Reply to decision letter reviews: #373
Tsang (2006) replication and extensions

We would like to thank the editor and the reviewers for their useful suggestions. Below we provided a detailed response as well as a tally of all the changes that were made in the manuscript.

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/GBOJWonNvgKG

A track-changes manuscript is provided with the file: "PCIRR-RNR-Tsang2006-replication-extension-main-manuscript-track-changes.docx" (https://osf.io/wnu7z)

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	Ed: We corrected the typo mistakes. R2: We rechecked the in-text citations and ensured they are all included in our reference list. R3: We went through the manuscript again and rearranged the paragraphs to increase the readability
Introduction	Ed: We clarified the rationale of our extension. Ed: We updated the hypotheses table and highlighted the core hypotheses. R1: We added explanations to justify why we replicate Studies 2 and 3 instead of 1. We also modified the rationale and adopted the suggested literature of Peng et al. (2020) in our explanation of the extensions.
Results	Ed: We reported the version of R used in our research. R3: We rechecked and corrected the numbering of all tables and figures. Ed: We updated the results to clarify the analyses for the core hypotheses. We added interactions examining core hypotheses.
Supplementary materials	Ed: We added a section under "Additional information about the study" to describe the subject recruitment process.

Note. Ed = Editor, R1/R2/R3 = Reviewer 1/2/3

[Sidenote: 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. Zhang Chen

> **I have now received comments from three expert reviewers, including the author of the original paper. As you will see, reviewer 1 (Tsang) and 2 (Field) are generally very positive about the current replication and extension, while reviewer 3 (Peng) is a bit more critical of the literature review and the methodology. All reviewers provided helpful and constructive comments that can be used to further improve the manuscript. Based on the reviews and my own reading, I would therefore like to invite you to submit a revised version.**

Thank you for the reviews, feedback, and the invitation to revise and resubmit. The feedback was very valuable, and we appreciate your and the reviewers' time and support for our manuscript.

> **.1. Both reviewers 1 and 3 have concerns about combining study 2 and 3 within subjects. I think these are valid concerns - combining both studies will make the current investigation less of a 'close' replication, and potentially complicate the interpretations of results. Of course, by looking at the first study only by each participant, you will still be able to examine each study separately, but with only half of the original sample size. Since data will be collected online via Prolific, I think running study 2 and 3 as two separate studies will not increase the total monetary cost and the time needed for data collection. As such, I agree with reviewers 1 and 3 that running study 2 and 3 as two separate studies seems to be a better option.**

Thank you for raising this concern. We ask for your understanding to allow us to proceed with our original design and below we will try and explain why.

We have successfully implemented this design many times in our replications, also with PCIRR submissions, and this issue often comes back. In all those we were able to convince the recommender/editor to allow us to proceed and then 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 original's. 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 a bias), resulting in similar findings. 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). [PCIRR Stage 2 recommendation/Open peer review] [PCIRR Stage 1 recommendation/Open peer review] [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*. [Article] [Open materials/data/code]

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*. [Article] [Open materials/data/code]

To address your point we added the following in our “data analysis strategy”, to pre-register examining order effects in case we fail to find support for our hypotheses, with a compensation for alpha:

> 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 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 (without outlier exclusions). 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.

> [TBD conclusion based on our experience with a unified design so far: We found [no] differences in conclusions]

Finally, you pointed out two challenges. First is that this may add 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 is indeed any. You will see in the Soman (2001) replication cited above that this was a straightforward analysis to address and share with the readers.

The second was regarding the loss of power if only the first study displayed is analyzed or with an examination of an added order moderator. In the previous submission we already noted that we:

> “multiplied the estimated required sample of 264 by 2.5 to result in 660. Accounting for possible exclusions of 0-10% based on our previous experience with the target sample, our integrated design, and allowing for the potential of additional analyses, we aimed for a larger total sample of 800 participants, over four times larger than the combined samples in the target article.”

And provided a sensitivity analysis showing that we are very well powered to detect effects much weaker than in the target, even if we include the order moderator, which does little to affect the sensitivity analyses we conducted for the ANOVA. The sample is so large in comparison to reported effects and target’s sample, that even with half of our sample we still have *n* = 125 per condition (half of 250), more than 2.5 the sample size of the target’s Study 2 (92/2 = 46 per condition) and more than 4 times the sample in the target’s Study 3 (86/3=28), well above the Simonsohn (2015) Small Telescopes rule of thumb. However, running this design actually allows us to determine and control for order and by doing that achieve much higher power than running those separately even in the unlikely case that order has an impact.

(Please note: This reply has been used in some version in other replies to PCIRR feedback)

> **.2. The manuscript is overall well-written. However, I agree with reviewer 2 that the short paragraphs at some places impede the overall readability. Furthermore, reviewer 3 provided many useful references that can make the literature review more comprehensive, and further strengthen the motivation for the current replication. Please also double-check the revised manuscript to make sure all in-text citations are included in the references, and vice versa.**

Thank you for your suggestion.

Based on our reviewers' recommendations, we rearranged the paragraphs to try and increase the overall readability. We also revised the manuscript and made sure that the in-text citations are included in the reference list.

> **.3. American students will be recruited from Prolific for this study (Table 2). Please provide more details on how the participants will be selected from the overall population on Prolific (e.g., what pre-screening options on Prolific will be used to recruit student participants from the US). This will help address the comment by reviewer 1, namely the scenario in study 3 is mostly relevant to students, but less so for non-students.**

Thank you for the suggestion. We welcome the opportunity to elaborate further.

We added the following to the Methods section in the main manuscript:

> We targeted US American students using Prolific’s filters. We restricted the location to the US using “standard sample”, we set it to “Nationality: United States”, “Country of birth: United States”, “Student status: Yes”, “Minimum Approval Rate: 90, Maximum Approval Rate: 100”, “Minimum Submissions: 50, Maximum Submissions: 100000”
>
> [We will first pretest the survey duration with 30 participants to make sure our time run estimate was accurate and adjusted pay as needed, 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.]
>
> [...]
>
> We employed the Qualtrics fraud and spam prevention measures: reCAPTCHA, prevent multiple submissions, prevent ballotstuffing, bot detection, security scan monitor, and relevantID.

Additionally, we have placeholders in the supplementary materials' "Additional information about the study" that aim to provide all the key information regarding our subject recruitment and all the parameters we used following data collection in Stage 2, including information about duration and compensation.

> **.4. Reviewer 1 also raised another interesting point, namely 200 dollars in 2006 would be worth almost 300 dollars today. When it comes to replications, there may be a tension between using the exact same stimuli, versus using stimuli that have similar 'psychological' meanings. I do not have a clear idea on this. For me, whether the amount matters or not in this case is eventually an empirical question, that can be tested by e.g. giving half of the participants the original 200-dollar version, and the other half the updated 300-dollar version. The amount can be included in the analyses as an extra factor, to (1) examine its potential influence, and (2) test whether the findings hold in both conditions. However, this comes with the cost of making the study less of a 'close' replication. I am curious to hear your thoughts on this.**

This is an interesting dilemma. If we consider the target's claims and theory there was nothing that we saw in the article or in the follow-up literature that indicates that this is an important factor, and in a way, the reason why we are doing replications is to try and assess that.

As you pointed out, this could go either way, changing from the original means that the change may lead to unexpected differences due to unforeseen factors embedded in the change, whereas not changing from the original may lead to a failure given the change in meaning of the stimuli. Inflation is a very generalized factor and it is not the only factor that changed over the years, it is just that this is the one that is salient in this scenario. If one starts changing the scenario, how can we determine what to keep and what to change? Once you've decided that you do want to change things, how do you know what to change it to and how do you know that it will have the same impact on participants? (e.g., does it need to be a round number? does it need to have a certain digit?)

In our view, in this specific scenario the change in value of money does not seem to fundamentally impact the overall meaning of the scenario. It would be somewhat disappointing if we were to find that a classic effect like the one reported in Tsang (2006) were only to hold in a specific scenario for certain sums of money. This is also part of why direct replications are so important, because we want to know if a certain paradigm does not work as well today as it did back in the day. If this does not hold, then we will know to adjust our priors regarding the strength and generalizability of the phenomenon, and follow-up research that wishes to investigate further can examine various moderators. However, whenever possible, we would consider it best to start from stimuli as close as possible to that of the original, and if that does not work, then follow-up with adjustments, rather than the other way around. Starting with changes to us seems like a much riskier strategy than starting with the target's own stimuli, seeing what does not seem to work and why, and adjusting accordingly.

As a sidenote, in our experience with replicating many JDM effects dating as far back as the 1970s, most of the effects held regardless of time, context, or value of money. It does not mean that it would not matter here, it is just a general observation regarding our experience so far.

We see the value of discussing this, and added this as a planned discussion in Stage 2:

> Planned discussion in Stage 2 following Dr./Prof. Jo-Ann Tsang's comment: $200 in 2006 is fairly equal to $300 in 2023. We will discuss the dilemma of whether to change stimuli in both our study and more broadly for replications, and our decision not to change the stimuli, with calls for future research to conduct more regular replications, to state clearly theoretical factors and predictions that might impact the effects and future replications, and to examine moderators like amount of money involved.

> **Below are some notes/thoughts that I had while reading the manuscript myself:**
>
> **.5. Hypotheses 1 (1a-1c) and 6 in Table 1 are about the correlations between gratitude and indebtedness in different helper intention conditions. It is not entirely clear to me how these correlations may address the core hypothesis in Tsang (2006) and in the current replication, namely "benevolent (versus selfish) intentions were more strongly associated with gratitude than with indebtedness" (Page 12).**
> **I can see how the results of ANCOVA and regressions can answer this question, but I am not sure which pattern of correlations between gratitude and indebtedness would support or refute the core hypothesis. I may have missed it, but the introduction also does not discuss the correlation between gratitude and indebtedness. If the correlation between gratitude and indebtedness is an important piece of evidence for the core hypothesis, you may need to discuss this more explicitly and extensively in the introduction, especially on how it is related to the core hypothesis. However, if the correlation is not central, it may be better to move them into the supplementary materials, or make a distinction between primary vs. peripheral tests.**

This is an excellent comment and we agree. In our revision, we tried to be clearer about this and the criteria for replications. We updated Table 1 with bolding the core hypotheses and adding in the note that “Bolded hypotheses are the core hypotheses which will be used to test the replicability of the target article.”. We also added the following text:

> We provided a summary of the hypotheses and their corresponding findings in Table 1 (see “analysis of the original article” subsection of the supplementary materials for further details). The target article had many hypotheses and many associated analyses, and we therefore pre-registered that our replication criteria will focus on the following. In our replication of Study 2 our focus was on the comparison of Hypotheses 2 and 3: Impact of intent (benevolent > selfish) on gratitude is stronger than on indebtedness.” . In our replication of Study 3 our focus was on the comparison of Hypotheses 7b/c and 8b/c: “Impact of intent (benevolent > ambiguous > selfish) on gratitude is stronger than on indebtedness.”.
>
> Given the two studies, we pre-registered our overall strategy to conclude a successful replication if the findings of the two studies are aligned with a signal in the same direction as the target article by Tsang (2006), mixed results if only one of two is supported, and failed replication if we fail to find support for the hypotheses in both studies.

Given that all these hypotheses and analyses are sub-analyses underlying the combined analyses, and were reported in full in the target article, we feel it is important to provide a full account of the findings in the main manuscript, rather than in the supplementary.

We also added the following note in the results section:

> [Addressing the comment by Editor Dr./Prof. Zhang Chen we will structure the replication section such that we flag and focus our reporting on the core hypotheses.]

We also added subheadings to each results subsection to make the core analyses more salient:

> [Study 2] Core hypothesis: Impact of intent (benevolent > selfish) on gratitude is stronger than on indebtedness.
>
> [Study 3] Core hypothesis: Impact of intent (benevolent > ambiguous > selfish) on gratitude is stronger than on indebtedness.

In addition, we also used this opportunity to clarify some of the hypotheses in Table 1 that were not clear enough in their predictions.

Finally, in the results section for both Study 2 and Study 3 we added “Interaction between intent and emotions (gratitude vs. indebtedness): Extension analysis of a direct test to core hypothesis” to directly test the interaction between intent and emotions to directly contrast the two. We added these analyses to our R code, and to Table 8 summarizing the findings mirroring the hypotheses in Table 1.

> **.6. For the regression analyses, it is not entirely clear to me what the predictor 'helper intention' refers to, either (1) the different conditions that participants are assigned to, or (2) the perceived helpers’ motivations (i.e., DV 4). If the former, I think the ANCOVAs and regressions are essentially the same, but presented in slightly different ways. Both use helper intention conditions and the magnitude of a favor as predictors, and gratitude or indebtedness as the outcome. Squaring the t value from the regression should give the F value from the ANCOVA, and the p values should be the same for each effect. If the authors can verify that these two tests are equivalent, combining them in the results section will simplify things. Table 1 can also be simplified (i.e., no need to repeat the same predictions twice).**

Thank you, this is valuable feedback.

In our revision, we took the following steps:

1. We reframed and clarified the hypotheses. For example, we reframed the combination of 2 and 3 to: “Combined: Impact of intent on gratitude (benevolent > selfish) is stronger than on indebtedness.”
2. We agree that the ANOVAs and the regressions are very similar, and that was meant to follow the target’s analyses and to report their findings clearly. We left it in the table in the same way, to allow an easier link between the replication and the target, and to allow us effect size comparisons following data collection.
    a. We now numbered Hypotheses 4, 5, and and 4 + 5 as 2r, 3r, and 2r + 3r in parentheses.
    b. We added the following note to the table: “Hypotheses 4, 5, and and 4 + 5, are re-analyses of the hypotheses 2r, 3r, and 2r + 3r.”
    c. We added “[Regression complementary analysis]” with the same text.
3. We made it clear in the table which are core hypotheses and added a paragraph explicitly stating that.

> **.7. I feel the proposed statistical tests do not provide a direct and formal test of the core hypothesis. For instance, in study 2, the predictions 2+3 (and other predictions involving comparisons between gratitude and indebtedness) seem to rely on a descriptive comparison of effect sizes between both conditions, but not formally tested.**
> **One may test this directly, e.g. by using (1) the helper intention condition, (2) the magnitude of a favor, and (3) the type of emotion examined (gratitude vs. indebtedness; within-subjects) as predictors, and the reported strength of an emotion as the dependent variable. Is it correct to say that the core hypothesis would then translate into a statistically significant interaction between factors (1) and (3)? If yes, I think it would be informative to conduct such an analysis, as another 'extension' of the original findings.**

Thank you, great suggestion. Yes, we agree, that is something that was missing in the target article.

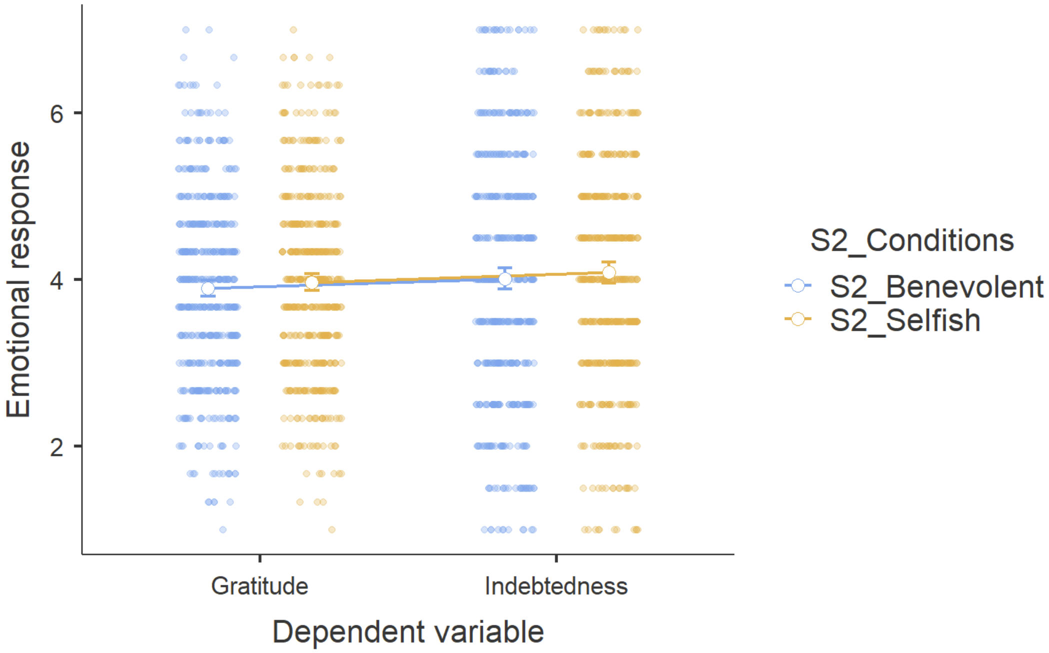
As a close replication, we initially prioritized replicating the procedures and analyses in the original article as closely as possible over modifying the original research method. However, we agree that this is an interesting and important extension that would allow a more direct test of the target's hypotheses.

We added this to our R/Rmarkdown code, added this to our "Data analyses strategy" with dedicated sections in the results section for both Study 2 and Study 3. The following is the example for Study 2:

> ***Interaction between intent and emotions (gratitude vs. indebtedness): Extension analysis of a direct test to core hypothesis***
>
> [To be completed in Stage 2]
>
> We conducted a mixed ANOVA examining the interaction between intent (benevolent versus selfish; between-subject) and emotion type (gratitude versus indebtedness; repeated) and found… (Figure 5)

Figure 5

Study 2: Interaction between helper intention and emotions



Note. Scale is from 1 to 7, higher values indicate stronger feelings of the emotion.

.8. I am not sure if I fully get the predictions when combining the findings from Tsang (2006) and Watkins and colleagues (2006) on Page 15. Watkins et al. found that high expectations for reciprocity would increase indebtedness but decrease gratitude.

Assuming that "benevolent giving may be associated with lower expectations for reciprocity than selfish giving", my chain of reasoning is that benevolent giving -> lower expectations for reciprocity -> decreasing indebtedness and increasing gratitude. It is unclear to me why "according to the findings by Watkins et al. (2006), benevolent giving may result in more indebtedness than gratitude, the opposite of the predictions by Tsang (2006)."

Thank you for pointing this out. Good catch.

We changed it to the following:

> We built our extension on the findings by Watkins et al. (2006) who, like Tsang (2006), argued that gratitude and indebtedness are distinct and that expectations for reciprocity would increase indebtedness but decrease gratitude. Tying these findings together with the experimental paradigm of Tsang (2006), we aimed to examine the associations between benevolent intent and expectations for reciprocity. If benevolent intent is associated with higher expectations then it would, according to Watkins et al. (2006), be associated with increased indebtedness and decreased gratitude. However, if benevolent intent is associated with lower expectations then it would, according to Watkins et al. (2006), be associated with decreased indebtedness and increased gratitude. Therefore, if we were to try and tie the two sets of findings together then the more theory consistent association seems to be that higher benevolent intent is associated with lower expectations and therefore higher gratitude than indebtedness.

> **.9. Please provide more details on potential data exclusion criteria. E.g., do participants need to pass all comprehension checks in order to be retained in the analysis? I wonder if there are other data quality checks. Especially for study 2 where participants have to recall their past experience and type it into open-ended questions - I can imagine some online participants may not be very motivated to do this. Are there any other measures that may be used to filter out low-effort responses, such as extremely fast responses or short answers?**

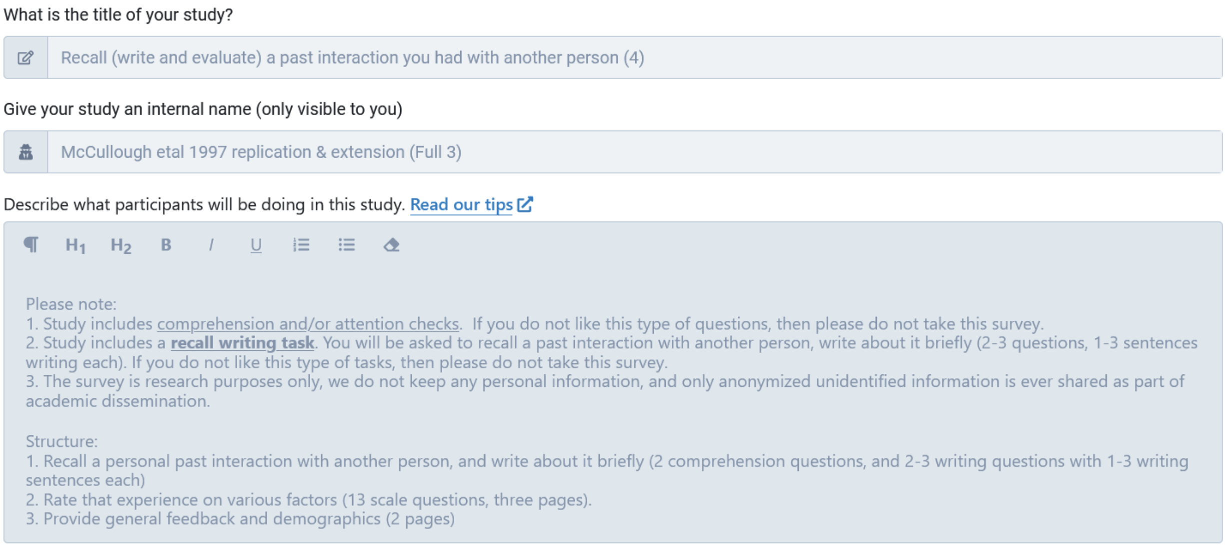
Thank you. This comment helped us realize we could improve on clarity.

We had both scenario/task comprehension questions, which participants had to answer correctly in order to proceed to the main task. In addition, there were manipulation checks, in order to ensure that the manipulation was successful.

Our first action in the revision was to make the manipulation checks clearer and so in the design tables and in the “Helper intention” heading we added “(manipulation check)” and in the results section made it clearer that “Helper intention” was the manipulation check.

The task recall open questions are very brief, with each question being 1-3 or 1-2 sentences long. We therefore are aiming for short answers. The writing task itself is not the important part, but rather the recall itself and the evaluations and the attributions regarding that situation. We ran far more demanding writing tasks with the Prolific population with very good results. In our experience this comes down to aligning expectations with the target sample in advance and so in our recruitment we make it clear that there is brief writing involved, and in the consent screen qualifying questions there are two specific questions that participants must answer with a "yes" and a copy-paste acknowledging that they know and understand that the task involves writing.

We provide an example for a Prolific recruitment for a study with similar writing task of a replication and extension of McCullough et al. (1997):



What is the title of your study?

Recall (write and evaluate) a past interaction you had with another person (4)

Give your study an internal name (only visible to you)

McCullough etal 1997 replication & extension (Full 3)

Describe what participants will be doing in this study.

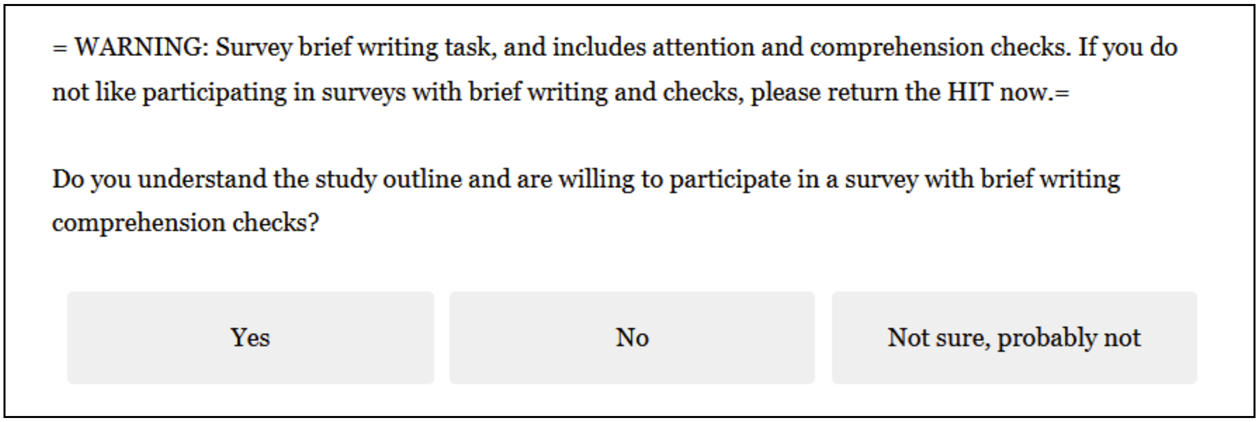
Please note:
1. Study includes comprehension and/or attention checks. If you do not like this type of questions, then please do not take this survey.
2. Study includes a **recall writing task**. You will be asked to recall a past interaction with another person, write about it briefly (2-3 questions, 1-3 sentences writing each). If you do not like this type of tasks, then please do not take this survey.
3. The survey is research purposes only, we do not keep any personal information, and only anonymized unidentified information is ever shared as part of academic dissemination.

Structure:
1. Recall a personal past interaction with another person, and write about it briefly (2 comprehension questions, and 2-3 writing questions with 1-3 writing sentences each)
2. Rate that experience on various factors (13 scale questions, three pages).
3. Provide general feedback and demographics (2 pages)

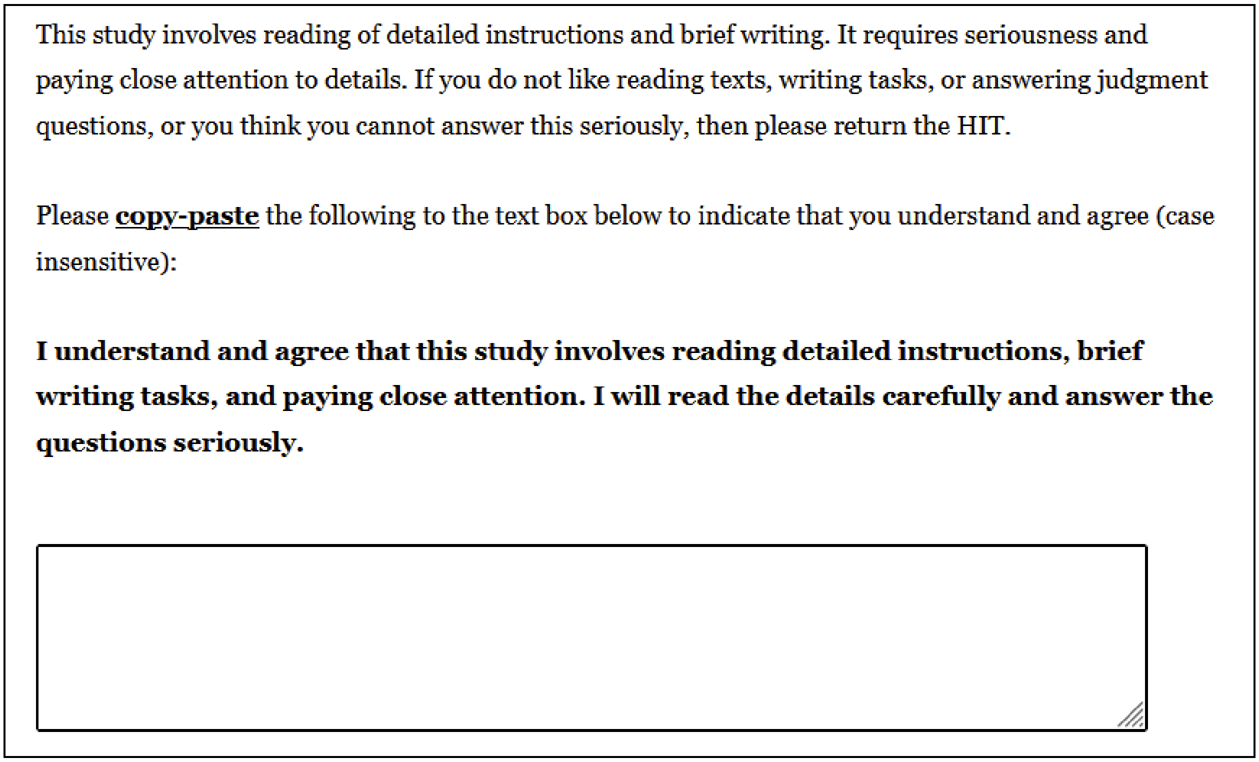
The above project (same lead and corresponding authors as this project) is currently in final revisions of PCIRR Stage 2 revise and resubmit, and has been one of our clearest smoothest successful replications, showing that the Prolific sample participants take writing tasks very seriously and with results comparable to the target 25 years later and with meaningful causality extensions:

> Chan, C., & Feldman, G. The impact of Empathy on Forgiveness: Replication and extensions Registered Report of McCullough et al. (1997)'s Study 1. [Stage 2 preprint] [In-principle acceptance/Open peer review] [Open materials/data/code]

In addition, we provide a screenshot of the study design consent page in the Qualtrics which includes two sections relating to aligning expectations regarding writing:


= WARNING: Survey brief writing task, and includes attention and comprehension checks. If you do not like participating in surveys with brief writing and checks, please return the HIT now.=

Do you understand the study outline and are willing to participate in a survey with brief writing comprehension checks?


This study involves reading of detailed instructions and brief writing. It requires seriousness and paying close attention to details. If you do not like reading texts, writing tasks, or answering judgment questions, or you think you cannot answer this seriously, then please return the HIT.

Please **copy-paste** the following to the text box below to indicate that you understand and agree (case insensitive):

I understand and agree that this study involves reading detailed instructions, brief writing tasks, and paying close attention. I will read the details carefully and answer the questions seriously.

In addition, we provide details regarding our comprehensive checks for both studies in the supplementary. For your reference, we copied the relevant section below:

> Comprehension Check
>
> *Study 2*
>
> We added two comprehension-check questions for Study 2. The two questions were designed based on the instructions. Participants were not allowed to proceed to the next page unless they answered the questions correctly. The answers were presented in randomized order.
>
> 1. What type of helping behavior are you asked to recall?

   a. Unselfish (benevolent) help
   b. Selfish help
   c. Any kind of help

   Ans: a. Help someone else gave to me in *Benevolent Condition*

   b. Selfish help in *Selfish Condition*

2. Whose helping behavior are you asked to recall?
   a. Help I gave to someone else
   b. Help someone else gave to me
   c. Any kind of help

   Ans: b. Help someone else gave to me

Study 3

We added three comprehension-check questions for Study 3. One was directly extracted from the original study, with additional two new questions. Participants were not allowed to proceed to the next page unless they answered the questions correctly. The answers were presented in randomized order.

1. How much money did the friend offer to give to help pay for the textbooks?

   Ans: 200

2. What was the favor offered in the scenario?
   a. birthday gift
   b. helping you with your homework
   c. paying textbooks for you

   Ans: C. Paying textbooks for you

3. According to the text: Why is your friend offering to help you?
   a. I know without doubt it is because my friend wanted to borrow my car.
   b. It is not clear about the two being related, but the weekend after helping me this friend asked to borrow my car.
   c. My friend is really concerned about me

   Ans: a. for *Benevolent Condition*

   b for *Selfish Condition*

   c for *Ambiguous condition*

Finally, we did not plan to pre-register any exclusions but rather run additional exploratory analyses in case we fail to find support for our predictions, yet we see the value in making those plans explicit in advance. We therefore added the following to “outliers and exclusions”:

> We pre-register that if we fail to find support for our hypotheses that we rerun exploratory analyses for the failed study by excluding participants who failed the manipulation checks. To compensate for multiple comparisons and increased likelihood of capitalizing on chance, we will set the alpha for the additional analyses to a stricter .001. We plan this as a second level analysis only after the “Order effects” analyses above also fail to find support for the analyses.

.10. Some minor points:

Page 4: In the abstract, the effect size of helper intention on indebtedness in Study 3 is outside of the 95% CI ("η2p = .14, 95% CI = [0.00, 0.03]").

Thank you. Corrected.

Page 7: "We then discuss our motivations for the current replication review and review Tsang (2006)...". The first 'review' should be removed?

Thank you. Corrected.

Page 11: "Especially so given that the target article sometimes theorized using null effect language and concluded no differences from null effects.". This sentence is not entirely clear to me.

Thank you. We reframed it to the following:

> In addition, the target article presented a theoretical model that predicted no effects for the impact of intent on indebtedness. In their findings they also reported failing to find a signal in support of rejecting the null hypothesis of finding no differences for indebtedness between the benevolent and selfish intent conditions, and built on that to conclude no effects. However, Null Hypothesis Significance Testing (NHST) methods are not well suited for testing and quantifying support for a null hypothesis. We felt it important to revisit the theoretical model by reframing the null hypothesis to differences in effects between gratitude and indebtedness, to rerun the studies with well-powered samples, and to add additional analyses that address the null hypothesis issue to gain deeper insights into the phenomenon.

> **Page 16: "Therefore, our extension ties and contrasts the predictions by Tsang (2006) and Watkins et al. to examine how helper intentions are tied." This seems to be an incomplete sentence?**

Thank you for catching that. We changed it in the following way:

> In summary, our extension ties and contrasts the predictions by Tsang (2006) and Watkins et al. (2006) and by Bartlett and DeSteno (2006) and Peng et al. (2020) to examine: 1) the associations between helper intentions and expectations for reciprocity, and 2) the relationship between reciprocal behaviors, gratitude, and indebtedness.

> **Page 18: "Effect size and confidence intervals were all calculated with Rstudio (Version: 1.4.2)". I think it's important to also report the version of R used - after all, R is doing all the computing, and RStudio is mostly an IDE for R.**

Yes, we agree. We changed to report the Version of R, i.e., 4.1.2.

> **Page 19: The planned sample size is inconsistent, being 800 at some places but 1000 at other places.**

Thank you for your question. 800 is our target sample size, whereas 1000 is the sample generated with our simulated data for demonstrating our result section. We stated at the beginning of the "Participants" subsection under the method section.

> *[To demonstrate what the results would look like after data collection we simulated a dataset of 1000 participants using Qualtrics and reported our analyses below based on that dataset. Results will later be updated in full to a sample of 800 and the real data.]*

This is all meant as a simulation of what the manuscript would look like following data collection in Stage 2, which will be updated to consistently report the real data.

> **Page 25: "including questions about what flavor was offered in the scenario". "flavor" should be "favor".**

Thanks for catching that. Fixed.

> **Page 31: "we used correlation tests (Pearson's Correlation) to examine the association between helper intention conditions (benevolent and selfish) and emotions (gratitude and indebtedness)".**
> **If I understood this correctly, the correlation tests are to examine the association between gratitude and indebtedness across different helper intention conditions (but see my comment above)?**

Yes, thank you, we appreciate the feedback to improve on clarity. The sentences you quoted here were meant to mirror what we reported in more detail in the results section on the simulated random data, yet we realize that our framing was confusing and not entirely in line with the results section.

We therefore modified that section to the following ("Data analysis strategy -.Replication: As in the original")::

> In both Studies 2 and 3, to mirror the target's analyses we first ran (Pearson's) correlation tests to examine the associations between gratitude and indebtedness across conditions and then in the separate benevolent and selfish helper intention conditions.
>
> In Study 2, we used ANCOVAs to examine the effect of helper intention (benevolent versus selfish) on gratitude and indebtedness, with the magnitude of favor as the covariate. We supplemented those with regression analyses using the same factors which served a similar purpose to the ANCOVA, and merely meant to mirror that target article's analyses and reported effects.
>
> In Study 3, we used one-way ANOVAs to examine the impact of helper intention (benevolent versus selfish versus ambiguous) on gratitude and indebtedness. After that, we conducted planned comparisons to examine the differences in emotions between helper intention conditions.

Reply to Reviewer #1: Dr./Prof. Jo-Ann Tsang

> **I am excited to see this replication and extension of work on intention, gratitude, and indebtedness. I am glad that the authors are recruiting a bigger sample size, adding more manipulation checks, and measuring the additional dependent variable of reciprocity intentions. Below are some suggestions and questions that I have which I hope will helpful to the researchers in conducting their study.**

Thank you for the positive opening note and the detailed feedback. We are very grateful for your openness throughout the process and for sharing your original materials - they were extremely valuable in preparing this study.

> **.1.- I may have missed it, but I don't think the authors specified why they were replicating Studies 2 & 3, but not Study 1. When reading Tsang (2006), it is obvious that one would try and replicate Study 3 rather than Study 1, but readers might not be familiar with the methods of the original studies and therefore might miss this.**

Great point, thank you for your suggestion. We added a sentence under the section “Original hypotheses and findings in the target article” to briefly explain why we chose to replicate Studies 2 and 3 but not 1:

> We focused our replication on Tsang’s (2006) Studies 2 and 3, given that Study 3 contains all the essential experimental designs of Study 1 with an extra condition of ambiguous helper intention for investigation.

> **.2. - The hypotheses in Table 1 (p. 13) that were reworded from the null (1c, 3, 5, 8a, 8b, 8c, 10) were confusing to me. I understand the need to reframe the original null predictions, but the reframed hypotheses were making predictions that the original paper did not make. The combined hypotheses made more sense to me as a reframing. I'm not sure if there is a way to clarify this.**

Good point, we agree this can be confusing. To address this specific point we added brackets with the words indicating the null and added the following to the Table 1 footnote:

> Hypotheses 1c, 3, 8a, 8b, 8c were originally null hypotheses, yet we reframed those to a testable alternative to the null, with indication of the null hypothesis in brackets (e.g., “[not]”). Similarly, the combined Hypotheses 1b+1c, 2+3, and 7b/c+8b/c reframed a the

the null hypotheses from from 1c, 3, 8a, 8b, 8c to a testable hypothesis expecting stronger effects for gratitude compared to indebtedness.

Hypotheses 1c, 3, 8a, 8b, 8c are now written in the following form:

- Gratitude is [not] associated with indebtedness in Benevolent condition. [Reframed from the original's null hypothesis]
- Benevolent favors [do not] result in more indebtedness than selfish favors, even after controlling the magnitude of favor. [Reframed from the original's null hypothesis]
- Indebtedness is [not] different between the three conditions (Benevolent, Selfish, and Ambiguous). [Reframed from the original's null hypothesis]
- Indebtedness is [not] higher in Benevolent condition compared to Ambiguous condition. [Reframed from the original's null hypothesis]
- Indebtedness is [not] higher in Benevolent condition compared to Selfish condition. [Reframed from the original's null hypothesis]

> **.3. - at the bottom of page 15, the authors theorize about the relationship between intention and reciprocity, but then on p. 16 make predictions about gratitude and indebtedness. this was a little unclear to me; I was expecting the predictions to be about intention and reciprocity given the theorizing. perhaps there is more the authors can say about theory related reciprocity and gratitude/indebtedness before they get to those predictions that will make the argument a little more clear.**

Thank you for the comment, we agree that more clarity is needed. We followed on this point in our reply to the editor's point number #8, and revised the manuscript accordingly.

> **.4. - Method: I see the rationale for running Study 2 & 3 within-subjects, but I am worried that this may subtly influence the results. For example, if a participant is assigned to a benevolent intention condition in Study 2, but then is assigned to a selfish condition for Study 3, this may introduce a contrast effect--selfish favors may seem more selfish after writing about a benevolent favor. Participants who are asked to read a textbook scenario first, might then be influenced by this scenario when they recall their own received favors in the subsequent study. Additionally, gratitude or indebtedness may be primed in ways not primed by the original studies if the studies are run together. Thus, running these studies together might compromise the closeness of the replication.**

Thank you for the comment, we agree that more clarity is needed. We followed on this point in our reply to the editor's point number #1, and revised the manuscript accordingly.

> **.5. - oddly, given my previous comment, I am also worried about the directness of the replication of the scenario study, in that the population from which the authors are recruiting are not all students, and also given the passage of time from 2006 until now. Specifically, the scenario from Tsang (2006) was designed to be relevant to the student population from which the participants were recruited. However, the current authors plan to recruit from a broader population, and it is likely that students will be in the minority in their participant pool. Thus, the scenario may be less relevant to their participants, and induce less gratitude and indebtedness.**

Thank you for the comment, we agree that more clarity is needed. We followed on this point in our reply to the editor's point number #3, and revised the manuscript accordingly.

> **The amounts used in the original scenario also mean something different today. For instance, $200--the amount lent to the protagonist in the original study--would be worth almost $300 today. Thus, using the same exact scenario today as Tsang used in 2006 would lead to participants reading about a less valuable favor.**

Thank you for the comment, we agree that more clarity is needed. We followed on this point in our reply to the editor's point number #4 on how we decided to address this issue.

> **.6. - it makes sense to measure reciprocation intentions as an extension. The authors might look at Peng et al. (2020) to inform predictions regarding gratitude, indebtedness, and reciprocation.**

Excellent, thank you very much for suggesting this citation. We agree, this is very relevant for our added extension. We therefore revised to include the following:

> Findings in the literature about the associations between gratitude and reciprocal prosocial behavior have so far been mixed. For example, a seminal study by Bartlett and DeSteno (2006) illustrated that gratitude is positively associated with reciprocity whereas Peng et al. (2020) failed to replicate Bartlett and DeSteno (2006) and did not find any support for links with reciprocity for both gratitude and indebtedness. Therefore, our extension could be thought of as a conceptual replication of the Bartlett and DeSteno (2006) and Peng et al. (2020) directions to try and determine whether reciprocity might play a role, using an empirical design from a different study.
>
> Therefore, our extension ties and contrasts the predictions by Tsang (2006) and Watkins et al. (2006) and by Bartlett and DeSteno (2006) and Peng et al. (2020) to examine: 1) the associations between helper intentions and expectations for reciprocity, and 2) the relationship between reciprocal behaviors, gratitude, and indebtedness. To the best of our knowledge, there is no investigation into the effect of helper intention on reciprocation magnitude, this report aims to extend Tsang (2006) in this direction.

Reply to Reviewer #2: Dr./Prof. Cong Peng

> **The current paper intended to offer a replication and extension of Tsang (2006) investigating the effect of helper intention on gratitude and indebtedness. Tsang (2006) suggested that perceived benevolent intention would trigger higher gratitude but would not affect indebtedness. Tsang (2006) is indeed a pioneer work and inspired many later researchers to differentiate gratitude and indebtedness.**

Thank you for the feedback and for the thoughtful comments.

> **However, there are severe problems of the current manuscript that blocks me from recommending to proceed to stage 2.**
>
> **.1. First, I think the current literature review on gratitude and indebtedness is very limited and far from comprehensive. There are accumulated literature suggesting the relation-oriented function of gratitude to promote intimate bonds (Algoe, 2012; Algoe et al., 2013; Bartlett et al., 2012; Gordon et al., 2012; Kubacka et al., 2011; Lambert et al., 2010; Ng et al., 2017; Peng et al., 2018; Williams & Bartlett, 2015), which is important to help clarify why benevolent intention is important to trigger gratitude. There are also accumulated literature suggesting the exchange-oriented function of indebtedness (Adams & Miller, 2022; Goyal et al., 2022; Naito & Sakata, 2010; Peng et al., 2018), which is important to clarify why beneveolent intention is NOT associated with indebtedness. Moreover, the current manuscript give me the impression that the authors lack to provide a systematic review of the literature and clear arguments but are listing literature loosely.**

We appreciate this suggestion, and we see the importance of a summary of the literature.

Our scope for this direct replication with extensions was rather narrow and focused on the empirical effort to reproduce and replicate the original findings, and so we initially kept our literature review very brief, mostly to explain how the target article was embedded in the broader literature. We see our main responsibility in choice of citations to be about conducting a rigorous empirical close replication of one target article rather than to provide a comprehensive review of the literature covering any/all papers published since. The task of reviewing and summarizing a literature is a major undertaking, and should have a dedicated systematic review/meta-analytic effort, preferably as a Registered Report, to address it comprehensively and make sure it lives up to its importance. Doing that in a replication manuscript runs the risk of distracting readers from what the manuscript was meant for - an empirical replication of one classic article in that literature.

To address this comment, we added a call for a systematic review of the literature as a future direction in the planned general discussion in Stage 2 under the “Limitations of our replication and directions for future research” subsection:

> [Planned discussion in Stage 2 following Dr./Prof. Cong Peng comment: We will discuss the need for a systematic review and meta-analysis of the literature pointing to the findings in the literature that built up on the target article.]

In addition, we added an extra paragraph under the section “Relationship and differences between gratitude and indebtedness”:

> Furthermore, research indicated that these two emotions play different functions in sociality. For example, accumulated literature suggested that gratitude contains a relation-oriented function to promote intimate bonds (e.g., Algoe et al., 2013; Bartlett et al., 2012; Kubacka et al., 2011; Peng et al., 2018; Williams & Bartlett, 2015), whereas indebtedness contains an exchange-oriented function (e.g., Goyal et al., 2022; Naito & Sakata, 2010; Peng et al., 2018). These functional differences may explain why helpers’ intentions are influential to one’s gratitude and indebtedness.

.2. Second, the manuscript writing lacks basic scientific rigor. There are many literature presented in the introduction but not listed in the reference, and I could not find them either on google scholar to judge their validity (e.g., Gray et al., 2001 on p7; Ortony et al., 1988 and Mathews & Green, 2010 on p8; Maureen & Jeffrey, 2009 on p9; Ames et al., 2004 and Welsh et al., 2021 on p10).

Thank you for catching that, we appreciate that. The references you pointed out are all from the target article - Tsang (2006). We have rechecked the in-text citation and ensured that they are in our reference list.

> **.3. Meanwhile, in many cases, the authors fail to provide reference for certain claims (especially when the claims are big), making it difficult for me to make sense of it.**
> **Some examples are: The starting sentence in background on p7: "Gratitude and indebtedness are common emotions in response to receiving help. But studies suggested that they are experienced differently depending on situation". And on p8 line 5, "These two emotions have often been equated in the early literature, yet evidence showing that these emotions are elicited in different situations suggested the need to differentiate them.").**

Thank you for raising that.

We are focused on replicating Tsang (2006) and followed their review of literature, arguments, hypotheses, and methods. We were already citing the target article very often, and therefore tried to avoid repeating a citation to the target for every other sentence regarding every claim the target made. These broad claims were at the very core of the target article, as for example, the first sentence of the Tsang (2006) abstract:

> Gratitude and indebtedness have often been equated in psychology. Emerging research, however, suggests that these emotions are experienced differently and occur in response to different situations.

To address this specific comment, we added the relevant references for these specific claims.

> **.4. Third, the current replication is making things too complicated to be qualified as a replication. I wonder why the authors considered to combine two separate studies in Tsang (2006) into one study rather than replicating them separately. This is not a replication anymore, as the design in either study may affect the other. For whatever results in the end, it would be very hard to interpret and compare with the original study. Let alone the authors are extending it to mix with the design of Watkins et al. (2006), making it even further from a replication. I think a good replication design should be as close and comparable to the original study as possible.**

Please see our reply to the editor's #1 comment on this point. We have been very careful not to affect anything in the studies we replicated and instead to build on top of those in a way that would not affect the replication and instead offer additional insights.

In your own replication of Bartlett and DeSteno (2006) you did something very similar:

> This was an extended replication of Study 1 in Bartlett and DeSteno (Citation2006). We tried to stay as close as possible to the original study apart from the following differences. [...] We also added three items that were constructed to measure indebtedness. Fourth, following communication with the original authors, we slightly changed the dependent measure of helping behavior. [...]

Our combining of the two studies in randomized order does not impact the first run study, and actually allows us to further test whether there are any implications for order, and our added extensions were variables included on top of the target's dependent variables. A good replication is one that makes adjustments that can add to better understand the findings, whether successful or not, a strategy that we both implemented in our projects.

Reply to Reviewer #3: Dr./Prof. Sarahanne Miranda Field

> **I am generally very positive about this Stage 1.**
>
> **- The replication protocol is very clearly set out, and it appears to me as though, providing the replication study is conducted closely to how it has been described here, the replication has a very good chance of reinforcing the effects in question, should they 'exist'. Importantly, I think this replication study protocol leaves little room for flexibility or bias, which is important for meaningful and high-quality replication studies.**
>
> **- The theoretical background is clear and follows logically, and motivates the study sufficiently. The target sample size is motivated also, and seems reasonable.**
> **- The planned statistical approach seems appropriate to me (although I will freely admit that I am no expert on these kinds of analyses).**
>
> **- Although I typically suggest using Bayesian methods to compare replication targets (as they allow one to quantify pro-null evidence), I am interested to see how the authors use the LeBel method for this study.**
>
> **- Manipulation checks and controls seem sufficient for purpose, from what I can tell. All in all, I look forward to what the results show and whether Tsang's original findings are supported.**

Thank you for the positive opening note and the detailed and constructive comments.

> **I have only three tiny quibbles (in no particular order):**
>
> **1. This article, while generally clearly written and free of obvious typing errors, is plagued by very tiny paragraphs. I'm not a nit-picker usually, but these paragraphs are so small as to be distracting and sometimes make the reading more difficult than it should be. Particular examples of where paragraphs could be merged are on pg 11 ("We chose..." might be merged with "The article has...") and pg 12 ("Tsang (2006) examined..." might be merged with "We focused our...") and so on. This isn't a huge deal-breaker, but readability would be improved, in my opinion, if the structure of the article were revised with this in mind.**

Thank you for your recommendation. As you suggested, we combined the paragraphs. We also went through the manuscript and tried to better reorganize its structure with the aim of combining disjoint paragraphs and increasing readability.

> **2. I find some of the figures a little hard to read. Violin plots are great and the figures generally look very good, however the raw data points on some of them are quite large and very transparent, which makes the distributions hard to see underneath the boxes. Figures 3, 4 and 8 for instance are great - the data points in those are smaller and you can clearly see the way the data are distributed, however Figures 5 and 6 (etc) are harder to make out. Not a huge issue, but given that the data in some of the figures seems quite evenly distributed, it's harder to see the distributions' nuances.**

Thank you for raising this point. Our present analyses were based on random simulated noise data, and that is why their distributions look similar and without any specific patterns.

This is something that we aim to examine and optimize in Stage 2. This should look very different once we replace the simulated data with the real data, and we will work towards improving the figures to make sure they can relay the findings well.

> **3. The figures are numbered weirdly, unless I'm missing something. It appears as though there are two Figure 8s?**

We renumbered and rechecked all tables and figures.

> **.4. As I mentioned above, this replication study plan looks very well thought-out to me, and will be a reasonable 'test' of Tsang's Study 2 and 3 (to the extent that a single replication can be, of course!). I wish the authors luck with the data collection, should they get the go-ahead!**

Thank you very much for your time and feedback. We appreciate it very much.