sides. Nonetheless, this methodological decision would be unlikely to change the rather humble conclusions of our meta-analysis: that we cannot currently make much sense of the demand characteristics literature. We describe this issue more in a footnote in our Limitations section because it sparked so much reflection among our co-authors. |

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>.

|  |
| --- |
| The updated manuscript includes ratings from a modified variant of a checklist described in the editorial: Downs and Black (1998). We found that many checklists contained a large number of items that were not applicable to literature we reviewed (e.g., baseline allocations, deviations from trial protocols, etc.). The same limitation applied to the Downs and Black (1998) checklist, but we felt we were able to assess most of the included items. The three scales we coded (reporting quality, external validity, and internal validity) were not significantly associated with observed effect sizes, so they are only discussed descriptively. |

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

|  |
| --- |
| Done. |

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 briefly address points 1 and 2 in the first paragraph of the *Methodology* section, wherein we argue that excellent narrative reviews exist but don’t allow us to quantitatively evaluate the magnitude, consistency, and potential moderators of demand effects. These evaluations informed what we believe is a fundamentally important and novel conclusion: that demand effects appear to be inferentially consequential, unreliable, and difficult to explain.  Will add a deviation from pre-registration section at the end (once we know the full scope of the deviations). Perhaps as supplemental material. |

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*.

|  |
| --- |
| We implemented this recommendation in the updated manuscript, and also noted it as a deviation from our pre-registration. |

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.

|  |
| --- |
| We apologize in advance for some of the issues you encountered while navigating our OSF page. The platform is a work in progress.  It appears that the OSF repository tab that would usually link to the pre-registration is not functional when we provide reviewers with an anonymized link. Nonetheless, we were able to create a separate anonymized link to the pre-registration:  https://osf.io/kwpys/?view\_only=8910e0dda30b4784b8ee6e3cfa60ba34  Please note that we did not use an OSF template for the pre-registration because none existed when we began our meta-analysis (van den Akker et al., 2023). For reasons we don’t understand, this seemed to prohibit us from creating time-stamped edits, which were instead performed on GitHub. Nonetheless, in an attempt to rectify the situation in a relatively streamlined manner, we have uploaded a single document describing the differences between our original pre-registration and the methods reported in our final manuscript.  Thank you for your attention to this issue. We appreciate the value you place on open science. |

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.

|  |
| --- |
| Updated manuscript uses the robust function from the metafor package to calculate CHE-RVE models. |

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

|  |
| --- |
| We were somewhat reluctant to try this approach because it’s unclear how cluster-robust estimation procedures would change the simulation results from Rodgers and Pustejovsky (2021). We relied heavily on these simulation results when developing our approach to assessing publication bias, and we feared that adding cluster-robust estimates to the 3-level precision-effect test might unacceptably reduce its power to detect small study bias (i.e., be too liberal).  That being said, we ran the analyses in an exploratory fashion and found that it did not change our inference. We now report this in a footnote. |

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.

|  |
| --- |
| We used the robust function from the metafor package in the updated paper. |

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).

|  |
| --- |
| Thank you for reviewing the effect size calculation script!  Code for all analyses are located in the script titled metaware\_manuscript.Rmd – which helps ensure that our manuscript is computationally reproducible. |

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).

|  |
| --- |
| To address this, the updated forest plot show one aggregated effect size per study. |

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.

|  |
| --- |
| Thank you for such exceptional attention to detail.  This is not a mistake. In the manuscript, we mention that we didn’t throw away an observation if they (a) described the direction and significance of the effect, but (b) didn’t provide any additional information needed to compute an effect size. (The exception was if the paper was recently published, in which we often were able to work with original authors to get the information needed to compute an effect size.) Instead, we (a) assumed *p*-values of .04 and .50 for significant and non-significant effects respectively*,* and then (b) estimated the effect size from that *p*-value.  In the Kenealy (1998) case, the entire paper only described the direction and significance of the effects. This is now described in the *Methodology* section.  Of relevance, papers that did not report enough information to directly compute an effect size received lower scores in the reporting subscale of the Downs and Black (1998) checklist. These scores were *not* significantly associated with observed demand effects. |

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.

|  |
| --- |
| This analysis was performed in the original paper, but perhaps not carefully described. When a different hypothesis was communicated, demand effects were significantly larger. |

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.

|  |
| --- |
| Thank you for catching this. This analysis is described in the updated manuscript. |

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.

|  |
| --- |
| The editor would know best, but we believe our approach is consistent with their recommendations. See the below quote from Rodgers and Pustejovsky (2021).  “When testing for selective reporting, practitioners should consider taking an exploratory approach by reporting 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 hesitate to do this because our publication bias analyses were informed by the simulation results from Rodgers and Pustejovsky (2021). We do not believe that Pustejovsky and Rodgers (2019) simulated instances where there are dependent effect sizes – so we view their 2021 work as a more authoritative guide for the present meta-analysis.  Perhaps the editor is willing to weigh in, though. We’re not sure if there is a package-based implementation of this method in R, but we believe we could implement the method based on openly-available code we found on the Open Science Framework. |

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 apologize for this oversight. We did consider that people could simply look at the document properties! The updated OSF pages blinded document properties.  P.S. Please note that OSF pages can also be unblinded by tinkering with the URL! Unfortunately, we suspect this is something we are unable to address. |

# 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.

|  |
| --- |
| Studies 1 and 2 have been combined in the updated manuscript, per the recommendation of the editor. To continue the conversation, however, we will refer to Study 2 as “vignette methodology” in our responses to your comments.  The updated manuscript describes the potential importance of this methodology, limitations we discovered, and the implications of those limitations. We summarize these points below.  One of the most seminal figures in the demand characteristics literature suggested that participants themselves may be able to help researchers understand demand characteristics (Orne, 1969). Vignette methodology has been one way that researchers have done so – and this methodology has recently experienced renewed interest (Corneille & Bena, 2023). For example, if participants are able to predict how they would respond in a psychology experiment, Corneille and Bena (2023) suggested that this may indicate that the effect is driven by demand characteristics. Indeed, this methodology has been used to bolster concerns about demand characteristics in some of the most seminal studies in the history of psychology – e.g., the so-called “Stanford prison experiment” (Bartels, 2019).  In the original manuscript, assumptions about the validity of this methodology led us to make claims about the mechanisms underlying demand effects. However, reviewer feedback helped us challenge these assumptions. In the updated manuscript, we cast doubts on this methodology and conclude that neither we nor participants themselves can currently explain how demand characteristics operate. We believe this is an extremely important methodological claim given that demand characteristics are sometimes characterized as an omnipresence threat to the validity of experimental psychology (Orne, 1962). |

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.

|  |
| --- |
| We believe that this is a case where reliance on such a sample is justified. Most studies in our meta-analysis relied on non-representative convenience samples, and we believe it is reasonable to argue that using a similar sample for vignette ratings increases the likelihood that such ratings will be predictive of the original participants’ responses. |

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.

|  |
| --- |
| Studies 1 and 2 are combined in the updated manuscript, but we will describe them separately in our response below.  We were pleased to read that you found Study 1 promising in itself. This was encouraging as we went through the painstaking effort of expanding and re-running the search, coding for quality, updating analysis plans, etc.  For reasons we described above, we believe the Study 2 vignette methodology still adds value to the manuscript -- although we are certainly open to being convinced otherwise.  The editor raised concerns about the appropriateness of Study 3 for Psychological Bulletin. Coupled with reasonable critiques from reviewers, we have opted to remove the study from the manuscript. However, because it meets our inclusion criteria, it is included in the meta-analysis and briefly described in the Discussion section. |

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

|  |
| --- |
| Agreed. This is implemented in the updated manuscript. I.e., tests of the moderating role of participant ratings are tested through three-level meta-analysis with robust variance estimators. |

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

|  |
| --- |
| Other people who we consulted also found this figure confusing. So, we opted to remove it and focus on describing the methodology in-text. |

# 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.

|  |
| --- |
| This study is dropped from the updated manuscript but included in the meta-analysis.  In case you are still curious, there was indeed a convenience factor in choosing this study: we had all the materials to run a close replication. That being said, there was also a more principled rationale: it is one of the only studies in this literature that (a) is pre-registered and (b) consistently produced strong demand effects. Thus, it seemed like a safe choice for attempting to replicate and understand the mechanisms that underlie demand effects. |

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.

|  |
| --- |
| This concern is no longer applicable now that we have removed the study.  In case you are still curious, we agree that the characterization is problematic (and would have adjusted if we kept the study). |

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.

|  |
| --- |
| This concern is no longer applicable now that we have removed the study.  In case you are still curious, we don’t believe there was a particularly strong justification for the choice to compare different types of demand characteristics (vs. a control). We simply followed the simplest study from the original paper, which did not include a control condition. |

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.

|  |
| --- |
| We don’t think this concern is applicable now that we have removed the study. However, we are happy to provide additional clarification if requested! |

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

|  |
| --- |
| We removed the numbering in our response to reviewers, but points 19 and 20 are the first two comments for “Study 2”.  This concern is perhaps longer applicable now that we have removed the study. But our response would have been similar to the ones we made for “Study 2”. |

# 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! |

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".

|  |
| --- |
| We think this is a great point and have updated our language accordingly. |

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.

|  |
| --- |
| We think this is a great point and have updated our language accordingly.  We also briefly discuss this issue in the Discussion section. There, we discuss why we feel that broader definitions of the demand characteristic construct present additional conceptual challenges that further deepen our concerns about the state of the literature. |

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>

|  |
| --- |
| We saw a pre-print of this paper shortly before finalizing the original manuscript and are thrilled to see that it is now published! We now cite and discuss the connection in the updated manuscript. |

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).

|  |
| --- |
| In the updated manuscript, we now primarily focus on Rosnow and colleagues’ framework. In our revision, we did initially attempt to more attention to Corneille and Lush (2023) and Coles et al. (2023). However, there were four reasons why we were convinced to narrow our focus:   * Discrepancies between frameworks often involve mechanistic details we were unable to evaluate in our review (e.g., whether demand effects emerge through faking vs. phenomenological control). * We became convinced us that our attempts to test mechanisms through participants ratings had notable limitations. Although we believe these ratings have value – e.g., in terms of whether participants can help us understand demand effects – their limitations convinced us to describe mechanism testing less centrally. * We were able to streamline the manuscript by (a) centrally focusing on a the most influential framework, and (b) making relatively brief remarks about ideas more recently proposed by Corneille and Lush (2023) and Coles et al. (2023). * Reviewer 1 offered encouraging remarks about the value of a manuscript draft that less centrally focused on participant ratings. |

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.

|  |
| --- |
| This is an excellent point that we, unfortunately, believe we cannot convincingly weigh in on in our current review. Nonetheless, we do discuss the role of mechanisms that can impact seemingly ‘uncontrollable’ outcomes (e.g., imagination and/or conditioning) in both the Introduction and Discussion. |

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>)

|  |
| --- |
| Discussions of the IAT are removed from the updated manuscript. |

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).

|  |
| --- |
| Thank you for this point. You’re absolutely correct.  In case you are curious about our initial confusion, we initially interpreted “opportunity to adjust response” to be a physical process (e.g., intentionally changing a self-report). However, as we re-read Corneille & Lush and several papers by Rosnow and colleagues, we realized we were mistaken. We’ve updated our description of this mechanism – and also acknowledge the intriguing possibility that participants may sometimes be unaware of their intention to produce experiences. |

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.

|  |
| --- |
| Excellent point. We now mention this in the updated manuscript. |

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.

|  |
| --- |
| Thank you for pointing this out.  When creating the vignettes, we generally attempted to match the original description of the methodology. In the case you reviewed, the original authors did indeed tell participants that “We are interested in the beneficial effects of listening to Mozart on test-taking capabilities.” 91% and 94% of participants interpreted this as meaning that the researcher hypothesized improvements in spatial and verbal ability respectively. (However, we admit this percentage likely would have been lower if we asked participants to use open-ended descriptions of the researcher’s hypothesis.)  When we asked participants to identify the researcher’s hypothesis, we initially intended to use it as an attention check. However, your comment reminded us that participants may not always correctly interpret the communicated hypothesis. We believe the likelihood that our participants correctly identified the researcher’s hypothesis might serve as a proxy for what Rosnow and colleagues describe as “receptivity” – and what Corneille and Lush describe as Level 1 of their framework. In an exploratory analysis we describe in the updated manuscript, we found that higher receptivity was indeed associated with larger demand effects. |

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?

|  |
| --- |
| Another excellent point. We now discuss this in the updated *Discussion* section. |

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.

|  |
| --- |
| Thank you for these recommendations. We followed both of these recommendations in the updated manuscript. |

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?

|  |
| --- |
| This finding did not replicate in our updated search.  If you are curious, though, here is a more thorough explanation of the original results.  Compared to a control condition where no hypothesis is communicated, participants responses to the dependent variable become attenuated when explicit demand cues suggest that no effect is expected. For example, if studying a truly effective mood-boosting procedure, the mood-boosting effect becomes weaker when participants are told that the researcher *does not* expect the intervention to work (nil demand). Similarly, the mood-boosting effect becomes stronger when participants are told the researcher *does* expect the intervention to work (positive demand).  So, imagine a scenario where participants receive nil demand, no demand, or positive demand. Our original results suggested that (a) the difference between the nil and no demand condition is bigger than (b) the difference between the positive and no demand condition.  That being said, we do not describe this result in the updated manuscript because it did not replicate in a larger sample. |

# 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".

|  |
| --- |
| We have removed Study 3 from the updated manuscript, as the editor pointed out that it would be unusual to publish a multi-study paper at *Psychological Bulletin*. Nonetheless, we want to engage with this comment in case it ends up being relevant elsewhere in the manuscript.  First, we are not intending to imply support for the null hypothesis when parameter estimates are non-significant. This is why we use phrases like “we did not find evidence…” (p. 40 of original submission; p. X of updated submission). We do agree, though, that we should not subsequently conclude that a specific prediction is “contradicted”. In the updated manuscript, we instead use phrases like “not supported”.  We believe your critique of our original characterization of non-significant interactions also hinges on our problematic use of the term “contradicted”. As you noticed, we frequently observed instances where:   1. a parameter testing prediction A (e.g., belief) is significant, 2. a parameter testing prediction B is non-significant, but 3. the difference in the magnitude of these parameters (i.e., the interaction) is likely non-significant.   In this instance, we believe that the most accurate way of describing these results is that:   1. we found support for prediction A, 2. we did *not* find support for prediction B, but 3. we did not find that the parameter testing prediction A was significantly stronger than the parameter testing prediction B   As awkward as this set of conclusions may be, we feel it is the most descriptively accurate. |

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 think these are fair points and have overhauled our discussion of the literature accordingly. For example, we (a) no longer make claims about the centrality of expectancy effects in previous theorizing, (b) have dropped the distinction between “motivation” vs. “expectancy” vs. “hybrid accounts”, and (c) refocused our review on the main components of the framework described by Rosenthal and colleagues. |

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.

|  |
| --- |
| Excellent point. We now discuss this at length – both when introducing our use the instruction-based replications and discussing the results. You’ll see that the updated manuscript exercises more caution when interpreting the results. |

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?

|  |
| --- |
| In the updated manuscript, we more carefully distinguish between the *observed* vs. *estimated* distribution of the effects. (The latter of which comes from the parameters estimated via three-level meta-regression.) The part of the figure and descriptive statistics you reference are based on the *estimated* 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.

|  |
| --- |
| In the Methodology section (Effect size index subsection), we mention that we didn’t throw away an observation if they (a) described the direction and significance of the effect, but (b) didn’t provide any additional information needed to compute an effect size. (The exception was if the paper was recently published, in which we often were able to work with original authors to get the information needed to compute an effect size.) Instead, we (a) assumed *p*-values of .04 and .50 for significant and non-significant effects respectively*,* and then (b) estimated the effect size from that *p*-value.  This set of circumstances applied to every effect size we attempted to extract from Kenealy (1988), who reported a large number of null effects without the additional information we would need to calculate an effect size.  As an aside, to address a concern that Reviewer 1 raised |