Reply to PCIRR decision letter reviews #496: Norton et al. (2007) replication and extensions

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response as well as a tally of all the changes that were made in the manuscript. For an easier overview of all the changes made, we also provide a summary of changes.

Please note that the editor's and reviewers' comments are in bold with our reply underneath in normal script.

A track-changes comparison of the previous submission and the revised submission can be found on: https://draftable.com/compare/EulXCCinzRtH

A track-changes manuscript is provided with the file: PCIRR-S1-RNR-Norton-etal-2007-rep-ext-main-manuscript-track-changes.docx (https://osf.io/r4jgz)

Summary of changes

Below we provide a table with a summary of the main changes to the manuscript and our response to the editor and reviewers:

Section	Actions taken in the current manuscript
General	R1: ● Reviewer comment 2- Correlational hypothesis H4-2 removed from PCIRR-Study Design Table. ● Reviewer comment 8- Amendment of wording for extension hypotheses in PCIRR-Study Design Table and in 'Extensions' and 'Results' sections. R2: ● Reviewer comments 1-3- Amendment of language and spelling errors in abstract ● Reviewer comment 5- Amendment of formatting for in text references to previously established effects throughout manuscript ● Reviewer comment 6- Amendment of language used to describe findings from previous studies and current replication throughout manuscript, to provide uniform inference of causality.

Section	Actions taken in the current manuscript
Introduction	R1: ● Reviewer comment 10- Additional information regarding conceptual replication of Study 3 and omission of Study 5 provided (p.10, see footnote [2]). R2: ● Reviewer comment 4- Introduction of *less is more* effect revised to distinguish between 'strangers' and 'acquaintances', with amended subsequent uses of term 'stranger' . ● Reviewer comment 7- Amendment of wording when referring to findings by Ullrich et al. (2013). ● Reviewer comment 8- Additional justification provided for conceptual replication of Study 3.
Methods	R1: ● Reviewer comment 6- Revision of additional information provided regarding use of pilot data.
Results	R1: ● Reviewer comment 2- Correlational analysis for Study 4 moved to supplementary materials.
Reporting	R2: ● Reviewer comment 6- Amendment of language used in reporting findings throughout manuscript to establish uniform inference of causality.
Supplementary materials	R1: ● Reviewer comment 2- Correlational analysis for Study 4 moved to supplementary materials. R2: ● Reviewer comment 9- Analysis code updated

Note. Ed = Editor, R1/R2 = Reviewer 1/2

[We note that we are not familiar with the titles and ranks of the reviewers, and looking for that information proves tricky. To try and err on the side of caution, we refer to all reviewers with the rank Dr./Prof. . We apologize for any possible misalignments and are happy to amend that in future correspondence.]

Reply to Editor: Dr./Prof. Yuki Yamada

> **First, I apologize that it has taken somewhat longer to collect the peer review reports. The reviewers all submitted their reports very quickly after accepting our request. What took time was the rest of the process, for which I bear the responsibility.**
>
> **Now, I have just received very helpful peer review comments from two experts. As you can see, both are very positive about this study. And at the same time, they focus on almost the same aspects: power analysis and adjustments for test multiplicity. Please see their specific comments, but I believe their points are in line with current standards of research practice. I encourage you to carefully consider them. The reviewers also made some really constructive comments on the wording of the text, so please take those into consideration as well.**
>
> **I very much look forward to receiving your revised manuscript!**

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit.

Reply to Reviewer #1: Dr./Prof. Zoltan Kekecs

> **The manuscript describes the protocol for a replication of Norton et al. (2007)'s lure of ambiguity effect. The registered report is thorough and shows not only the protocol but also the results for a simulated scenario. The replication attempt makes a reasonable effort at testing the replicability of the critical results of Norton et al. (2007), and includes extensions to the original research questions that help further evaluate the mechanisms underlying the effect. I really like that materials, data, and analysis code used to produce the manuscript are made openly available by the authors, enabling a thorough evaluation of the work. All in all I think this is going to be a valuable project that has a good chance to replicate the effects if they exist and that can provide deeper insight into the influencing factors and mechanisms at play. Below I list a number of suggestions that may help the authors improve the manuscript and the protocol.**

Thank you for your kind comments and helpful suggestions.

> **1. It seems to me that all participants will do all experiments. However, the subsequent experiments might influence each-other. For example a person who first claimed that they think more traits lead to more liking might respond accordingly in the second study to make their responses more consistent. It seems sensible to me to at least separate Study1 and the rest of the studies to prevent such effects.**

Response:

We have successfully implemented this design many times in our replications, also with PCIRR submissions/publications. In all those we were able to convincingly demonstrate why this approach was beneficial and important for advancing the literature and our understanding of the phenomenon.

We consider this design to have major advantages, building on but going beyond the target article's design. One of the things that this design would help us to specifically test would be whether there would be carry-on effects and the impact of order combining several paradigms.

A unified study design embeds the original's separate studies, for the first study displayed to participants (like you pointed out), but goes beyond that in allowing for additional insights by performing additional exploratory analyses either only examining the first displayed study (which would mirror the original's) or with order as a moderator of the different effects.

In addition, most importantly, this helps address concerns regarding the sample and attentiveness. When we ran replications of studies from the same article separately and then some of those failed whereas some of those were successful, then reviewers often raised concerns that the failed experiments were due to sample/time/context, and then asked us to again repeat the failed replication (reflecting hindsight/outcome biases), and once we did these ended up resulting in similar findings to the original run. Yet with a single unified design, that concern is fully addressed, with the much more likely explanation that the failed replications are because of the differences between the studies, not because of the context or the sample.

Furthermore, we are able to run additional exploratory analyses linking the two studies to examine consistency in responding and gain a better understanding as to whether the two studies truly seem to tap into the same phenomenon, atleast from the perspective of participants' responding.

There are many examples, but we will give one recent example that just completed a PCIRR Stage 2:

> Petrov, N., Chan, Y., Lau, C., Kwok, T., Chow, L., Lo, W., Song, W., & Feldman, G. (2023). Comparing time versus money in sunk cost effects: Replication Registered Report of Soman (2001). *International Review of Social Psychology*, 36(1): 17, 1–18. DOI: 10.5334/irsp.883 [PCIRR Stage 2 recommendation/Open peer review] [PCIRR Stage 1 recommendation/Open peer review] [Article] [Preprint] [Open materials/data/code]

In this project, we conducted direct replications of Studies 1 and 2 and a conceptual replication of Study 5 in the article Soman (2001) claiming that money sunk costs are larger than time sunk costs. We used a similar unified design running the three in a single unified data collection, with the order of Studies 1 and 2 randomized. The final result was that Study 1 was successfully replicated, whereas in Study 2 there were sunk cost effects for both time and money, yet no differences between the two, which we summarized as a failed replication. Therefore, this was not an issue of lacking power, but rather the detection of effects even when none were expected (time sunk costs). We conducted order effect analyses and analyzed the data from the studies in which the study was displayed first, and across all these analyses the results were very similar and consistent.

In the past when we ran these separately, editors and reviewers would ask us to rerun the failed replication, with various post-hoc claims regarding the reason having to do with the sample or time/context. However, in the case of the Soman replication, the unified data collection clearly showed that the sample was attentive and careful, with one successful replication, which means that the failed replication was not due to issues with the sample or context/time. In addition, combining the two allowed us to get better power for the less money invested, and additional

analyses can be run to further identify participants who do not answer consistently across the two different scenarios in the two different studies. Additionally, it shows that order did not impact these studies.

We ran many replications with this design and across all the replications that implemented this approach we have yet to see any order effects, yet have been able to gain important insights regarding the phenomenon.

Additional recent examples with a unified design and diverging findings between studies:

Chandrashekar, S., Adelina, N., Zeng, S., Chiu, Y., Leung, Y., Henne, P., Cheng, B., & Feldman, G. (2023). Defaults versus framing: Revisiting Default Effect and Framing Effect with replications and extensions of Johnson and Goldstein (2003) and Johnson, Bellman, and Lohse (2002). *Meta Psychology*, 7. DOI: 10.15626/MP.2022.3108 [Article] [Preprint] [Open materials/data/code] [Open peer review]

Yeung, S. & Feldman, G. (2022). Revisiting the Temporal Pattern of Regret: Replication of Gilovich and Medvec (1994) with extensions examining responsibility. *Collabra:Psychology*, 8 (1): 37122. DOI: 10.1525/collabra.37122 [Article] [Preprint] [Open materials/data/code]

Vonasch, A., Hung, W., Leung, W., Nguyen, A., Chan, S., Cheng, B., & Feldman, G. (2023). "Less is better" in separate evaluations versus "More is better" in joint evaluations: Mostly successful close replication and extension of Hsee (1998). *Collabra:Psychology*, 9 (1), 77859. DOI: 10.1525/collabra.77859 [Article] [Preprint] [Open materials/data/code]

In our first submission we attempted to clarify this choice by noting the following in the manuscript:

> "Combining several studies from a single target article in a single data collection has previously been successfully tested in several replications and extensions conducted by our team (e.g., Chen et al., 2023; Petrov et al., 2023; Vonasch et al., 2023; Yeung & Feldman, 2022; Zhu & Feldman, 2023), and is especially powerful in addressing concerns about the target sample (naivety, attentiveness, etc.) when some studies replicate successful whereas others do not, as well as in the potential in drawing inferences about the links between the different studies and consistency in participants' responding to similar decision-making paradigms.".

We agree that this is a deviation from the target's and adds some complexity, and yet we believe that the risk reduction (in interpretation) and value added in exploratory insights is well worth

additional complexity, if there would indeed be any. You will see in the Soman (2001) replication cited above that this was a straightforward analysis to address and share with the readers.

Action: We added the following subsection to the Method->Data analysis strategy -

> **Order effects**
>
> One deviation from the target article is that all participants completed all scenarios in random order. We considered this to be a stronger design with many advantages, yet one disadvantage is that answers to one scenario may bias participants' answers to the following scenarios.
>
> We therefore pre-register that if we fail to find support for our hypotheses that we rerun exploratory analyses for the failed study by focusing on the participants that completed that study first, and examine order as a moderator. To compensate for multiple comparisons and increased likelihood of capitalizing on chance, we will set the alpha for the additional analyses to a stricter .005. Our planned sample size (1383) is large enough to provide sufficient statistical power to conduct moderation analyses and examine the order effects, if needed.
>
> [TBD conclusion based on our experience with a unified design so far: We found [no] differences in conclusions]

We also added the following as a planned discussion in Stage 2 in the discussion section:

> [Planned discussion in Stage 2: Discuss the unified data collection and possible order effects]

(Please note: This reply has been used in some version in other replies to PCIRR feedback, such as in our recent PCIRR IPA replication of Tsang, 2006:

> Chan, C., *Lim, H., *Lau, F., *Ip, W., *Lui, C., Tam, K., & Feldman, G. (2024 expected). Revisiting the Effects of Helper Intention on Gratitude and Indebtedness: Replication and extensions Registered Report of Tsang (2006). DOI: 10.17605/OSF.IO/GHFY4 [In-principle acceptance/Open peer review] [Preprint] [Open materials/data/code])

> **2. "However, we found the choice of analytic strategies somewhat arbitrary; to directly test the effect of the quasi-experimental condition on liking, it is sensible to conduct a t-test rather than computing the correlation. Thus, while we aimed to replicate the correlation, we also planned to test the relationship with a t-test to see whether the quasi-experimental condition influenced liking."**
> **– There is no point in replicating an inappropriate analysis, especially that this is already a conceptual replication. You should use the best analysis method available to answer the research question.**

Response: We appreciate you sharing your views on this.

This is a common dilemma we face when replicating older articles. We first aimed to replicate the target's original analyses (the correlation) so that readers can compare the target and the replication using the target's own terms and analyses, which may have been considered appropriate at the time, atleast to the authors/editor/reviewers of that manuscript.

Given your feedback, we decided to move the correlation analysis to the supplementary materials. That way the emphasis is on the best analysis, yet readers can still compare the results should they want to.

Action: Correlation analysis for Study 4 moved from main manuscript to supplementary materials, t-test analysis retained in main manuscript. Hypothesis H4-2 is thus removed, and H4-3 is renamed to H4-2.

> **3. I really like the fact that the analysis codes and power analyses codes are available.**

Response: Thank you for the support and for reviewing those. This just makes more sense, especially with Registered Reports.

> **4. I don't see why there is not sample size calculation (power analysis) for H2-2 and for H4-3. Instead of effect sizes provided in the original study (not available in this case), you can use smallest effect size of interest, or some other effect size estimation method to gain the required numbers. Simulation can also help. You can do a simulation-based power analysis (of course that would also require setting effect sizes and variances for the simulation).**

Response: Our main aim was the replication, and so to have sufficient power to replicate Norton et al's (2007) key arguments with much weaker effects. The corresponding hypotheses were not conducted/possible in the original study, and were meant as extensions. In our power analysis we

therefore focused on the target's effects and determined our target sample size using the Simonsohn's (2015) small telescopes recommendation. We indicated that with our target sample size of $N = 720$ (after planned exclusions), we can detect a small effect size of $d = 0.25$ ($r = .12$), typically considered weak to medium effects in social psychology research (Jané et al., 2024), and about half of the effect sizes reported in the target article.

We note that aiming to determine SESOI estimates for extensions is a somewhat subjective process, far trickier to determine for new hypotheses, and runs into the challenge of trying to balance power and resource/budget constraints. We much prefer a solid quantifiable approach with needed adjustments, as the small telescopes approach. Therefore, our priority is the replication. The findings from this investigation using the already large sample well-powered for the replication would allow us an initial estimate of the effect for the extensions.

Action: No action taken

> **5. In the power analysis for H3 the authors write: "Since the paper does not offer information about standard deviations, we assumed they were 1 and conducted the analysis." This seems arbitrary. I am not a domain expert so I do not know whether this is a reasonable assumption. This should be supported somehow, for example data from another study, or data from a pilot study. Alternatively, a range of reasonable SDs could be tried in this analysis and the range of estimated samples sizes could be reported in the paper.**

Response: Thank you for raising this, we understand the concern.

We believe that is a reasonable assumption, yet if there are more reasonable/plausible estimations of SDs, we would be happy to adjust our power analysis given clear editorial guidelines.

Following your suggestion, we conducted the same power analysis with SDs = 0.1 and 5, and found that a sample of 410 is sufficient to achieve 90% power. We should be sufficiently powered even with other assumptions.

Action: No action taken

> **6. "…the data of the 30 participants will not analyzed other than to assess survey completion duration, feedback regarding possible technical issues and payment, and needed pay adjustments. Unless in the case of serious technical issues that affect data quality and require survey modification, these participants will be included in the overall analyses" These two statements seem to be contradictory. Please reconcile.**

Response: Thank you, we understand how what we wrote could be confusing, and we appreciate the opportunity to clarify this.

What we meant was that we did not intend to test our hypotheses on the responses of these 30 participants prior to the completion of the full data collection, beyond the checks outlined above - payment and technical issues, to ensure no issues and sufficient data quality. We intended to only run our planned full analysis on the full data.

Action: We adjusted the wording of the 'Participants and design' section in the main manuscript to make this clearer:

> [We will first pretest the survey duration and technical feedback with 30 participants to make sure our time run estimate was accurate and adjusted pay as needed. The data of these 30 participants will not be analyzed to test the outlined hypotheses in this paper prior to full data collection, other than to assess survey completion duration, feedback regarding possible technical issues, and needed pay adjustments. Unless in the case of serious technical issues that affect data quality and require survey modification, these participants will be included in the overall analyses conducted with the full sample.]
>
> [An example placeholder, to be updated in Stage 2: We first pretested survey duration with 30 participants to test time run estimate and adjusted pay based on the duration. The data of the 30 participants was not analyzed other than to assess technical issues, survey completion duration, and needed pay adjustments, and were included in the final data analysis.]

> **7. The Participants and design section indicates that 1383 participants were included in the data analysis. This is much higher than the target sample size. Why is this the case? If you exceed the target sample size, it may seem as optional stopping, collecting data until you get the desired results. I suggest that for the confirmatory analyses you only take into account the first X responses, X being the target sample size. If you want you can repeat the analyses on the full sample as a robustness check/sensitivity analysis. Also, the exploratory analyses can be conducted on the full sample.**

Response: The 1383 participants are simulated random data created by Qualtrics. It is only meant for demonstration purposes, nothing more. We had 1383 data points because we used Qualtrics for random response generation, and Qualtrics sometimes glitched and timed-out, and we just decided to include the full generated dataset.

We aim for 800 participants, yet we note that with labor markets it sometimes goes a bit beyond or under that (e.g., someone completed the survey but timed out before submitting the HIT, and therefore not included in the general count but receives compensation using a compensation HIT afterwards). We see no reason to exclude valid responses, as long as we are transparent about everything in the process, and therefore intend to report all valid collected data points.

We see no reason to worry about or suspect optional stopping, first because this is a Registered Report with an in-principle acceptance regardless of outcomes, and so there is no incentive or reason for us to do anything that would affect the outcomes in any direction. Also, open-science practices address that - we will make all of the data (without workerIDs, IPs, or locations) available in our dataset, including date/time-stamps, allowing everyone to assess the data collection.

Action: No action is taken.

> **8. You have marked H5 through H9 as exploratory "hypotheses". Something is either an exploratory analysis or a confirmatory hypothesis test. In the "Extensions" section you use inferential statistics and confirmatory hypothesis testing language for these analyses. Please, decide whether these are exploratory analyses or confirmatory hypothesis tests. If exploratory analyses, do not use p-vlaues, or testing language. Just focus on the descriptive results, effect size estimates, dispersion statistics, and visualization of these. If they are confirmatory tests, do not mark them as exploratory, and provide sample size calculation (power analyses) for these as well.**

Response: We appreciate the note and feedback.

The label ‘extension’ hypotheses is a better label than ‘exploratory’ hypotheses and what we used in the hypotheses tables in the introduction. We realized that was not well-aligned with the PCIRR design table.

Action: We amended the study design table to align with the hypotheses table, adjusting ‘Exploratory’ to ‘Extension’ for H5 through H9.

> **9. The power analysis provides sample size targets to reach at least 90% power for each replication hypothesis tests individually. This seems to assume that you will have at least 90% power to detect the effects in this study. However, this is incorrect, since you are testing multiple hypotheses. For example if you have two hypotheses and have a 90% power to detect each effect, you only have 81% probability to detect both effects, and thus, 19% probability to miss at least one of them. With more hypotheses, this missing effect chance can stack up quickly. You could either power your study to have a 90% probability to detect ALL effects, or be explicit about the probability of missing a number of effects in the Power and sensitivity analysis section.**

Response: We appreciate this feedback. In our many other replications, also with PCIRR, we did not face this comment, nor did we find this to be the case for our findings for those replications. We rarely saw such adjustments to alpha in testing multiple hypotheses in publications, unless there is some link or dependency between the tests or these are included in the same analysis. Yet, that is ofcourse not an indication that this practice is correct.

We therefore consulted on this point with the community on : https://twitter.com/giladfeldman/status/1700408453335028072?s=20 . As you can see, there are lots of mixed opinions, with a somewhat heated debate, and even experts and editors indicate different at times conflicting views on this issue and the needed remedy. We can generally see that the common practice is what we have implemented so far in our other replications and in this manuscript, especially given that the hypotheses and their tests are considered separate and are not dependent on one another. We follow Rubin (2021) to adjust p-values whenever we use multiple tests to test a single hypothesis.

We also think that in the case of Registered Reports, with everything open and transparent, power analyses aiming for 95% power, and with a focus on effect size comparisons, this becomes less of a concern (Schimmack, 2012).

Action: No action taken

Note: We would be happy to revise given clear editorial guidelines and instructions on what to amend. If the reviewer or editor feel that an adjustment in sample target is needed - then we ask

that you please provide us with relevant citations and an example or two of other Registered Reports (preferably PCIRR, preferably replications) that has done something similar, and taking into consideration cost/benefit of going beyond the already large planned sample of 800.

References:

- Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. Synthese 2021, 1–32. https://doi.org/10.1007/S11229-021-03276-4
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*(4), 551–566. https://doi.org/10.1037/a0029487

> **10. "We did not include Studies 3 and 5 as targets of direct replications as these involved experiments using real online dating platforms" – do you mean 4 instead of 5? study 5 was not mentioned until this point.**

Response: Thank you, we tried to clarify this point.

We do not directly replicate Study 3 by Norton et al. (2007), but instead conduct a conceptual replication of Study 3 within the replication of Study 2. We agree that it is relevant for us to elaborate a bit more about the omission of Study 5.

Action: We added the following footnote to address Study 5:

> We chose not to include Study 5 by Norton et al. (2007) in our replication. Study 5 explored whether findings from Studies 1-4 would be replicated in real-life settings, using individuals from a real-world dating platform who had either recently been on a first date, or were going on a first date in the near future. Findings replicated effects reported in Studies 1-4. We felt that we should first focus on Studies 1-4 and once we are able to establish the replicability and robustness of those findings, inspire the more costly and ambitious follow-up of Study 5 in real-life setting.

We also added a planned discussion for Stage 2:

> [Planned discussion in Stage 2: Our focus on Studies 1-4 and implications for follow-up research addressing real life settings, such as a replication of Study 5.]

Reply to Reviewer #2: Dr./Prof. Philipp Schoenegger

> **The paper under review provides a commendable effort in directly replicating Norton et al. (2007). The authors have done an excellent job in motivating the study and setting up the project. The manuscript is not only well-structured but also remarkably transparent, with a plethora of resources and data made available at the OSF. The research question and hypotheses are clearly articulated, and the methods employed are appropriate. Furthermore, the manuscript is characterized by a high level of detail, making it easy for the reader to follow the research process and understand the nuances of the study.**

Thank you for the positive opening note and the helpful and constructive feedback

> **However, I have a number of points that I would like to see the authors address before running the study. Some are minor while others are major, but I believe that this study is very much worth running and would be a great addition to the scientific record. The former are suggestions that may be disregarded with some reasonable explanation, but the latter should be followed or at least be rejected with a detailed argumentation as to the reason for doing so.**

> **1) For the abstract, the use of 'results' seems too strong to me, especially given the fact that you use much more associational language throughout the paper. I would suggest you use more uniform language to avoid misunderstandings.**

Response: We agree that this causal language does not reflect the associational nature of the analyses conducted. We have amended this in line with language used by Norton et al. (2007) in their original manuscript.

Action: Abstract:

> "Norton et al. (2007) introduced the less is more phenomenon wherein knowing more information about others results in less liking."

amended to

> "Norton et al. (2007) introduced the less is more phenomenon wherein knowing more information about others is associated with less liking."

> **2) Additionally, the term 'less liking' should be set-up better in the abstract, a less informed reader may not be able to follow.**
>
> **3) There is also a small typo in the abstract, it should be 'Overall, we found'. I generally suggest reworking the abstract for clarity.**

Response: Thank you for the feedback. We adjusted and amended the typo.

Action: Abstract- Amendment of typo in abstract to read 'Overall, we found', and the intro now reads:

> Norton et al. (2007) demonstrated a counterintuitive phenomenon that knowing other people better and/or having more information about them is associated with decreased likings. They summarized it as - ambiguity leads to liking, whereas familiarity can breed contempt.

> **4) In the introduction, I would improve the set up early on somewhat better motivate the term 'stranger'. It seems to me that one may not be able to meet the same stranger regularly (without that person not becoming a stranger anymore). At least this may be the case with the standard definition.**

Response: We appreciate the opportunity to improve our introduction of these concepts and made revisions. We believe our amendments helped strengthen the manuscript. For greater clarity, we now utilize the term 'acquaintance' in lieu of 'stranger' when referring to subsequent interactions with an individual. We retain the original 'stranger' term for initial encounters only.

Where information in the manuscript may be relevant in both initial encounters and subsequent interactions, we use 'individual', to avoid implications of reference to only one of these contexts.

Action: Page 1, Introduction- Now distinguishes between strangers and a newly included term, 'acquaintances' in main text and footnote [1]. Wording throughout the first paragraph is amended accordingly (tracked changes available in main manuscript).

Section 'Direct replications of Studies 1a, 1b, and 2'- amended word 'strangers' to 'individuals' in last sentence of second paragraph.

> **5) Additionally, it is worth keeping an eye on consistent writing of the 'less is more' effect. This can be with ' or cursive, etc. Just keep it consistent.**

Response: Thank you. We amended the formatting of all mentions of this effect throughout the manuscript to italicize the effect name (i.e., the less is more effect).

We have also amended formatting of the reference to the lure of ambiguity effect and person positivity bias in our manuscript, in accordance with this suggestion.

Action: Amendments of formatting for references to the less is more effect (throughout manuscript, tracked changes available in main manuscript file), the lure of ambiguity effect, and person positivity bias.

> **6) In the 'Target for Replication' section, I would again point out that there is a stark difference in the presentation of the results, particularly in the language used to describe the findings. The manuscript alternates between associational language, such as 'tended to report,' and causal language, represented by terms like 'results.' This inconsistency could lead to confusion regarding the level of causality that the study aims to establish. I suggest adopting a more cautious and consistent approach to causal language. Specifically, it would be prudent to decide on a uniform level of causality that the study aims to establish and maintain this consistently throughout the manuscript. This is particularly important because some of the studies that you are replicating or referring to are associational in nature. Using inconsistent or overly strong causal language could risk misleading interpretations and should therefore be avoided.**

Response: In our revision, we amended language that may indirectly suggest causality in the discussion of findings of previous research and our own planned analyses. We welcome any further suggestions for amending this within our manuscript.

Action: Amendments throughout manuscript to avoid causal language.

> **7) When you write that "Ullrich et al. (2013) also challenged Norton et al. (2007)," I would change it to a phrasing that suggests that a paper or finding is challenged, not a set of authors.**

Response: We appreciate this suggestion and make this amendment accordingly.

Action: Introduction- the sentence now reads

> "Ullrich et al. (2013) also challenged the findings by Norton et al. (2007)" .

> **8) Lastly, in the 'Conceptual replications of Study 3 and 4' section, I think your treatment of the conceptual replication of Study 3 could use more detail. I suggest you explain why you do not replicate Study 3 (I assume because of the specific sample, but simply having a different sample is not reason enough, the reason may be cost, temporal differences, effort, etc), and what potential differences this change in sample may bring with it with respect to interpreting any given results.**

Response: We have now added justifications for the conceptual replication of Study 3 in the main manuscript. We provide an overview of these justifications here for your reference.

Whilst the sample does deviate from the original Study 3 (which used the online dating platform to recruit participants), this study also required a second sample who generated a list of self-describing traits to be used in the main study. However, findings by Norton et al. (2007) suggest the use of self-generated traits did not have a significant influence on perceiving liking. Rather, the authors concluded that Study 3 replicated the effect found in Study 2, which used a more limited set of traits containing more negative attributes. We also felt the nature of this study was too similar to Study 3 to present both tasks in the same experiment. In light of these points, we opted to omit a direct replication of Study 3, and instead include the perceived similarity measure from this study in our replication of Study 2.

It is important to note that this means the number of traits presented to participants will vary slightly from that of the original Study 3, adopting the 2 vs. 4 vs. 6 vs. 8 trait conditions from Study 2, rather than any number of traits between 1 and 10 as used by Norton et al. (2007). Again, as findings for perceived liking were highly similar across both trait length variations in Studies 2 and 3, we do not anticipate that this will significantly impact our findings.

Action: Section 'Conceptual replications of Studies 3 and 4' now contain more detail justifying the conceptual replication of Study 3 (tracked changes available in main manuscript).

Below I outline my three bigger concerns:

9) In the methods section, while the inclusion of a power analysis is commendable and aligns with best practices in research methodology, there are several issues that need to be addressed. Firstly, the code provided for the power analysis is not immediately executable without minor modifications (at least for me). For instance, the line of code "esc_chisq(chisq = 112,67, totaln = 294, es.type = "r")" contains a syntax error; the comma should be replaced with a period to read "112.67" instead of "112,67." Although I was able to replicate your final sample size of 234 participants after making these adjustments, the code should be cleaned up to ensure that it runs seamlessly for other researchers who may wish to replicate or extend your work.

Response: Thank you very much for reproducing our analyses on your device and for pointing out the issue. It did not produce any errors when we ran the codes and it still does not. We believe this has something to do with language settings.

We note that we provide the knitted Rmarkdown output in HTML format together with the Rmarkdown code, so anyone aiming to assess reproducibility can see the code and outputs side by side. Minor cross platform/version/language issues are always tricky to address 100%, but in this case like the one you mention, it is fairly straightforward for anyone to understand the type of adjustments that need to be made.

Action: Amended code available on our OSF project page, replacing "," with ".".

> **10) Secondly, the choice of approach for determining effect sizes in the power analysis warrants discussion and perhaps revision. The manuscript seems to rely on expected effect sizes derived from a single paper, which is a methodological choice that could be problematic. Given that the very essence of replication studies like this one is to question the generalizability of such expected effect sizes, it would be more prudent to adopt a 'smallest effect size of interest' approach instead of using the observed effect size as the expected effect size. This would involve identifying the smallest effect that would still be of scientific interest and using that as the basis for the power analysis, rather than relying on potentially inflated or context-specific effect sizes from previous work. The 'smallest effect size of interest' could be anchored initially to the expected effect size but should be revised downwards to reflect a more conservative and scientifically rigorous estimate. This approach would align better with the overarching goals of replication studies, which aim to rigorously test the robustness and generalizability of previous findings. If the authors disagree with this suggestion, a detailed justification for the current approach would be beneficial, including why it is considered superior or more appropriate for this specific study. I am also aware that this is likely to reduce the effect size and thus increase the sample size needed, but for an important replication like this, these additional costs seem justified and indeed crucial.**

Response: We appreciate the feedback and suggestions.

Our planned sample size was determined in accordance with Simonsohn's (2015) small telescopes approach with the generalized rule of thumb of recruiting a sample size 2.5 times greater than the original study. Our planned sample size (N = 800) exceeds these calculations. Simohsohn's small telescope approach overcomes limitations of the smallest effect size of interest (SESOI) approach. Specifically, the two major issues associated with null hypothesis testing of the SESOI is that 1) the SESOI can only be arbitrary and 2) the SESOI often requires an unrealistic sample size (see Simonsohn, 2015, p. 560).

The Simonsohn's (2015) small telescope approach considers studies with less than 33% statistical power as being severely underpowered, as this indicates that the odds is 2:1 against obtaining a significant effect. Accordingly, it defines a small effect as the effect size that gives 33% statistical power to the original study. If a replication study finds an effect size that is statistically significantly smaller than the small effect, this suggests that the replication effect size is not large enough to have been detected with the original sample size. As such, the small-telescope approach is partly based on the SESOI approach as it uses the 33% cutoff as an arbitrary threshold. Simonsohn (2015) recommended 2.5 times the sample size of the original

study as this ensures replication studies have at least 80% statistical power to reject the hypothesis of the small effect (assuming the true effect is 0).

We also provided a detailed sensitivity analysis of the types of effects we are able to detect with the sample, showing that it will allow for a detection of what are considered to be weak to moderate effects, far weaker than those reported in the target article. Our sample size exceeds Simonsohn's (2015) recommendations with accounting for possible exclusions.

Reference:

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, 26(5), 559–569. https://doi.org/10.1177/0956797614567341

Action: No action taken

> **11) Lastly, I noticed that the 'Results' section (or any other place) does not include adjustments for multiple hypothesis testing. Given that this is a replication study with multiple hypotheses, it would be beneficial to consider some form of adjustment to maintain the integrity of the results. I checked the analysis code and, as far as I can tell, found no evidence of any such adjustments that may not have been mentioned in the paper, with any instances of 'p_adjust' set to 'none.' I would strongly suggest that the authors consider incorporating a method to correct for the multiple hypotheses being tested in this study.**
>
> **Section Results: Lastly, I am surprised not to see any adjustments for the multiple hypotheses that are tested in this replication. I went to the analysis code to ensure it wasn't just left out of the manuscript but I could find no adjustment either, with the few instances of p_adjust set to none. I would strongly suggest that the authors include some type of adjustment that corrects for the large number of hypotheses tested in this replication.**

Response: Thank you for the comment. We do adjust p-values whenever we use multiple tests to test a single hypothesis (e.g., H2-2), yet not for multiple hypotheses, as was discussed and answered above.

We would happily revisit this decision and make further adjustments if given clear detailed editorial guidelines as to what to adjust and how, and would then please ask for citations and links to examples of replications, preferably as Registered Reports or from PCIRR, that have previously made such adjustments.

We will also add a note about replications more broadly. Direct replications aim to follow the target article as closely as possible and simply rerun and re-test with a new sample. Even if some adjustments are needed but were not addressed by the target article, it is unclear whether the replicators should simply follow the target or improve on their analyses to tackle things they have not, often leading to a stricter criteria. This creates a very complex interpretation of the results of the replication compared to the target. In most cases we have seen, and in most of our replications so far, we try and run things as the target did to see if we can come to the same conclusion.