Replies to reviews (bold)

Reviewed by Maxine Sherman, 30 May 2023 15:45

I'm aware that this is a long and complicated manuscript, so I'm especially grateful to the reviewer for her attention to detail, which has greatly helped to improve some features of the methods and design plan that I had not thought through properly - thank you.

Michotte's studies are a great target for replication and I read the manuscript with interest. Though I think this will be a very important piece of work given the influence of Michotte's research, I found that keeping track of 14 studies (some with multiple manipulations) quite burdensome. Once the results are in, the Stage 2 manuscript will be considerably longer and I think attempting replication of this many effects may too much for one paper unless it's possible to make the Introduction and Methods considerably more concise. Perhaps some of Michotte's methodological details could be relegated to supplementary materials, and key information placed into a table/figure?

This reviewer asks for introduction and methods to be more concise and the other reviewer has asked more more sub-paragraphs in the introduction so I'm not sure how to find the best compromise between the wishes of the two reviewers. In view of the other reviewer's comment, I think relegating the passages on Michotte's methodology to supplementary materials might not be preferable - and of course his methodology (e.g. n of 1 and no statistical analysis) was a major motivating factor for this proposed replication, so for that reason I think the contrast between his methods and mine needs to be at the forefront. That is the main reason for the section on general features of method: my methods differ from Michotte's in several respects so there is a need to explain what changes were made and why, given that this is a replication study. Any researchers who may want to conduct further replications or extensions of the research will need to know all that.

I have reduced the last two paragraphs of the introduction to one shorter paragraph, as can be seen on pp. 7 - 8.

With the aim of trying to help the reader deal with the complex information about the experiments I have added two tables, as Table 1 and Table 2. Table 1 lists the features of method that are common to all experiments, with variations indicated. Table 2 briefly lists the variable or effect that is being replicated in each experiment, with the experiment(s) from Michotte on which each was based. Having this information in tables has resulted in some reduction and simplification of the main text, which I hope will help readers.
One possibility that might be worth considering is making one paper into two. Experiments 1 - 8 are on launching and experiments 9 - 14 are on entraining, so that could be a reasonable basis for dividing the paper into two. There would have to be some repetition between papers but I think not a huge amount. For now I will submit the manuscript as a single paper but, if editor and reviewers think that two papers, one for launching and one for entraining, would be better, then I'd be willing to go along with that.

The methodology is described in considerable detail, however the analysis plan isn't

sufficiently described to fully restrict researcher degrees of freedom. For example:
i. What data quality checks will there be, if any? For example, if a participant gives a high rating to both “The initially moving rectangle made the other rectangle move by bumping into it” and “The initially moving rectangle passed across the other rectangle, which moved little or not at all” what will happen to the trial and/or participant data?

There is no clear principle by which to exclude individual participant data from analysis. If a participant consistently gave maximal ratings on both measures in a given experiment then that might be grounds for inferring that they had not understood or engaged with the task; on the other hand, excluding such responses would not affect statistical comparisons between means. This is one of the reasons for the choice of a large sample size (50): a small proportion of participants who provide what might be regarded as poor quality data should not have too great an effect on the analysis. One can't remove participants just because one doesn't like the look of their data, so all will be included. I have added a sentence stating this at the end of the new section "Minimum effect size and sample size determination", p. 15. Participants will be excluded if there are incomplete data (e.g. because of equipment failure or experimenter error); if that happens they will be replaced so the final n will still be 50.

ii. Apologies if I have missed this, but I couldn't figure out what exactly the will DV be. Will it be mean rating? Table 1 suggests that there will be 1 ANOVA per study, however when there are ratings on multiple statements surely there will be multiple DVs and therefore multiple ANOVAs? What correction for multiple tests will be conducted? Alternatively, if the ratings are to be combined into one DV how will this be done?

The DVs are the ratings on the statements. These vary between experiments, but are described in the method sections of each experiment. There is also a link to the full set of instructions to participants (which includes the statements to be rated) in the method section (p. 23). There will indeed be multiple DVs and therefore multiple ANOVAs. The significance level for each will be determined by the Bonferroni correction based on the number of DVs for that experiment. I have added a sentence making this clear in the design section on p.18. I have amended what is not Table 3 (formerly Table 1) to make it clear that there will be a separate ANOVA for each statement (in most experiments - the exceptions are experiments 8 and 10, as described in Table 3.

iii. Following from the above, might it be more straightforward in terms of analysis to have participants give a categorial report on their impression (“what was your impression, A, B or C?”) and follow that up with a continuous intensity rating (“please rate the intensity of your impression from 1-10”)?

I have many years of experience running experiments on perceptual impressions of causality using separate analyses for different statements. Changing to a method with which I have no experience feels risky and I would not want to do it without extensive piloting to ascertain what problems there might be. The word "intensity" is not necessarily the best word to use and that alone would need careful thought and piloting. Also I am not sure how I could analyse the data if the statement chosen for the intensity rating differed between stimuli both within and between participants.

iv. Will any sphericity corrections be used?

Yes. I have added a sentence to the design section on p. 18 to say this.

v. Given the number of tests that will be conducted, and therefore the high type 1 error rate, how will unexpected interactions (e.g. between speed and width in Experiment 1) be interpreted?

"Significant interactions are not predicted for these studies. If any occur, the existing literature will be utilised as a source of possible interpretations; any interpretation proposed would be treated as a way forward for further investigation." That passage has been added to the design section on pp. 18. Obviously one would have to be very cautious about interpreting an unpredicted one-off result, so suggesting a need for further research would perhaps the best option, unless there are clear implications for any of the extant theoretical accounts of the effects.

I also have some concerns about the proposed analyses as described in Table 2, however it is possible I have misunderstood. If I have misunderstood I apologise and could the analyses to be conducted please be clarified in text.

Experiment 1: this suggests the ratings on different statements will be directly compared – "Transition from high passing ratings at low width to high launching ratings at high width would be successful replication". Could you explain how "transition" will be tested for

This will be tested by direct comparisons between the two statements using related means t test. This has been added to what is now Table 4.

Experiment 2: "Significantly higher launching ratings for standard than for camouflage stimuli would be successful replication. All other results would be failure to replicate. Reported effect of fixation were not interpreted by Michotte; interpretation here will depend on results" Could these possible interpretations please be given.

Interpretation would be conjectural and I would have to be very cautious: as with unpredicted interactions, it would be written basically as a suggestion for further research to investigate. The fixation point is essentially a manipulation of the focus of attention, so things that are attended are processed in more depth than things that are not attended. Maintaining fixation on the point where contact occurs means that little attention is given to the rest of the stimulus. That should mean that the camouflage part of the stimulus, which occurs away from the fixation point, should have reduced or even no effect. Therefore it can be predicted that the camouflage will be more effective in the no-fixation condition. There is of course a substantial literature on effects of attention in visual processing and I would have to study it in more depth if the results go that way. Brian Scholl has done some important and relevant work so I would start with that and follow the literature as much as is necessary. Michotte did not say anything about why a fixation point should make a difference which is why it was not discussed in the manuscript, but I have added a little to the text to give the general idea. I have added a short passage on

this to the method section on p. 31. I would be happy to go into more depth on the attention research literature if the reviewer wishes.

Also, the text here and in Table 1 suggests to me that the Experiment 2 analyses involve running a 1 way Fixation (yes, no) ANOVA and another 1 way Stimulus (camouflage, standard) ANOVA for each of the 5 camouflage stimuli, i.e. 10 ANOVAS (potentially x 2 for each statement reported on). Is that correct? If so, why ANOVA and not t-tests? To avoid running 10-20 ANOVAs, a linear mixed model that permits different means for the 5 different stimuli would probably be more appropriate. Something like Rating ~ Fixation + Stimulus + (1 | participant ID) + (1 | stimulus ID).

Yes, that is correct. And of course I could use t tests (independent means for the fixation manipulation and repeated measures for the comparison with the launching stimulus) - this is pretty much the same as one-way ANOVA and I stipulated ANOVA only for the sake of consistency with other analyses. I would be happy to specify t tests instead if the reviewer prefers. The linear mixed model is not appropriate because each stimulus has to be analysed separately, the reason being that they have major qualitative differences, so direct comparisons between them are not meaningful. In effect, experiment 2 is five separate sub-experiments each based on one of Michotte's experiments. And the repeated measures comparison can only be for the no-fixation condition because the standard launching stimulus will be run with no fixation point. I had overlooked this and have now made it explicit in Table 3 and the design plan.

Experiment 3: “Significant effect of size of either object on launching ratings would discnfirm Michotte's claim. Non-Significant effect would be consistent with it” If the null is being predicted then Bayes Factors would be more appropriate so that you can make inferences about H0

I know very little about Bayesian analysis but fortunately one of my colleagues is an expert on it. He has advised me that the analysis would be a Bayesian linear mixed model using the Bayesian ANOVA module in JASP. I hope this is sufficient for present purposes. I have changed Table 3 and the design plan accordingly.

Experiments 4, 5, 6, 13, 14: I’m not sure what the advantage of using Tukey posthocs is here. On my understanding it sounds like the appropriate test here is a single contrast testing for a linear trend

O.K., this can be done. The design plan has been modified accordingly.

Experiments 7, 8, 9, 10: An ANOVA won’t be able to test this hypothesis. Multiple one-sample t-tests against 5 for each condition may be more appropriate

For experiments 7 and 9 I'm not sure why ANOVA doesn't test the respective hypotheses. However I have changed the hypothesis to a weaker directional hypothesis that is tested by the main effect of fixation. On reflection, for experiments 8 and 10 it makes more sense to ask for choice between the statements rather than weighting of them. Then it can be predicted that, in experiment 8, the launching statement will be the least chosen and, in

experiment 10, that the entraining statement will be the most chosen, for every stimulus. Each stimulus can be analysed separately. I think the chi-square test would suffice for this. The manuscript (including Table 3 and the design plan) has been amended accordingly.

Experiments 11, 12: The ANOVA seems to be testing for effects of motion and speed (ie those are the factors), but the hypotheses pertain to differences between the scale ratings (launching vs pulling etc)

The reviewer is right. The design is the ANOVA as described but the hypotheses would be tested on comparisons between the statements. This will be done separately for each stimulus using one-way ANOVA with repeated measures. I have amended Table 3 and the design plan accordingly.

Other comments

i. Will each experiment recruit a new (and non-overlapping) sample or will some participants take part in multiple studies? I think it's important that participants only take part in 1.

This is not possible because of resource limitations. Fourteen experiments with 50 Ps each adds up to 700 Ps. Therefore some participants will take part in multiple studies. Order of experiments will be randomised for each participant. In my experience people report what they see for each individual stimulus and as far as I can tell they don't think about others they have seen before. I appreciate that this is anecdotal only. However, with a within design, even running all stimuli for a single experiment on each participant could be regarded as problematic (and randomising order of stimuli for each P is the way to deal with that). Multiple experiments is not any more problematic than that.

ii. P4-5, L 95-102: "The perceptual nature of the launching effect is shown by evidence that it 96 can influence other contemporaneous perceptual processing. […] Detection occurred sooner for launching stimuli than for non-causal controls, supporting the 102 hypothesis that causality is constructed at an early stage of perceptual interpretation."
I don't think this is correct – differences in breakthrough time doesn't mean the effect is perceptual. Participants must decide when to report a percept as present versus absent, and in that sense variations in breakthrough times may well be a function of differences in decision thresholds.

The reviewer's point seems fair. Moors et al. didn't discuss decision thresholds - perhaps they should have done. But the question should be asked, why would decision thresholds vary between the different kinds of stimuli they used? For example, in their experiment 2, a standard launching stimulus was compared with a stimulus that was similar except that the first moving object stopped at a location where it couldn't be the cause of the second object moving, and a difference in breakthrough time was found between these stimuli. Given that the stimuli were similar in every respect except for the stopping location of the first moving object, it would not be easy to explain why decision thresholds would differ between them. The present manuscript is not the place for a detailed evaluation of this issue but I have added a footnote acknowledging the reviewer's point.

iii. P11, L275-277: “It is, however, important to the replication study that participants should, as far as possible, report perceptual impressions and not products of post-perceptual processing”
I think the point being made here is clearer later in the manuscript, but on my first reading it sounded to me as though participants were being asked the impossible – to report on some pre-decisional state of perceptual processing. Could this be reworded to make clear the point that particpants are being asked to report what they saw and not what they think following conscious deliberation?

I use the term "perceptual impression" because the causal impression is not exactly "seen", at least not in the way that an object and its features are. But the word "see" is used, albeit with qualification, in the instructions because there is no alternative that would be unambiguous for a naive participant. Also I am reluctant to use the term "conscious deliberation" because I don't know what it means. If the reviewer would be willing to accept exclusion of that tricky word "conscious", then I am happy to agree to the suggested re-wording and I have amended the manuscript accordingly.

iv. Power analysis: In my view it’s important to conduct a power analysis for each study. Most/all of the hypotheses can be tested with a t-test, so hopefully the calculations should be far more straightforward than powering for the multifactorial ANOVAs.

I included power analysis for each experiment in the rationale column of the design plan so I'm not sure what needs to be added. I have probably misunderstood the reviewer here but t test can only be used when there is one IV with two values and all of the experimental designs have more than that. If the reviewer would like to clarify what I have got wrong I would be happy to make the appropriate changes.

v. Table 2 (Rationale column): why does the smallest effect of interest change in each study? Is this the effect size that would be powered for given n = 50?

This has all been changed now - please see the section "Minimum effect size and sample size determination" starting on p. 18. The design plan (specifically the rationale column) has been changed in accordance with what is in that section.

Reviewed by anonymous reviewer, 05 Jun 2023 10:58

Thanks to the reviewer for the positive comments - much appreciated.

In this paper, the author illustrates a series of experiments aimed at replicating (and extending) the ‘classic’ work conducted by Michotte on causality. As reported by the author, despite the great importance and impact of Michotte’s work, a systematic investigation of his experiments with modern and fully reproducible tools is lacking. The main aim of this registered report is to provide a wide range of experiments to fill this gap.

The paper is surely well written and organised; honestly, I have little to comment on at this stage. Several comments are reported below.

The introduction did an excellent job of revising and illustrating the main characteristics of Michotte's work. However, I have found it to be quite dense, so I am wondering if some additional subparagraphs could help the reader.

As can be seen, the other reviewer requested the opposite of this, that the introduction be more concise. I can't please both reviewers on this and I am reluctant to lengthen what is already a very substantial paper. However, if the reviewer could give some direction on where extra material would help, I would be happy to put it in.

The second and more relevant comment is about the sample size. I fully understand that power analyses may be problematic, but at the same time, I also feel that determining sample size on the basis of an arbitrary decision, albeit motivated, can be challenging. So, I am wondering if there is still a possibility to see the power analyses reported in this work. I am aware this request implies great effort, so I'd be happy to adjust to what the editor (or other reviewers) thinks too.

I hope that the new section "Minimum effect size and sample size determination" starting on p. 18 addresses this point, and the design plan has been modified in accordance with what is in that section. If this isn't what the reviewer wanted I would be happy to make further changes.