Response to Review for Stage 1 RR: https://osf.io/74gcn

Anoushirvan Zahedi, PhD: Recommender, Peer Community in Registered Reports, Universität Münster:

First, the reviewers were doubtful about the applicability of the proposed project for addressing the declared hypothesis. Specifically, several points regarding the comparability of control and intervention groups that need rigorous consideration were highlighted.

As we detail below, we have changed both groups to be now more comparable except with respect to the feature we wish to test.

Second, the reviewers had concerns regarding the demand characteristics as a confound that needs to be addressed thoroughly. Relatedly, the instructions used in the study could be clarified better, and the reviewers have several suggestions on how to do so.

We have changed instructions for the control group to try to elicit the same expectations for change, which we now measure as an outcome neutral test, as we describe below.

Third, several critical points were raised regarding the implemented power analysis. Particularly, the reviewers were concerned that the analysis does not account for the uncertainty of the effect sizes, considering the differences between the experiment and the pilot. I encourage the authors to consider this and implement a power analysis that accounts for the uncertainty.

We now model the uncertainty in the standard error in estimating sample sizes, and provide code in case others would like to do so as well.

Finally, the reviewers wanted to access codes and pilot data, which is reasonable. I would strongly suggest that the authors use an online repository to do so.

The pilot study data and R code were already referenced in the “Pre-registered Results” sub-section of the “Pilot Study” section. Incidentally, the code has since been updated to aid reproducibility. They can be found at https://osf.io/cf7kh/ along with a readme.txt file describing the files and software versions used. All analyses have been checked for reproducibility by an independent statistician, whose reproducibility report we now also cite. We will use their service again for the full Stage 2.

Review by Zoltan Kekecs, 08 Feb 2024 15:04:

This study is testing the effectiveness of a suggestibility-enhancing training which is based on the cold control theory. The experiment is transparently reported, with appendices allowing for a direct replication of this study proposal (except for missing analysis code). I wish all research papers would be like this. My main concern is that the goal of this study is not clear. It seems as if the authors want to do a crucial test of the theory or at least one of its predictions. However, the proposed experiment is not doing that, it would take more matching between the control and intervention groups and some more blinding to circumvent alternative explanations of why group differences can occur. Below are specific suggestions which might improve the manuscript.

- *Previous attempts at enhancing PC (suggestibility) also include using sensory deprivation (Darakjy, Barabasz & Barabasz, 2015), and using reversible inhibition of the DLPFC (for example the authors’ own work). Reference: Darakjy, J., Barabasz, M., & Barabasz, A. (2015). Effects of dry flotation restricted environmental stimulation on hypnotizability and pain control. American Journal of Clinical Hypnosis, 58(2), 204-214.*

References added and addressed in the Enhancing Phenomenological Control sub-section:

> Some attempts to enhance phenomenological control involve short term changes due to non-invasive disruption of brain operation by rTMS (Dienes & Hutton, 2012; see Coltheart et al., 2018; Faerman et al., 2024; Kekecs & Souza, 2024); restricted environmental stimulation (Darakjy et al., 2015); or use of psychoactive substances such as LSD (Carhart-Harris et al., 2015), nitrous oxide (Whalley & Brooks 2009), or alcohol (Semmens-Wheeler et al., 2013).

- *Calling low suggestibles „lows" is abrupt and is not explained in the text. It is of course clear to people well-read in the hypnosis literature, but the use of this phrase could be introduced for those who are not familiar with it.*

Good point; manuscript modified as a result.

- *Based on the description of the procedures of the pilot study, it seems that the control group only had one try with each suggestion, while the intervention group had 5 tries. Why was this difference between groups introduced? This seems to be a possible confound. Group differences might arise simply due to practice/fatigues effects. Similarly, the two groups might have different response expectancies, and different beliefs about the role of the practice phase. I suggest that the practice phase of the two groups should be matched very carefully, with the only difference being the instructions for imagined involuntariness. That means that the groups should be matched in having to imagine the enactment of the suggestions. The only difference should be that people in the intervention group should imagine not only the enactment of the suggestion, but also that this enactment is involuntary.*

Originally it was assumed that repetitions of the same suggestion (in the control group practice) would only lead to fatigue/boredom effects (e.g., Fassler, O., Lynn, S. J., & Knox, J. (2008). Is hypnotic suggestibility a stable trait?. *Consciousness and cognition, 17*(1), 240-253.); the training, however, was assumed to be sufficiently novel for each repetition that any effect would be minimised. As a result, it appeared prejudicial to potentially bore the control group into lower responsiveness in the test phase.

To address this comment, the control group participants will now be given the same number of opportunities to repeat each suggestion, in a similar decision-making format to the training group participants. In addition, response expectancies will be recorded for each group after a description of their involvement has been given. We will treat these as a covariate in the analysis.

- *Importantly, the conditions should also be matched in what is the implied purpose of this practice run. It should be made explicit in both cases and these statements should be matched exactly. Something like: "The role of this practice is to enable you to respond to the suggestions as well as possible. By imagining the enactment of the suggestion and by practicing the enactment, you will became more capable of responding to suggestions." – Same for both groups. This might help matching response expectancies between the groups.*

Instructions like this can create strong expectancy and or demand characteristics if they only appear in one group: "Now, the idea is that we're going to try to make that feel involuntary through the use of imagination. So, I'd like you to do it again, when I indicate, but this time I'd like you to also imagine that you're not involved in the process at all, as if your hands are moving all by themselves. Can you imagine that while you do this? Okay, please do that now." The instructions should match between the groups as much as possible except for the manipulated mechanism (imagining involuntariness). If you say "we will try to make this feel more involuntary" to one group, you have to say the same in a credible way to the other group. For example you could tell the other group: "Now, the idea is that we're going to try to make that feel involuntary through the use of imagination. So, I'd *like you to do it again, when I indicate, but this time I'd like you to also imagine that you are responding to the suggestion completely, just make it happen with your imagination. Can you imagine that while you do this? Okay, please do that now."*

Good point. The instructions (in Appendices G, H, J and K) have been modified to take into account these comments. We have attempted to make the test fair while also giving as little instruction to the control group of how to make the suggestions feel involuntary so as not to unduly bias the test.

- *Also, I think that the trainer can influence the outcomes by explicitly or implicitly implying desired results. So it would be ideal if there was no human in the loop, or if the humans in the loop would be blinded to either group allocation or at least expected study outcomes.*

To address this concern, we have redesigned the experiment so that it will be entirely automated. Participants will be greeted by the experimenter over Zoom who will provide some introductory information and the link to the questionnaire. The participants will then leave the Zoom session and work through the questionnaire, which includes all the instructions and exercises as pre-recorded audio. The questionnaire will randomly allocate the participants to groups only after being greeted by the experimenter, thereby removing any possible bias from the experimenter's introduction. On completion of the questionnaire, the participants will return to the Zoom session. It was assumed that, by including the experimenter at the start and end of each session, participants would be more attentive to the study and therefore provide higher-quality data rather than simply sending out links to participants to do in their own time.

- *One possibility for the different results between motor and hallucination suggestions could be that enacting the hallucination suggestions might requires imagination in itself, and there* *might be a competition for the imagination resource between the actual suggestion enactment and the suggestibility-enhancement strategy. So it is possible that this particular suggestibility-enahncement strategy would not work, or would have limited utility for (especially positive) hallucination type suggestions. If this is true, I would predict that negative hallucination type suggestions would benefit more from this strategy than positive hallucination type strategies, because negative hallucination type suggestions might be enacted in different ways other than through imagination, and also because I expect they require less imagination resources (I have no actual evidence to back this up). So it might be worth considering to include a negative hallucination type suggestion in the registered experiment.*

This is an interesting point. To address whether the training only works for motor suggestions, we have divided the experiment into two. The first will train on motor suggestions (two direct and one challenge) and test only on motor suggestions (two direct and two challenge). The second, which will only be run if the first produces evidence for H_1, will train on motor and hallucination suggestions (one direct and one challenge motor suggestion, plus one direct hallucination and one challenge hallucination) and will test only on hallucination suggestions (two direct hallucinations and two challenge hallucinations). Details of the suggestions can be found in the manuscript and specifically in Appendices G, H, I, J, K, and L.

We will save the comparisons between individual suggestions for a future study. We will run enough participants for the current studies to ensure good evidence for our main tests, but that won't be enough for testing and comparing individual suggestions.

- *Another possible explanation is that hallucination type suggestions are usually "hard suggestions", and that the imagination-based strategy only works for easier suggestions. This could potentially be tested with including suggestions in the experiment that have similar difficulty. Or at least including a hard motor suggestion in there as well.*

It is important to note that the difficulty of a suggestion might not only depend on the expected behavior or response itself. For example if I am not mistaken the arm immobilization suggestion in the SHSS:C is on the "harder side", maybe because the test suggestion is short. In comparison a very similar arm immobilization suggestion in the EHS is on the easier side, maybe because there is a lot of repetition and formulation of the suggestion in different ways. So it might require prior data or extensive pretesting to figure out what is the difficulty of a given suggestion in the particular experiment.

Again, interesting points, but ones that would require considerably more participants than we are running to properly investigate.

- *The authors describe their sample size rationale very clearly. This is exemplary. However, I think the calculations do not take into account sampling error. That is, the actual power to detect the effect if it exists is small, because the authors did not account for the noise. I recommend running a simulation to assess power, and aim to achieve at least 80% power (preferably 90%).*

To address this comment, we have modelled the uncertainty in the standard error. We have now run Monte Carlo simulations with a standard error sampled from the posterior distribution of the pilot standard error, for 1000 iterations per given sample size. Our results indicate that with 120 participants we have 100% probability (to an accuracy of 1%) for achieving $B > 5$ for both subjective and involuntariness measures, assuming there is an effect (of the sort found in the Pilot for motor suggestions). Our original calculations indicated 64 participants would be sufficient for a 50% probability, so this level of power for 120 participants is not unexpected. Our simulations indicate that an 80% probability could be achieved with 73 participants.

For testing the null hypothesis, 120 participants achieved 98% probability for the subjective measure $B < 1/3$, but only 67% for the involuntariness measure $B < 1/3$. Increasing to 130 participants increased the probability to 81% for *B* < 1/3. We have therefore increased the upper limit to 130 to accommodate these calculations.

The simulations also showed that 387 participants would be required to achieve 80% probability for *B* < 1/5. To achieve this number would require full automation of the study (without experimenter greetings and conclusions) and a larger pool of potential participants than a single academic year of psychology students at our university can provide. Thus, we have stuck to a threshold and stopping rule of *B* < 1/3.

- *It is not clear why highly suggestible individuals are involved in this experiment due to the likely ceiling effect such a training would have with them. Including only lows and mediums could increase the impact of the intervention, and thus, statistical power (of course with the trade-off of having to do a pretest or some sort of screening pre, during, or post study session). I see how this could interfere with the practicalities of the experiment, so I don't expect this to be necessarily adopted, but something that the authors might consider (especially if they are running other experiments from which it is easy to pre-screen individuals). Maybe it would be good to propose an exploratory analysis to look at the correlation of the effect of the training with baseline performance (on the three training suggestions). This can be done with no cost and could aid future studies/trainings using similar protocols.*

Good point. To address this comment, we will now recruit from the pool of pre-screened students, excluding those who are highly responsive (PCS score in the top 10%).

- *I expect that with improved matching between conditions, which will make expectancies and demand characteristics closer between the groups, and also with blinding (or automation) of the trainer, the effect size would drop substantially. The changes in effect size alter sample size targets, and the changes in the procedure might involve unexpected events or participant* *reactions. So I suggest running a new pilot study with this new protocol, and only then engage in a full crucial test. I know this is very expensive, but engaging in the crucial test without the modifications to the protocol will not really be a crucial test the theory (it would "only" be an efficacy study of a training that is based on the theory). And running the study with so many modifications after the initial pilot without piloting is risky. Because of the time and resources cost of running an extra pilot I would accept if the authors would decline this, but then they need to acknowledge the limitations of the implications of the findings of this experiment regarding the theory.*

We have incorporated closer matching between the groups as noted earlier. One can argue whether or not this will reduce or increase the expected effect. We noted evidence (from SJ Lynn's lab) that repeated testing over a short period reduces hypnotic response; this consideration leads to the prediction that the new design should have larger effect sizes than our originally planned experiment. Running a pilot wouldn't add anything other than getting a better estimate of the difference between groups; but we may as well get that estimate with the number of participants needed to make it precise by actually running a pre-registered study. The model of H_1 assumes a range of effect sizes are possible, anywhere from 0 to 1 Likert unit difference, and this still seems a reasonable range given the scientific context. We will of course report robustness regions for all Bayes factors.

- *Relatedly, I would suggest that the authors pre-formulate brief conclusions for a scenario where Bf 3 was achieved on the involuntariness scale, for where bf 1/3 was achieved for the involuntariness scale, and for the test being inconclusive.*

The following text was added to address this comment:

> The two measures (subjective realness and involuntariness) permit testing of two hypotheses, namely that imagining suggested behaviour as

happening involuntarily will facilitate the behaviour being experienced as more "real", and more involuntary. Cold control theory predicts that suggested behaviour that feels more involuntary will also feel more "real" (and vice versa), and also that suggested behaviour that feels more "real" will also feel more involuntary (and vice versa). If one measure finds evidence for H_1 but the other measure finds evidence for H_0, then this would challenge cold control theory as it currently stands.

Evidence for H_1 for both measures would suggest that the capacity for phenomenological control can be increased through training in imagining suggested behaviour as occurring involuntarily. Evidence for H_0 for both measures would suggest that the training made no difference to the capacity for phenomenological control. As the hypotheses are independent, the evidence for each can be reviewed separately, especially if one or both are insensitive.

- *It would be great to include link to proposed analysis code (together with simulated data or designed to work with raw data from the pilot study).*

The pilot study data and R code were already referenced in the "Pre-registered Results" sub-section of the "Pilot Study" section. The code has since been updated to aid reproducibility. The files can be found at https://osf.io/cf7kh/ along with a readme.txt file describing the files and software versions used.

- *"We estimated the need number of subjects in the following way." – "ed" missing from "needed".*

Corrected.

- *"Participants will be randomly assigned to the either the control group or the intervention group." – no "the" is required before "either".*

Corrected.

Review by DR. Sophie Siestrup, 26 Feb 2024 11:06:

In this proposed study, the authors plan to investigate whether not having higher order thoughts of intending facilitates not having higher thoughts of intending. They have collected pilot data with a protocol that was highly similar to the one they propose for their study. The authors plan to collect data from two groups of participants, a control and an intervention group. Both groups will be trained to imagine certain behaviors from verbal suggestion. In the intervention group, these behaviors will be actually performed by the participants. Afterwards, participants will be presented with more test suggestions and asked to imagine these behaviors as well. They will rate the strength of the effect and how involuntary the effect felt. The authors will use Bayesian statistics to evaluate the results. While the general idea of the study is interesting, I am at the moment not convinced that the authors will be able to answer their research question with the presented experiment. My largest concerns are (1) the fact that participants in the control group are never actually performing the imagined behaviors and (2) the fact that participants in the intervention group are aware that they should achieve high involuntariness. Please find a list of my major and minor suggestions below.

Major points

Introduction

-The introduction is made up of rather short paragraphs, each listing some previous findings/theories. I found it complicated to follow the authors' line of thought and suggest they make stronger connections between individual aspects relevant for their study.

The introduction has been streamlined, the logic made more explicit, with guidance for where the argument is going added in various places.

-Related to my previous point, I feel like the wealth of information provided in the introduction is overwhelming and often appears incoherent. Restructuring the introduction (see above) might solve this issue, but I also suggest the authors to critically evaluate which information is crucial to understand their research question. (For example, I found the short insertion on trance on page 4 more confusing than helpful)

That particular section on trance has been deleted, as have some other passages, so that the remaining text is more related to the flow of the argument.

- "Participants score their subjective responses to each item on a Likert scale from 0 to 5, with 0 indicating not feeling the effect at all and 5 indicating feeling the effect as if it were completely real." (page 5) – what effect are you referring to here? The exact nature of the rating does not become clear to me throughout the manuscript.

The subjective rating indicates how strongly they felt the effect of the exercise. In the control group practice phase and the test phases for both groups, the exercises involve imaginative suggestions that encourage automatic behaviour to occur. These scores relate to how strongly they felt that behaviour occur. In the intervention group training phase, the exercises involve making movements or imagining scenarios and also imagining that those movements and scenarios feel like they are happening involuntarily. The theory predicts that if they feel involuntary then they will feel as if they are happening to the participants. In this phase, the subjective rating indicates how strongly they felt this. Hopefully, this is made clearer in the text now.

-From the information presented in the introduction, I do not entirely understand the motivation for the planned research and the research question per se. This is probably due to the incoherent structure of the Introduction, as mentioned above. Similarly, I do not understand what the motivation for the pilot study was. I think the manuscript would profit from a short section where the authors summarize the most important information and make clear what is not known, i.e., needs further research, and clearly formulate the research question.

Additional text at the beginning of the introduction addresses this.

-I do not find the design of the (pilot) study convincing. Participants in the control group did not physically produce any behavior. Therefore, differences between groups could simply arise due to the lack of any motor response in the control group. One would need an additional control group that (1) imagines a behavior actively, carries out that behavior but (2) imagines it was voluntary.

The control group do indeed produce behaviour in response to the suggestions; they are on average of medium ability on phenomenological control. Note also the dependent variable is not the amount of movement. Of the motor suggestions, one is hands stuck together, where successful responding implies not producing overt behaviour, and the other is hands moving together, where successful responding does imply overt behaviour. A range of types of suggestions are used in the tests (hallucination, direct motor, challenge motor). Thus, the selection of suggestion types means displaying overt behaviour or not is not an issue.

The reviewer is right however, that the design was not completely controlled (specifically, in number of attempts for each suggestion), as also pointed out by the first reviewer. As we detailed in our response to them, we have now addressed that confound.

Methods

-I think it is not a convincing argument to limit the number of participants based on "the number of undergraduate psychology students available and the number of experiments competing for their participation", even though the analysis shows that > 300 participants would be optimal. Another solution would be to stretch the experiment over multiple terms or to recruit also non-students, as mentioned below.

We typically recruit students because they are obliged to take part in a set number of psychology experiments in each of their first two academic years. Recruiting from outside of this pool is expensive and will present practical difficulties.

-In the section "Pre-registered Experiment": why will the music hallucination be swapped with the hand lowering suggestion? I see the authors' point that they do not only want motor suggestion in the practice phase, but like this, it is still not balanced (i.e., still more motor suggestions in practice than hallucinations). Similarly, there will be more motor suggestions in the test phase as well. I think according the pilot findings, it would be highly important to keep the types of suggestion balanced.

To address this concern and also to address a concern of the first reviewer, we have divided the proposed experiment into two. In the first experiment we will practice/train with motor suggestions and then test only with motor suggestions. This will be a close replication of the pilot study. In the second experiment (only run if the first produces evidence for H_1) we will train with a mixture of motor and hallucination suggestions, but test only with hallucination suggestions. This approach will demonstrate if the training works for motor suggestions and hallucination suggestions separately.

-The authors note that they will not place any constraints for the selection of participants. I think it would make sense to carefully think about this again. E.g., it might be difficult to recruit deaf participants or such with a history of auditory hallucinations due to for example psychiatric diseases. I also wonder whether all students are by default over 18?

The details have been amended to indicate that deaf students will be excluded due to the reliance on pre-recorded audio, and blind students will be excluded from the second experiment due to the inclusion of visual hallucinations. Over decades of testing phenomenological control, we have not explicitly screened out people on psychiatric grounds, and have not found problems. The vast majority of students will be over the age of 18; a small proportion may be 17, and none will be younger.

-It does not become clear why the changes to the experimental protocol were made after the pilot study.

The pilot study discussion explains that while the training was focused entirely on motor suggestions (with no hallucination suggestions), the training appeared to only generalise to motor suggestions. It was therefore assumed that in order to train participants for hallucination suggestions, that a hallucination suggestion would need to be included in the training set. Following review, we have now divided the experiment into two where the first tests the training on motor suggestions and, if it produces evidence for H_1, the second will test if the training works for hallucination suggestions.

-In the intervention group, participants are always informed in the training that the aim is to make the behavior feel involuntary. This likely influences their ratings of Involuntariness (for training and test) since they know involuntariness is expected from them.

This is a good point, also raised by the other reviewer. As a result, the instructions to the control group have been modified so that both groups are aware that the goal is to make the behaviour feel involuntary.

Minor points

Title

-I found the title quite confusing and had to read it multiple times to get an idea of what the manuscript is about. The authors might consider revising the title.

We thought about the title and consider it to most directly say what we want to say.

Introduction

-The Introduction starts with a paragraph about the early history of hypnosis. This seems unnecessary (and even a bit confusing) since none of these historical facts are relevant to understand the present research.

We have now streamlined the introduction.

-What exactly are "imaginative suggestions"? (first introduced on page 3)

The addition of text at the start of the introduction hopefully helps by providing a definition:

> The response to hypnotic suggestions therefore appears to not require hypnosis at all, and instead requires only the application of the capacity for phenomenological control (Dienes et al., 2022). That is, participants can simply be asked to make their experiences feel automatic or involuntary and, if sufficiently

> motivated, they will do so to the best of their abilities. This capacity is roughly normally distributed, whether outside or inside the hypnotic context (i.e. whether or not a hypnotic induction is used and the suggestions are called hypnotic; Lush et al., 2021).
>
> Suggestions presented with no induction and no mention of hypnosis are not “hypnotic” as no hypnotic ritual is involved; instead, following Braffman and Kirsch (1999), we refer to them as “imaginative suggestions” to distinguish from other forms of suggestion that are unrelated to these investigations (e.g. self-affirmation; Sherman & Cohen, 2006).

-*What are “successful participants”? (page 5)*

The text has been changed to “participants who were successfully responding”.

-*Please avoid colloquial language (e.g., “doesn’t” on page 6, “none were still highs but that almost half were mediums” on page 8)*

Contractions in the main text have been expanded. (Those in the scripts have been left in as the style of the scripts is more natural.) The references to “highs” and “mediums” have been expanded.

-*I do not understand what is meant by “Given the theory that experienced involuntariness is the basis of the experience feeling real” – please clarify*

Hopefully the addition of text to the introduction has clarified that the participants will be presented with imaginative suggestions that will cause some of them to experience involuntary behaviour that they are unaware of causing. Cold control theory predicts that the involuntariness associated with these behaviours makes the behaviour feel as if it is happening to them. As the nature of the presented suggestions is that things are happening to the participants, how strongly they feel this is expressed by participants as how “real” it feels. For example, if participants are told “Your hands are stuck together” then the extent to which they feel this is true and “real” is predicted by the theory as being based on how involuntary their behaviour feels in failing to separate their hands.

Methods

-Why is the availability of participants strictly connected to academic terms? Did the authors consider also recruiting non-students?

Recruiting psychology students is trivial as we have systems in place purely for this purpose, and the department also requires their participation in a number of psychology studies as part of their degree. In these cases, the students are rewarded with credits that they need to obtain. Recruiting non-students would typically require payment for which there is insufficient budget.

-In my opinion, one could find better words to say that one participant “failed to state their gender”. Did they actually try but did not succeed? Or rather prefer not to the state their gender? I would consider expressing this in a more neutral way, e.g. “one participant did not declare their gender”

Corrected.

-Were the groups in the pilot study really assigned randomly? Genders seem oddly balanced for a random attribution to groups.

Additional text has been added to describe the randomisation process:

> A simple computer algorithm based on the Unix rand() function (seeded with the time the program started to cause each run to be different) generated random pairs of allocations, each either {control, intervention} or {intervention, control}. The program was configured to generate at least 60 pairs of allocations.
>
> The Unix rand() function is pseudo-random; each allocation pair was decided based on the least significant bit of its output. By generating pairs of allocations, the algorithm guarantees the same number of allocations for each group, leading to a 50:50 ratio of allocations.

-In the section "Pre-registered Experiment" I do not understand the paragraph starting "Based on the theory that…". Please consider rephrasing, this sentence is very long.

Noted.

-What about ethics approval?

The Participants sub-section of the Pre-registered Experiment states that the research was approved. A similar statement has been added to the pilot study.

-Why will the experiment be conducted in Zoom? Meeting the participants in person might guarantee a more controlled environment to do the experiment. If it needs to be done in Zoom for important reasons, this might even be another argument to also recruit non-students, since they do not even have to be physically present at the authors' institution.

Zoom allows more convenience for participants while ensuring engagement with the experiment for its full duration, unlike for example simply sending out a link. In the initial Zoom meeting, the experimenter establishes that the participant is alone in a quiet environment where they will not be disturbed.

Results (pilot study)

-Aside from giving information about evidence for H1, I think readers would profit from a short summary of what exactly this means (about differences between groups/conditions).

Additional text has been added.