Response to editor and reviewers

Dear Dr. Cabbai,
I am pleased to report that I have received three expert reviews of your manuscript, "Investigating relationships between trait visual imagery and phenomenological control: the role of context effects".
Overall, the reviewers mainly had a positive reaction to your work. I agree with their suggestion that the topic is important and interesting, and your line of research appears to constitute a significant advance in teasing apart individual differences in different aspects of the possible connection between imagery and phenomenological control. I also appreciate several methodological pluses of your approach, as well as your care in having an independent statistician check the findings. After making some revisions in response to the points of the reviewers, I hope you will submit a revised version for further consideration at Collabra: Psychology.
To highlight one of the issues raised by the reviewers that you should respond to, I note that part of the purpose of your Study 2 was to do a comparison to some results of Study 1, e.g. "we predicted lower PCS or SWASH scores in self-reported aphantasics tested in the present study, compared to aphantasics tested via single-blind recruitment in Study 1." Reviewer 1 expressed several concerns about comparing Study 1 and Study 2 findings given both the differences in sampled population and the methodological differences of the two studies. For example, Reviewer 1 wrote: "So the question is then, how to tease apart demand characteristics from simply mis-classifying oneself, or any of the other possibilities? I think to answer this, a control sample would need to be recruited from facebook, reddit, etc. in the same way as the self-assessed aphantasia sample." Please address this point and clarify which conclusions rely on similarities between Study 1 and Study 2 that we can't be certain of.
In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers' comments, outlining each change made in your manuscript or providing a suitable rebuttal. Please see the instructions below for submitting your revision.
Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This may be the last opportunity for major editing, therefore please fully check your file prior to re-submission.
If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.
We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.
Sincerely,

Alex Holcombe

Reply: We thank the editor and the reviewers for their helpful comments and suggestions. We have replied to the editor and each reviewer's point below, and we have highlighted in the manuscript all the changes we have made following the reviewers' comments.

EDITOR

To highlight one of the issues raised by the reviewers that you should respond to, I note that part of the purpose of your Study 2 was to do a comparison to some results of Study 1, e.g. "we predicted lower PCS or SWASH scores in self-reported aphantasics tested in the present study, compared to aphantasics tested via single-blind recruitment in Study 1." Reviewer 1 expressed several concerns about comparing Study 1 and Study 2 findings given both the differences in sampled population and the methodological differences of the two studies. For example, Reviewer 1 wrote: "So the question is then, how to tease apart demand characteristics from simply mis-classifying oneself, or any of the other possibilities? I think to answer this, a control sample would need to be recruited from facebook, reddit, etc. in the same way as the self-assessed aphantasia sample." Please address this point and clarify which conclusions rely on similarities between Study 1 and Study 2 that we can't be certain of.

Reply: We have added discussion of this limitation and clarified potential differences between the groups contrasted in Study 2 (this is a contrast between undergraduate students and an older sample of internet volunteers). We have addressed this point in response to Reviewer 1 by drawing on prior evidence that, while a large range of variables have been tested over many decades, hypnotizability scores have not been found to be influenceable to any great degree. We have also added descriptive statistics and analysis (in the supplementary material) of a PCS sample which, like the self-reported aphantasic group in Study 2, is composed of older internet volunteers recruited from the general public and tested online. Additionally, this sample was recruited in Germany (rather than the UK) and tested on a German translation of the PCS.
Despite the demographic differences, the PCS scores are very similar across this German sample, the PCS norms sample (Lush et al., 2021) and Study 1 student sample (all three samples show a mean score of 1.9). Hence, while the study design does not rule out the possibility that differences other than participant beliefs arising from demand characteristics may account for the

results of Study 2, these data (and prior evidence) speak against this interpretation.

REVIEWER 1

Major, potentially requiring additional data collection:
The authors' conclusion is that some performance differences between aphantasics and typical imagery controls may be the result of demand characteristics, using SWASH/PCS as a particular example of this. I think this is a valid and important point to address, but I feel the current study could do a better job of identifying and addressing potential confounds.

Mainly, the manuscript currently does not consider many other possible confounds, including but not limited to:

1. ***the use of the VVIQ to measure mental imagery ability in Study 1***
2. ***the many demographic differences between the two samples from Study 1 and Study 2***
3. ***the temporal confound of administering the VVIQ and SWASH/PCS separately (Study 1) or at the same time (Study 2)***
4. ***a lack of attention checks/quality control in both studies***

I will go through each of these points in more detail.

1. ***It is possible that the VVIQ is not the best measure of aphantasia, and especially if individuals are not aware of imagery differences (and unaware of their own aphantasia) they may rate themselves higher because they don't know any differently. Individuals who are self-professed aphants know about the imagery spectrum, and can better compare their abilities to "typical". Given this, it is possible that the imagery sample in the first study was confounded with aphantasics, and the aphantasic sample was confounded with imagers.***

Reply: We assessed aphantasia following the standard approach adopted in the field of mental imagery (i.e., by scores on the VVIQ, e.g., Zeman et al., 2015, 2020; Milton et al., 2021; Dance et al., 2021, 2022; Pounder et al., 2022; Kay et al., 2022; Bainbridge et al., 2020; Keogh & Pearson, 2018; Wicken, Keogh & Pearson, 2021). Whether or not aphantasia can be validly

assessed by the administration of VVIQ, it is a routine procedure in the field to do so. We better clarified this point by adding the following considerations in the Discussion, p. 30, paragraph, 1 as follows:
"Our conclusions regarding relationships between trait phenomenological control and visual imagery are based on the assumption that VVIQ scores can be interpreted as evidence of visual imagery ability. While we acknowledge this assumption, we note that our results are potentially informative for any research that employs the VVIQ as a measure of aphantasia. At present, within the field of mental imagery, the VVIQ is the standard measure used to identify aphantasic participants (e.g., Dance et al., 2021, 2022; Pounder et al., 2022; Zeman et al., 2020; Keogh & Pearson, 2018; Keogh, Wicken & Pearson, 2021; Liu & Bartolomeo, 2023; Bainbridge et al., 2021; Dawes et al., 2020; Kay et al., 2022; Wicken, Keogh & Pearson, 2021; Milton et al., 2021). We note that most of the aforementioned studies also recruited aphantasics from aphantasia-related forums, thereby confirming the self-reported aphantasics via VVIQ scores. Indeed, aphantasia was defined with regard to a specific range of VVIQ scores in the paper in which the term was introduced (Zeman et al., 2015). Whether or not aphantasia can be validly assessed by the administration of VVIQ, it is a routine procedure in the field to do so."

Concerning the reviewer's point that someone might not be aware of imagery differences, the VVIQ instructions clearly indicate that "there are marked individual differences in the population in strength and clarity of reported visual imagery and that these differences are of considerable psychological interest", see Marks, 1973, 2019. Hence, we do not think our participants' replies were affected by being not aware of imagery differences, as that is clarified by the procedure of the VVIQ itself.

---

2. ***Relatedly, were any attention checks made with the student sample? The SWASH/PCS were performed in-class, but the VVIQ was done online for EPR points – how was quality controlled in this case?***

Reply: Only participants who fully completed the VVIQ were included in the VVIQ database and matched to the PCS/SWASH score. Following studies in the imagery field (e.g., Dance et al., 2021, 2022; Zeman et al., 2020; Milton et al., 2021; Bainbridge et al., 2021; Pounder et al., 2022; Floridou et al., 2022), we did not exclude participants based on other criteria. To address the concern that our VVIQ results may not be representative of other reported

VVIQ results, we have added a comparison of VVIQ scores in our study and existing literature to the Methods section, page 12, as follows:

“Since researchers in the field of mental imagery typically report VVIQ using total scores (sum of all items), we calculated the total score in our sample for comparison with other studies. Mean total score of the VVIQ sample matched with PCS (M = 55.68) and the of VVIQ sample matched with SWASH (M = 53.6) are comparable to those reported in other studies in the field (e.g., Milton et al., 2021 report M = 56.65; Liu & Bartolomeo, 2023 report M = 57.38; Zeman et al., 2015 report M = 57.92). While alpha is less frequently reported, Cronbach’s alpha of 0.95 for the VVIQ sample matched with PCS and of 0.91 for the VVIQ sample matched with SWASH is also consistent with other studies that reported alpha (e.g., Floridou et al., 2022, alpha of 0.93; Burton & Forgarty, 2003, alpha of .95; Nelis et al., 2019, alpha 0.89).”

---

3. ***Secondly, aside from being different in terms of imagery ability, the groups from Study 1 and Study 2 are also different in many other ways – the main one being students vs internet volunteers, who will have different levels of education and age (which the authors point out briefly in the discussion), but also familiarity with psychology studies, incentive to perform (were they paid?), attention to the task, understanding of instructions, etc, that could impact the group differences. So the question is then, how to tease apart demand characteristics from simply mis-classifying oneself, or any of the other possibilities? I think to answer this, a control sample would need to be recruited from facebook, reddit, etc. in the same way as the self-assessed aphantasia sample.***

Reply: Neither Study 1 nor Study 2 samples were paid. However, as the reviewer points out, the study design leaves some other differences (demographic) uncontrolled. A general finding from the field is just how difficult it is to find variables that impact hypnotic response to any large degree; the correlates of hypnotisability often vanish when tested out of context (e.g., see Lynn et al. 2021). After decades of trying, it has just been hard to find variables that meaningfully impact hypnotisability. That does not mean the reviewer’s point is not technically valid; just that it is implausible given the literature that a standard demographic variable could change scores as much as we found. To address the reviewer’s point on comparing scores of Study 1 with another sample characterised by demographic differences, we discussed and added in Supplementary a comparison with a PCS sample (n = 240) collected by another lab, which, like the self-reported aphantasic group in

Study 2, was composed of older internet volunteers recruited from the general public and tested online. Of note, this sample was recruited in Germany (rather than the UK) and tested on a German translation of the PCS. As we report below, despite the many demographic differences, the mean PCS score of this sample is very similar to that of the student sample in Study 1.

We thus have revised and extended our discussion of these issues in the manuscript (General Discussion) to clarify these limitations and to explain why we consider differences in participant beliefs to be a more plausible explanation for the difference observed between groups, on page 29:
"A limitation is that our study design does not rule out a role of demographic differences between groups, such as age, level of education and familiarity with psychological experiments. Of these factors, we measured only age, for which we observed a difference. However, we note that mean scores on imaginative suggestion scales (in the hypnotic context) are fairly stable across age (M at 19.5 years old = 5.9, M after 10 years = 6.0, M after 25 years = 6.5, Piccione et al., 1989). While there is some variation in scores across hypnotizability scale samples for scale translations in different countries (e.g., de Saldanha da Gama, Davy & Cleeremans, 2012), we are not aware of any potential demographic differences which could account for the size of the difference in scale scores between groups in Study 2. Regarding the Phenomenological Control Scale, we report in the supplementary material comparisons with a sample of the general public (n = 240) recruited and tested online on a German translation of the PCS. This German sample of internet volunteers (mean age 33.42, SD = 12.31, https://osf.io/b7yfx/; see supplemental material for descriptives and analyses) showed an average PCS score of 1.9 (SD = 0.7). As Study 2 sample, the German sample is older than Study 1 sample (mean difference of 13.5 years). Despite the difference in age and recruitment and other potential demographic differences, the German internet sample PCS score is similar to that of the student sample in Study 1. In contrast, the Study 2 self-reported aphantasic PCS score is lower than that of the German sample, with a mean difference of 1.4. As such, general demographic differences linked to online recruitment from the general population and age cannot easily explain the lower PC score in Study 2 sample. Therefore, while we cannot exclude the possibility that uncontrolled factors are involved, we consider systematic differences in beliefs arising from demand characteristics to be a more plausible explanation."

---

4. ***Footnote 1: The critical distinction in phenomenological control results, until now, seemed to be 'awareness of aphantasic status', but the addition of this footnote raises an important confound. Would the authors expect that in Study 2, known aphantasics would have performed differently if they had not been asked to take the VVIQ? If the point is rather that it is critical to temporally separate the VVIQ and SWASH/PCS, then this should have been tested in Study 2 (e.g., 1 group VVIQ/SWASH together; 1 group VVIQ/SWASH separated in time).***

Reply: The central issue is the presence of cues that can inform beliefs about the experimental hypothesis. Participants of Study 1 could not have known that their PCS or SWASH scores would be related to VVIQ scores when they completed these procedures. The demand characteristics for self-reported aphantasics of Study 2 include their pre-existing beliefs about aphantasia, knowledge that they were recruited by an imagery lab for whom they had agreed to be included on an aphantasia database and an information sheet that informed them that the study would involve a measure of visual imagery. Each of these cues could contribute to participant beliefs about the hypothesis. However, given the first two cues would likely be sufficient to support beliefs that the study would involve relating visual imagery to response in the phenomenological control procedures, the influence of the third cue is likely to be minimal in this case. As such, we do not believe that attempting to disguise the hypothesis by presenting the measures in separate sessions would be sufficient to prevent the self-reported aphantasics from forming beliefs about the hypothesis which relate to visual imagery.

Following the reviewer's comments, we acknowledge that the use of the word "unaware" to refer to the aphantasic sample of Study 1 may create confusion in the reader. For this reason, we have now changed "unaware" to "blinded" across the manuscript (including Figure 3 and Figure S1) to refer to the aphantasic sample of Study 1, which better reflects that this group was recruited following a single-blind procedure and as such was likely to be blind to the study goal (that is, they were unlikely to be able to work out that their scores on VVIQ would be related to PCS).

---

5. ***Expand in the discussion:***
   Phenomenological control is likely tied to many demand characteristics, so the effects/lack of effects would be amplified in various individuals who think they should perform a certain way,

not just known aphantasics. Lower suggestibility scores would be expected for individuals who identify as rational/logical, disbelieving in magic/UFOs/etc – many aphantasics take on this logical identity, but many do not, so there could be some confound here. This limitation could be discussed.

Reply: The reviewer's suggestion that demand characteristics for the self-reported aphantasic participants include beliefs about not being the kind of person who would respond to imaginative suggestions is consistent with our interpretation of Study 2. However, we are not aware of any evidence showing that aphantasics people take on a logical/rational identity. Furthermore, available evidence only shows a weak relationship between supernatural beliefs and hypnotisability (e.g., r = .17; Wagner & Ratzeberg, 1987). Besides fantasy proneness (e.g., r = .29 out of context; Silva & Kirsch, 1992) imaginative suggestion is not reliably predicted by any trait construct (including dissociation and absorption, which have historically been considered highly conceptually related to phenomenological control). For example, the "big five" captures very little of the variance in trait responding to imaginative suggestions (e.g., Nordenstrom, Council & Meier, 2002). See Lynn et al., 2020 ("Myths and misconceptions about hypnosis and suggestion") for a discussion of the influence of personality traits on phenomenological control in a hypnotic context (hypnotisability).

6. ***It is not so trivial to simply use single-blind recruitment for studies of aphantasia – I appreciate that the authors have raised the problem that there will always be an issue of power using this method. So what about in-person experiments that rely on specialized equipment and non-parallel testing?***

Reply: We agree with the reviewer that for more time-consuming methods, minimizing confounding demand effects will present practical challenges. We added the following point to the discussion to implement reviewer's point, after the sentence "Concerning aphantasia, unfortunately, this is particularly challenging to achieve given the low prevalence of this group in the general population (3.9%, Dance et al., 2022)." on page 31:

"*Where possible, it would be ideal to adopt blinding recruitment of special populations, for example, in undergraduate samples this may be achieved by pre-screening entire cohorts on VVIQ. However, this may not be practical in all cases (i.e., for labour intensive and time consuming in-person procedures). In*

such cases, potential demand effects should be acknowledged."

---

7. ***Furthermore, as soon as individuals find out they have aphantasia (even via a pre-experiment questionnaire), they will go to the internet to find out more about this, which may immediately change their demand characteristics for good, probably within even a day of testingThis would potentially change their responses on SWASH/PCS regardless of temporally separating the tests. This limitation could also be discussed.***

Reply: We have replaced "unaware participants" with "blinded participants" throughout (See point 4) to clarify that we are not claiming to have evidence that Study 1 does not contain any aphantasic participants who are aware of their aphantasic status.

---

8. ***Minor:***
   Study 1, methods- which was tested first, VVIQ, SWASH, or PCS? Or was test order randomized? How long of a temporal gap was there between tests? E.g., were students told to go online to do the VVIQ immediately after the SWASH/PCS session? Separated by a term of study?

REPLY: In Study 1, The VVIQ data collection started after the SWASH/PCS, which was collected early in the academic year, in mid-October (as explained in the manuscript, during a lab class). The VVIQ data collection started at the end of October. We added this information in the Method section of the manuscript, end of page 10.

Importantly, the students were not told to complete the VVIQ after SWASH/PCS. As explained in the manuscript (page 10), the SWASH/PCS were collected by different labs, and as part of ostensibly separate studies. These steps were taken to minimize the risk of participants' hypothesis awareness arising from demand characteristics. We explained this in Study 1 introduction "*To minimize context effects, our measures of interest were collected in separate studies and no mention of the other measure being investigated was present in the study invitation or instructions, making our participants blind to the experimental goal*." and in the Methods section "*VVIQ and PCS/SWASH databases were collected by different researchers*

belonging to different labs". As explained in the manuscript, page 11, the VVIQ was accessed by students via SONA recruitment, whereas PCS/SWASH was administered during a lab class.

---

9. ***Study 1, results page 14- Why is there a comparison of the strength of correlations between SWASH/VVIQ and PCS/VVIQ? Which r-to-z test was used to calculate these scores? Regardless, the relationship between PCS and VVIQ is rather modest even with nearly 4x the N of the SWASH/VVIQ sample, so it is not really clear what can be gleaned from this additional analysis. (Page 15 indicates this analysis will not be discussed – perhaps remove it completely?).***

Reply: The aim of this analysis is to test whether there is evidence for a difference between correlations, evidence of no difference, or no evidence either way. It is intended to prevent the misinterpretation of evidence for a difference in one condition and no evidence for a difference in a second condition as evidence for a difference between conditions. As explained in the Analysis section, end of page 13 and start of page 14, we used a Fisher z as r-to-z test *"We tested for the difference in the strength of correlations between VVIQ and PCS, VVIQ and SWASH using Fisher's z (1921, p. 26) and 95% CI using Zou's confidence interval (2007) as implemented in the package cocor in R (Diedenhofen & Musch, 2015)*"

---

10. ***Study 2, top of page 18- 'We collected as many participants as we could in one Term and as with Study 1' do you mean to say '…Term, as with study 1'? And wasn't the data collected across 3 Terms in study 1?***

Reply: We thank the reviewer for highlighting this issue and we apologize for the confusion - there was a typo in this sentence (we aimed to say that as in the case of Study 1, in Study 2 we also tested for sensitivity at Bayes Factor >3). We corrected the sentence on page 20, as follows:

"*We collected as many participants as we could in one Term and, as for Study 1, we tested for sensitivity for the main analyses at the conventional Bayes Factor* > 3"

---

11. ***General discussion, end of page 25, top of page 26: because PCS scores of 0 could indicate reactance (effectively suggesting 35% of participants in Study 2 may not have even tried to perform the PCS), what happens to the Study 1 v Study 2 comparison if the authors remove these scores from analysis?***

Reply: Since reactance is one way in which demand characteristics can play out at the level of motivation (see Corneille & Lush, 2022), reactance in self-reported aphantasics of Study 2 is consistent with our hypothesis (indeed we discussed this as a possible partial explanation for results, on page 28). However, we compared the two groups without the 0 PCS scores, and we confirm that results do not change: $t(44.75) = 4.47$, $p < .001$, $SE = 0.19$, 95% CI [0.47, 1.26], $B_{H(0,1.60)} = 3023.14$, $RR_{BF>3}$ [0.1, 1.6], with mean PCS of Study 1 aphantasics being 1.60 (SD = 0.71), and mean PCS of Study 2 self-reported aphantasics (with 0 scores removed) being 0.73 (SD = 0.67).

REVIEWER 2

This was a very interesting manuscript that cleverly linked research on visual imagery, ability to respond to suggestions, and demand characteristics.
I think this was a very well conducted and reported set of studies and I have just a small number of comments:

Reply*:* We thank the reviewer for their positive evaluation of our work!

---

12. ***On the first page the definition of phenomenological control comes across as a bit loose. The reader is told that phenomenological control is the ability to alter conscious experience and behaviour in order to meet goals and expectations. We are told that this can be a response to direct imaginaitive suggestions but that it can also be a response to indirect, implicit suggestions. Next we are told that this ability in a hypnotic context is known as hypnotisability. But are PC and hypnotisability really the same thing? Doesn't hypnotisability refer speficially to the capacity to respond to explicit suggestions (whereas PC refers to capacity to respond to both explicit and implicit suggestions.) I feel setting up some distinction between***

phenomenenological control and hypnotisability is quite important . At the very least surely hypnotisability must be something like phenomenological control ability PLUS response biases related to beliefs about hypnosis.

There were a few places in the manuscript where phenomenological control and hypnotisability are used almost interchangeably and it makes me uncomfortable.

Reply: The reviewer is correct that hypnotisability can be described as phenomenological control ability plus response biases related to beliefs about hypnosis. We agree that this is unclear in the text. We have added the following clarification to the first paragraph (page 3):

"*The term phenomenological control was introduced to avoid potential sources of confusion and to better reflect scientific consensus than existing terms (Lush et al., 2021; Dienes et al., 2022). First, hypnosis involves a specific context which is associated with many potentially misleading myths (Lynn et al., 2020) and refers to sleep, which is not related to phenomenological control (e.g., Banyai & Hilgard, 1976). Second, suggestibility can be confused with other concepts which share this label, but which show little relation to phenomenological control (e.g., social compliance; Coe, Kobayashi, & Howard, 1973; Moore, 1964). Misconceptions arising from these terms present challenges when communicating ideas about these phenomena. See Dienes et al. (2022), Lush, Dienes & Seth (2023) and Lush et al. (2021) for detailed discussion of phenomenological control and hypnosis. Note that control here refers to the ability to change one's experience, rather than to resist suggestions (contrary to popular belief, response to imaginative suggestion does not involve a loss of control; Spanos, Cobb, & Gorassini, 1985; Lynn et al., 2020).*

Trait phenomenological control can be measured by response to a series of direct verbal imaginative suggestions (e.g., Lush et al., 2021; Oakley et al., 2021). Some direct verbal imaginative suggestions invite participants to imagine a counterfactual situation (e.g., that there is a magnetic force positioned between one's outstretched hands) and suggest something that will happen as a consequence, rather than require an intentional response (e.g., that one's hands will be drawn together, as if by a magnetic force). Other imaginative suggestions contain no direct appeal to imagine (e.g., that a hand feels too heavy to move, or that the participant cannot remember anything)."

We also better specified this sentence on the same page*: "We note that, although here we focus on responses to direct imaginative suggestions, phenomenological control can also occur for indirect and non-verbal, implicit suggestions, such as in 'mesmerism' (a historical precursor to hypnosis, Gauld, 1992), or for beliefs arising from cues surrounding experimental situations (demand characteristics – Orne 1962; Corneille & Lush, 2023)."*

We have also added the following text in the next paragraph, page 4: "*Hypnosis can be considered just one particular context for phenomenological control, in which the focus is on direct, verbal suggestion (Dienes et al., 2022).*"

---

13. ***The statistical analyses were generally well explained. There were just two issues where I felt confused. First, at the bottom of page 12, we are told that the aphantasia range is <1 on rescaled VVIQ. I didn't realise that VVIQ had been rescaled. Can you please explain this more clearly. Is this related to the calculation of difference in strength of correlations? If so, was only VVIQ rescaled or were all variables?***

Reply: Only the VVIQ was rescaled. The rescaling was performed to reflect the scale anchors of the scale more accurately. The lowest score on the VVIQ indicates "No image at all", which seems better represented by a score of zero than a score of one. A lower limit of zero on the scale also allows the interpretation of regression intercepts (when modelling the effect of VVIQ on SWASH or PCS in Study 1). This was explained in Study 1 Methods section, when introducing the VVIQ, page 12. "*To facilitate interpretation of the intercept, we rescaled the VVIQ to range between 0 to 4 (instead of 1 to 5) across all our studies.*"

---

14. ***Second, in study 2, I did not follow the explanation as to why and how the linear relationship from Study 1 was used to "define estimates for unaware aphantasics". It seems like this led to comparisons between***

mean SWASH/PCS scores in study 2 and the model intercepts in study 1. Could you explain what is going on here in more simple language.

REPLY: The linear models of Study 1 where PCS or SWASH are predicted by VVIQ provide an estimate of the effects of the VVIQ on these measures, as well as an intercept value for each, which corresponds to the SWASH/PCS score when rescaled VVIQ is equal to zero (no imagery at all). As such, the intercept value (which is 1.49 for SWASH and 1.59 for PCS) provides an estimate of Study 1 blinded aphantasics' PCS/SWASH score (in particular, a rescaled VVIQ = 0 corresponds to the extreme aphantasia cutoff). We explained this in the sentence on page 20 "*The intercept corresponds to the predicted PCS score when VVIQ (rescaled) is 0, corresponding to extreme aphantasia cut-off.*" An advantage of using the intercept is that it uses the full sample of Study 1 (N = 508 for PCS, or N = 131 for SWASH) to make an estimate of PCS or SWASH scores when participants' VVIQ is equal to zero. As such, it provides a more precise estimate than an average PCS or SWASH score that would be obtained from sub-selecting Study 1 data, which would also require arbitrary decisions on how to subset the aphantasics group (e.g., using different literature-based cutoffs, or by matching VVIQ scores).
We better clarified the sentence on page 21 to explain this and highlighted changes: "*By using all the data to estimate the PCS mean score for this Study 1 aphantasics, this approach provides a more precise estimate than the group mean that would be obtained from sub-selecting Study 1 data by using arbitrary cutoffs (e.g., literature-based cutoff or matching VVIQ means)".*
We also added the following information regarding how we did the t-test with the Study 1 model intercept on page 21: *"To carry out the t-test between Study 1 model intercept and Study 2 self-reported aphantasics, we first calculated the mean difference between the intercept value and the mean of self-reported aphantasics (e.g., mean difference was 1.12 for PCS). The standard error (SE) of the mean difference was calculated as equal to the square root of the sum of the squared SEs for each estimate (square of SE of Study 1 intercept + square of SE of Study 2 self-reported aphantasics). The t-value was thus calculated as equal to mean difference / SE of mean difference. The degrees of freedom (df) were calculated as df for Study 1 intercept + df for Study 2 self-reported aphantasics".*

---

15. ***Just noting in case its helpful for the future that blinding for this manuscript was incomplete. There were lots of parts were information***

was redacted to preserve anonymity but the statisticians report and project OSF links showed the authors' identities. I don't believe this information has influenced by evaluation in any way.

Reply: We thank the reviewer for letting us know about this - we flagged this to the reproducibility statistician.

---

16. Minor comments

- ***The writing on p7 could be a little clearer. Three specific examples:***
  – The first paragraph contains the phrase "considering that the PCS does not involve a hypnotic induction" but the PCS hasn't been mentioned in the previous paragraph. Following the context here requires a jump.

Reply: On page 7 (now page 8) we added PCS in parentheses in the sentence before to clarify its mention in the sentence after:
"First, to our knowledge, the role of imagery abilities in response to imaginative suggestions outside the hypnotic context (as in the PCS) has not been tested yet. The relationship between these two traits may be stronger than between imagery and hypnotizability, considering that the PCS does not involve a hypnotic induction (which may cause reactance in some participants, Lush et al., 2021) and emphasizes the use of imagination to generate experiences (which might encourage the use voluntary use of imagery abilities)."

- ***Second, in th***
  – In paragraph 2 on p7, the phrase "in that case is unclear". In what case?

Reply: We removed "in that case" and clarified the sentence in paragraph 2, page 8 as follows:

"*If imagery is necessary, aphantasics should display lower levels of phenomenological control compared to non-aphantasics.*"

The phrase "in the aphantasia range" is unclear. What scale is being referred to here?

Reply: We apologise for the lack of clarity - aphantasia range refers to the range of VVIQ scores that are considered to identify aphantasics. We clarified that as follows on page 8: "in the VVIQ aphantasia range".

REVIEWER 3

17. ***P7 'First, to our knowledge, the role of imagery abilities in response to imaginative suggestions outside the hypnotic context has not been tested yet'. To make this hypothesis more interesting and avoid implying that you want to know whether imagery is related to imagination, which sounds self-evident, it would be useful to give an example of an imaginative suggestion with some brief discussion of what the alternatives might be – what are the alternative explanations for individual differences in this response that don't evoke imagery? Defining imaginative suggestion more clearly would also help distinguish this concept from that of phenomenological control.***

Reply: We thank the reviewer for this helpful suggestion. We have added the following example of imaginative suggestion in reply to both this comment and a similar comment made by reviewer 2 on page 1:

"Trait phenomenological control can be measured by response to a series of direct verbal imaginative suggestions (e.g., Lush et al., 2021; Oakley et al., 2021). Some direct verbal imaginative suggestions invite participants to imagine a counterfactual situation (e.g., that there is a magnetic force positioned between one's outstretched hands) and suggest something that will happen as a consequence, rather than require an intentional response (e.g.,

that one's hands will be drawn together, as if by a magnetic force). Other imaginative suggestions contain no direct appeal to imagine (e.g., that a hand feels too heavy to move, or that the participant cannot remember anything)."

---

18. ***My understanding from what you wrote in the opening paragraph was that phenomenological control is the ability to respond to imaginative suggestion in a way that feels like a real experience rather than pretense. Why don't you call it imaginative suggestibility? I'm guessing that you see the concept encompassing and extending beyond that of response to suggestion, but some more discussion delineating the two concepts would be helpful and save readers having to follow up the references you've cited, as well as making it clearer that phenomenological control is about ability to change your experience rather than keep it constant despite suggestion.***

Reply: We agree that these points were unclear in our manuscript, and we have added the following clarification about the intended meaning of control to the first paragraph of the manuscript, page 3.
"*Note that control here refers to the ability to change one's experience, rather than to resist suggestions (contrary to popular belief, response to imaginative suggestion does not involve a loss of control; Spanos, Cobb, & Gorassini, 1985; Lynn et al., 2020)."*

Reviewer 2 raised similar questions regarding the motivation for the term phenomenological control and details about phenomenological control and hypnosis which were not clear in our manuscript. We reproduced the text added in this response to R2 on these points here, page 3:
"*The term phenomenological control was introduced to avoid potential sources of confusion and to better reflect scientific consensus than existing terms (Lush et al., 2021; Dienes et al., 2022). First, hypnosis involves a specific context which is associated with many potentially misleading myths (Lynn et*

al., 2020) and refers to sleep, which is not related to phenomenological control (e.g., Banyai & Hilgard, 1976). Second, suggestibility can be confused with other concepts which share this label, but which show little relation to phenomenological control (e.g., social compliance; Coe, Kobayashi, & Howard, 1973; Moore, 1964). Misconceptions arising from these terms present challenges when communicating ideas about these phenomena. See Dienes et al. (2022), Lush, Dienes & Seth (2023) and Lush et al. (2021) for detailed discussion of phenomenological control and hypnosis."

We also better specified this sentence on page 4 (edits highlighted)*: "We note that, although here we focus on responses to direct imaginative suggestions, phenomenological control can also occur for indirect and non-verbal, implicit suggestions, such as in 'mesmerism' (a historical precursor to hypnosis; Gauld, 1992), or for beliefs arising from cues surrounding experimental situations (demand characteristics – Orne 1962; Corneille & Lush, 2023)."*

---

19. ***Study 1***

    Participants: "we matched the scores of two databases (one consisting of solely VVIQ scores and one of solely SWASH scores) that were collected from the same pool of Psychology students". Please add a brief explanation of how you did this. E.g., did you retain identifying information until data collection was complete or did you use ID codes? If the latter, how many participants were excluded because their codes couldn't be matched?

Reply: We retained information until data collection was complete, and matched participants by the provided email. We added this information to the manuscript, page. 10: *"In both cases, we matched scores from the two databases by using participants' provided email. We retained this identifying information until data collection was complete."*

---

20. ***Materials. Please could you say more (p10) about the hypnotic induction for the SWASH. Is this done by a researcher, or via audio or written instruction?***

Reply: We added and highlighted the following information in the Materials Section of Study 1, page 11. "*The SWASH was administered by computer, with the induction and suggestions scripts delivered by audio recording, as in Lush et al. (2021*)."

---

21. ***I accept your argument that scores on VVIQ-2 'reflect an overall imagery ability that encompasses other sensory modalities' (p8) but given your tactile and taste examples of hypnotic suggestibility, it was disappointing that you didn't use a measure like PSI-Q (Andrade et al., 2014) that assessed imagery in other modalities and I hope you would consider that for future research as it could give a more nuanced picture of the relationships you are interested in.***

Reply: We agree with the reviewer, and we have added and highlighted the following consideration about this in General discussion, page 27. "*Furthermore, in light of the multisensory nature of the available imaginative suggestion scales, we suggest that future studies should take into account other questionnaires which allow assessing subjective imagery abilities in other sensory modalities, such as the Plymouth Sensory Imagery Questionnaire (PSI-Q; Andrade et al., 2014) to gather a more comprehensive picture of the relationship between this trait and phenomenological control."*

Analyses: I was intrigued by your use of formal checking by an independent statistician and followed the OSF link to the report. The reproducibility template provided there is a really useful tool, which I've circulated to staff at my own institution. Thank you for making me aware of it.

Reply: We thank the reviewer for this and have passed this feedback to the reproducibility team at Sussex.

---

22. ***Discussion – p15 'in spite of their self-reported lack of imagery, aphantasics can generate experiences in response to imaginative suggestions' and 'Ultimately, this result provides evidence that imagery is not a prerequisite for generating responses to imaginative suggestions'. This is where your use of VVIQ-2 becomes problematic, because the criterion for aphantasia does not preclude imagery in other domains. For example, Dawes et al (2020) reported that, in their study, 'aphantasic individuals report decreased imagery in other sensory domains, although not all report a complete lack of multi-sensory imagery.' (my underlining). As far as you know about your own participants, they lacked visual imagery but not imagery in the other sensory domains assessed by the SWASH and PCS.***

Reply: We agree with the reviewer. Although visual imagery might be used to reply to the imaginative suggestions in the SWASH/PCS scale, none of these (except perhaps the negative visual hallucination) would necessarily require it. We have edited these sentences on p. 17 (we highlighted the edits):
"*This suggests that, in spite of their self-reported lack of visual imagery, aphantasics can generate experiences in response to imaginative suggestions which (except for one item which requires being unable to see a picture of a ball) do not necessarily involve visual experiences. Ultimately, this result provides evidence that visual imagery is not a prerequisite for generating responses to imaginative suggestions, at least for imaginative suggestion scales that do not require participants to generate visual hallucinations (though we note that one PCS and SWASH item involves a negative visual hallucination)*"

---

23. ***Study 2 p21. The analysis of self-reported and unaware aphantasics is very interesting. I'm wondering if you are making a fair comparison though. The mean VVIQ for the self-reported group was 0.03 (SWASH) and 0.05 (PCS) but you used a cut-off of < 1 for defining unaware aphantasia in study 1. Presumably the mean imagery score for this unaware group was greater than it was for the self-declared groups. If PCS and imagery are positively associated, surely the better average imagery ability of the unaware group would lead you to expect higher PCS scores compared with the self-aware group? I appreciate that this critique only applies to the groups comparison and not to the intercept modelling, and that you report a supplementary data with a stringent cut off of VVIQ = 0, but I was curious about why you didn't try to match the mean VVIQ scores for the groups comparison. Would the comparison provided in the supplementary materials not be more convincing than this one?***

Reply: We originally considered the option that was also suggested by the reviewer (comparing scores between samples by matching VVIQ), however, we considered that this approach might involve greater researcher degrees of freedom (in selecting which participants to match scores across). We decided to report results using the cutoff that has been most commonly used to define aphantasia in the imagery literature (e.g., Zeman et al., 2015, 2020; Milton et al., 2021, Dance et al., 2021, 2022; Wicken, Keogh & Pearson, 2021; Kay et al., 2022; Keogh & Pearson, 2018; Liu & Bartolomeo, 2023) and we considered that it would therefore be the most informative option with regard to relating our results to that literature. The literature-based cutoff was also used to classify the non-blind and blind samples of aphantasics in Study 1 and 2. We are open to the suggestion that the VVIQ = 0 test might be more convincing, but we think that because it reduces the sample size considerably (to N = 10 for the blinded participants and N = 26 for the self-reported aphantasics), it would be better to leave it in the supplement.

---

24. ***In sum, this is a methodologically interesting paper but I feel it would make a stronger theoretical impact if the authors defined the concepts they are investigating more precisely and were more circumspect in their conclusions, acknowledging that they speak only to the relationship between visual imagery (which is what they measured) and responses to imaginative suggestion in other modalities, and that the relationship between imagery in general and PCS requires further research where measures take into account the sensory modalities of the imagery and suggestion tasks.***

Reply: We hope we have addressed the reviewer's point in our previous replies, and we thank again the reviewer for their helpful suggestions.