Reply to PCIRR decision letter reviews:
Barasch et al. (2014) 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.

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/QZCQPYFhjOgC

A track-changes manuscript is provided with the file: "PCIRR-S1-RNR-Barasch-etal-2014-replication-extension-Registered-Report-main-manuscript-trackchanges.docx" (https://osf.io/se7cg)

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
Introduction	Ed & R1: Clarified and added hypotheses in Table 1 and Study Design Table where relevant
Methods	Ed: Edited explanation of self-focus and perceived emotional benefits measures to reflect changes made on Qualtrics R1: Clarified that alpha value is set at 0.05 for all extension measures

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. Romain Espinosa

> **Dear authors,**
> **Thank you very much for your submission. I have read your paper with great interest and received feedback from two reviewers. Given this feedback and my own reading of the paper, I recommend a minor revision to address the relatively minor concerns that Thibaut raises and that I also noted while reading your manuscript. Angela finds your manuscript ready for IPA.**
>
> **Thibaut suggests to improve the presentation of the hypotheses (something that I also noted). He also suggests making a clearer distinction between the confirmatory and exploratory analyses, i.e., what you clearly replicate and what you explore. While you might have already stated this in some parts of the manuscript, it might be appropriate to increase its salience. Thibaut also suggests MHT correction if the extensions are part of your confirmatory analyses. More generally, even if you disagree, I think that it calls for making more explicit your decision rule (i.e., the significance threshold that you shall use for your registered hypotheses). Last, the referee is not convinced by the use of simulated data. I see it as a way to test your codes and understand why you included them. Please do not feel compelled to change the manuscript on this point.**

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

> **.1. I am not sure about one sentence in hypothesis 4 in Table 1. You say first that « expecting emotional rewards serve as a positive signal of moral character but expecting reputational and/or material rewards serve as a negative signal of moral character. » For me, there are three groups here: (i) people who don't expect any reward, (ii) people who expect an emotional reward, and (iii) people who expect a material reward. How do these groups sort? In the Design table, you say that people who expect to receive a material reward are perceived as less moral « yet less so when they expected to receive emotional benefits ». So, people from (ii) are perceived as less or more moral than people from (i)? (I understand that people from (iii) are the lowest, but I am not sure about your prediction about group (ii) relative to group (i).)**

Response: Great points, thank you. We can understand the confusion, and this feedback helps us to clarify things better.

Action: We revised to better align the study design table with Table 1 and across all hypotheses worked to make it clear which of the conditions are contrasted and to better align the methodological design with the hypothesis. Please see Table 1 for all the changes. Specifically, about the issue you raised here, the hypotheses were adjusted to the following:

> 4a: Donors who do not expect to receive material benefits for their donations are perceived as more moral than donors who expect to receive material benefits.
>
> 4b: Donors who do not expect to receive reputational benefits for their donations are perceived as more moral than donors who expect to receive reputational benefits.
>
> 4c: Donors who expect to receive emotional rewards for their donations are perceived as more moral than donors who do not expect to receive emotional benefits.

And we made similar adjustments in our framing of the extensions.

> **.2. I do not understand why there are 4 hypotheses in Table 1 and only 2 in the Design Table.**

We revised to better align the study design table with Table 1 and across all hypotheses

> **.3. I am puzzled by the pre-test participants. Of course, I completely understand that it is important to test the design, the code, the duration, etc. However, we can agree that this generates some degree of freedom for the researchers that is not observable by the reviewers. I think that one of the main ideas of RRs is that it is always (or at least very often) easy to find ex-post a rationale to justify a decision (like excluding some observations). The objective of RRs is to explicitly say which rule we are going to use once we start data collection. So, in my view, I would either recommend to (a) drop these 30 observations whatever happens, (b) to test the survey, duration, etc before the IPA, or (c) to explicitly say which rule you will apply to decide whether you want to keep these observations (« case of serious technical issues » is, in my view, too vague).**

Response: We are unsure what degrees of freedom you had in mind, we stated this specifically for the purpose of limiting our degrees of freedom, we were very explicit in how we plan to run those and what the pre-test is meant for. We wrote:

> [Stage 1 note: 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 the 30 participants will not be 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.]

The main concern as we understand it when it comes to flexibility is the concern that authors would evaluate whether to include or exclude the participants based on some kind of post-hoc outcome after running planned analysis on those participants. However, we explicitly write that we will not run any statistical analyses on the data of the pretest participants, and it would be pointless to do so on a sample of 30, these cannot tell us anything about the dataset or hypotheses (and given that this is a Registered Report with an in-principle acceptance regardless of outcome, this would serve no purpose). Therefore: 1) We only run our planned analyses on the full sample, 2) we only proceed to the full data collection when we see that there are no technical issues and pay is reasonable for the duration, and 3) in case of a technical issue the data is useless so we exclude the pre-test participants, and rerun the pretest.

The pretest is meant to ensure everything runs smoothly and participants are compensated fairly. Even if we pretest the survey beforehand, we cannot foresee all possible issues with tools like Qualtrics, Prolific, etc. at that specific moment in time for those remote participants. Therefore, we examine duration and the funneling sections that elicit feedback about technical problems.

When serious technical issues regarding the survey or platform, we exclude those. When there are issues that have to do with the survey content, we consult with the recommender.

> **.4. I am a bit surprised by the gender distribution in the US Prolific population. More than half of your participants report « other/did not disclose ». When I check on Prolific (US), there are 17,352 men (cis+trans), 27,802 women (cis+trans), 1,416 non-binary, and 271 « rather not say ». So, in prolific, there are less than 4% of Non-binary/rather not say, but you expect more than half of your sample. Why is there such a difference?**

Response: We believe this comment reflects a misunderstanding. We tried to be clear about how the data was generated, and wrote:

> [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 1164 and the real data.]

Therefore, all reported descriptives and statistics are based on simulated random noise created by Qualtrics (See Qualtrics's website on: Generate Test Responses). It has no relevance to real-life and nothing to do with Prolific or whatever we might expect from the real data. It will be replaced in Stage 2 with the real data.

> **.5. I am also puzzled by the treatment without donation (Study 3). I understand the original concern that we could not say much about a person's behavior when we do not know what she did. On the Qualtrics survey, in the treatment where donation is not said, there are two questions that I find intriguing:**
> **-To what extent is the person's behavior reflective of their intentions?**
> **-To what extent did the person donate to make him/herself feel better? [The person did not donate, no?]**
>
> **Regarding the first item, I do not have a strong opinion but I feel that « behavior » might refer to different things in different treatments (i.e., the physical response in case participants don't know the donation, and the donation in case they know it).**

Response: Thank you, great feedback.

Regarding the first item: We gave the word choice much thought and decided on 'behavior' as it signals both action (donating or receiving cash) and inaction (control condition; targets' decisions were ambiguous).

Regarding the second item, thank you for raising that, much appreciated! We realized an oversight in the Qualtrics. The item you referenced was our second self-focus question, planned to be combined with the first self-focus question (SF1Q; “To what extent was the person thinking about him/herself”) to form a measure of self-focus - We took this method from Study 2 of the original article.

However, we realized that this created a situation in which two questions were too similar:

- Self focus question 2: To what extent did the person donate to make him/herself feel better?
- Emotional benefits motivation: To what extent do you agree with the following statement: The person donated to the African Children's Fund to make himself feel better?"

Going back to the target article, the authors of the target article noted that they used the emotional benefits motivation measure from Study 1 and combined it with SF1Q to form the measure of self-focus:

> “Participants were also asked, “To what extent was the donor thinking about him/herself?”. **This item loaded with one item from the previous study** (“To what extent did the donor donate to make him/herself feel better?”) , and these two items were combined to create a measure of self-focus…” (p. 399).

Thanks to your feedback, we now made revisions to address this issue.

The original study seemed to have used the self-focus question 2/emotional benefits motivation both as a measure of self focus AND a measure of motivation to receive emotional benefits which is evidenced by the ‘Other vs self-focus’ and ‘Emotional benefits’ subheadings in the Results section (p. 400). We decided against it, and will analyze them as separate constructs: The question of “To what extent did the person donate to make him/herself feel better” seemed to align more with *receiving emotional benefits* rather than a tendency to focus on the self.

We believe that the existing questions: “To what extent was the person thinking about him/herself” (SFQ1) and “To what extent was the person thinking about others?” (Other-focus) adequately capture self and other focus that they were meant to measure.

Action: We deleted the duplicate self-focus question from all conditions, and added an emotional benefits measure to the control condition of Study 3. We adjusted the R code and regenerated a new set of simulated random data to ensure that everything works as intended. We reran the analysis and so the Rmarkdown output ensures that the code is well suited for the dataset after

the changes we made (but made no changes to the results section in the manuscript, no point in replacing one set of random noise for another).

A comparison of the changes we made regarding the aforementioned measures is provided in the table below:

Before	After

Donation condition:
- SFQ1 (To what extent was the person thinking about him/herself?)
- SFQ2 (To what extent did the person donate to make him/herself feel better?)
- Other-focused ("To what extent was Jeff thinking about others")
- Emotional benefits motivation (To what extent do you agree with the following statement: The person donated to the African Children's Fund to make himself feel better?")
 	
Cash condition:
- SFQ1, SFQ2, Other-focused, Emotional benefits motivation
 	
Control:
- SFQ1 & SFQ2, Other-focused.
	All conditions:
	- Self-focus (SFQ1; "To what extent was Jeff thinking about himself?")
	- Other-focused ("To what extent was Jeff thinking about others")
	- Emotional benefits motivation: ("To what extent do you agree with the following statement")
	  - "The donor donated to the African Children's Fund to make himself feel better" (Donate condition)
	  - "The donor opted for taking the cash to make himself feel better" (Cash condition; reverse-coded)
	  - "The person wanted to make him/herself feel better" (Control condition; reverse-coded)
	
	(treated as separate constructs/i.e. will not be combined to form a measure of self-focus as was done in the original article)

Suggestions (not mandatory):

.6. I would personally support the exclusion of the participants who fail the attention check because when they are forced to respond, they might then only look closely at the question which prevents them from going further, without paying attention to the rest of the questionnaire. In my view, accepting all participants could lead to an increase in the noise in the data and, thus, decrease the probability of detecting an effect (statistical power). I have no strong opinion here but I just wanted to make the point clear.

Response: Thank you for sharing your views on this.

We agree that there are trade-offs, but in our experience with online samples this has been the best way to communicate a serious survey that validates actually reading and understanding to minimize participants just clicking through the survey (random select, next, random select, next) that often results in unneeded high exclusion rates, waste, and smaller samples. The forced validation has been valuable in 1) ensuring participants read, understand, and process the important manipulation in the scenario, and 2) signaling to participants that we do checks and ensure attentiveness and seriousness. In our experience, it increases the chances of detecting an effect, given that: 1) mere fact that it has a larger sample than if some answers would be excluded (some fail attention checks even when very attentive), 2) allows serious participants to double check they attended to the crucial aspect of the scenario and their understanding of the scenario (everyone makes mistakes sometimes, and many Prolific/MTurk/CloudResearch participants take pride in the quality of their work), and 3) provides a more conservative test ensuring that the manipulation has been read and processed before attending to the dependent measures, rather than afterwards when it is too late. This is a noted deviation, but we feel it is a necessary one.

.7. I would personally like to see on page 26 a footnote saying the minimally acceptable Cronbach's alpha for you to run the study. It is, in my view, important to set up criteria for the internal validity of the estimators.

We appreciate the note, and see the value in setting a criteria in advance for things like reliability for new studies as to reduce flexibility.

However, this case is different. This is a replication, with data collected regardless of the results. Whatever results we have, they are worthwhile communicating, because they repeat the original. Setting an arbitrary threshold not specified in the target would create all kinds of challenges. For example - if we were to set an alpha of .80 but the alpha is .7949 do we simply conclude a failed replication and not proceed to analyze the rest of the data even if analyzing that data shows support for the original effect? Perhaps we should set it to .60? maybe .40? why .40/.60?

This is to say, that in a replication we aim to run the study like the original and as close as possible to their terms and report whatever we get.

To address this, we added the following as a planned discussion for Stage 2:

> [Planned discussion for Stage 2: Based on comment by Dr./Prof. Romain Espinosa, we note that we will discuss issues of scale validity/reliability if scale reliabilities in the replication are below Cronbach's alpha of .60. In such a case we also see value in a discussion of the challenges in how replications should approach issues of scale reliability thresholds.]

.8. I have tried the Qualtrics survey. It seems to work well (I've been randomized to several treatment variations and the order of the two studies was also well randomized). I would only suggest integrating the Prolific ID in the URL such that participants do not have to type it.

Response: Thank you for making the effort to help us double-check the flow of our Qualtrics survey. Yes, that has already been implemented. The Prolific ID was already an embedded field (imported from the URL passed from Prolific to the Qualtrics) and if it were completed by a Prolific user (using a redirect from the Prolific website) then it would have had that field already filled so that they would only need to click "Next".

.9. Misc:

Page 11: Typo: « Studies 3 Study 6 »

Page 12: Number 4 in the hypothesis number in Table 1 missing.

Page 19, Table 3: Are we sure that the age in Prolific starts at 0?

Response: Thank you for catching all of that. Regarding the comment about the age on Prolific: The range reported in Table 3 is based on the simulated data, Prolific only allows those 18 and above, so the actual data should not have any participants below 18. We now set the possible range in the Qualtrics to be 18-99.

Action: Typo on page 11 has been changed to "Studies 3 and 6", and the hypotheses in Table 1 have been appropriately labeled.

Reply to Reviewer #1: Dr./Prof. Thibaut Arpinon

Summary

This Stage-1 is a replication and extension of the study 3 and 6 from Barasch et al. (2014). The authors are planning to fully replicate the results from study 3 and 6, with minor deviations, and extend by adding measures of authenticity, rewards motivation, perceived self-focus, and perceived other-focus.

General comments

I would like to thank the authors for undertaking this very interesting replication project and I applaud the choice of using the Registered Report methodology to do so. Overall, I am impressed by the level of detail. All the deviations from the original paper are thoughtfully mentioned and justified. The proposed extensions are carefully designed and I appreciate the authors' goal to build a unified comprehensive comprehensive design. The justification for the sample-size rationale is also well defined and the power analysis well specified. I have some comments that I believe will further improve the clarity of the analyses to be conducted and will improve the statistical precision of the analyses.

Thank you for the positive and supportive opening note and the constructive feedback.

Major comments

.1. First, I believe that the authors could improve the hypothesis layout in the PCIRR-Study Design. For example, « Emotions towards victims predict moral character, yet more so when they result in prosocial behavior (compared when they do not) » could be divided into two separate hypotheses for clarity. The same goes for the second hypothesis, which could be divided into multiple hypotheses for clarity. This leads to my second comment.

We appreciate the feedback. Thanks to your and the editor's comments we have made comprehensive revisions of the hypotheses and aligned the tables. Please see our reply to the editor above.

> **.2. While reading the stage-1, I found it difficult to distinguish which analyses are part of the confirmatory analysis and which are part of the exploratory analysis. From my understanding, the authors will test the replicability of the hypotheses from the original paper (very well summarised in Table 1). I assume that these will be part of the confirmatory analysis. In addition, the authors will test the inclusion of additional measures (e.g., authentic prosocial motivation), but are these extensions part of the confirmatory analyses or purely exploratory?**
> **For example, on page 14, the authors plan to test the following « targets experiencing high distress in the donate condition will be seen as more authentically motivated than targets experiencing low distress in the donate condition. ». However, I fail to see this included in the PCIRR-Study Design Table. The distinction between confirmatory and exploratory analyses in a Registered Report are important, and I believe that the Stage-1 will gain in clarity and statistical precision if the authors provide more detail.**

Registered Report Stage 1 is meant to be confirmatory, given that the Stage 1 is officially what is pre-registered, and so tests that we specify are meant to be a confirmation of expectations, including our extensions. It is possible that in Stage 2 we will add exploratory tests on top of that, but those will be clearly marked as exploratory, as per the PCIRR guidelines.

We did note one exploratory analysis in our initial submission regarding the mediation analysis. We noted the weaknesses with their mediation analyses that were based on a correlational design with many variables of many contrasts, and these do not seem to be core ideas of the target article. To try and avoid confusion, in our revision we removed this exploratory analysis.

We have now better aligned Table 1 for the hypotheses with the PCIRR Study Design table, and now also clearly mark which hypotheses are for the replication and which are for the extensions.

> **.3. Following my previous comment, I am concerned about the number of hypotheses that the authors will be testing. Replicating and extending findings is good for science to provide more precise effect sizes and a more comprehensive understanding of a phenomenon. However, I am afraid that the extension here will lead to test many hypotheses and increase the likelihood of finding a positive result. If the extensions are part of the confirmatory analysis, I suggest that the authors include a Multiple Hypothesis Testing (MHT) correction procedure. If they are part of the exploratory analysis, I could understand reporting the explorations without but I would still suggest checking the robustness of such findings to a MHT procedure.**

Response: Thank you for the comment. The main aim of our paper is to replicate the findings of the original paper; and the extensions are meant with the hope of additional insights on top of that. Replications are commonly expected to run their analysis in similar terms to those in the target article, and as far as we can tell the target article has not made such adjustments.

However, to address MHT for the extensions, we set our alpha value to .005.

In addition, as can be seen in the Rmarkdown code we provided, for each of the tests that have post-hoc analyses we apply the Tukey adjustment to the p-value. We therefore added the following clarification:

> We note that we applied the Tukey p-values adjustment to all analyses that include post-hoc comparisons.

> **.3. The authors plan to replicate using a Prolific sample. Could the authors please clarify the filters that will be applied on Prolific? Will the sample be representative of the US population?**

We added the following:

> "We used Prolific's filters to restricted the location to the US using "standard sample", we set it to "Nationality: United States", "Country of birth: United States", "Place of most time spent before turning 18: United States", "Minimum Approval Rate: 95, Maximum Approval Rate: 100", "Minimum Submissions: 50, Maximum Submissions: 100000"."

> **.4. This is not a comment per say but more a question to the authors. I see the benefit of writing a result section with simulated data for the authors. However, I do not see the purpose of including this in the stage-1. In my opinion, a stage-1 should include as much elements as possible up until the result this section. The elements should be described clearly and a referee should be able to « visualize » the result section without seeing it. Adding a simulated data result section, especially with data simulated using random generation in Qualtrics, does not add important elements to the stage-1. If anything, it adds additional work the reviewers as they will have to review this section again for the stage-2. Could the author justify why they decided to include this section?**

This is an interesting perspective, and we appreciate you sharing that with us and asking. We do realize that adding a results section based on simulated data is still not a common practice, but we do hope that it will soon become a standard. We see many advantages to this approach.

Stage 1 Registered Reports are meant to ensure that everyone knows and understands what the plan is and to reduce flexibility or document deviations. If referees are left to try and visualize things on their own, the chances of a misalignment of expectations or of misunderstandings regarding the analyses increases significantly. Why would we deny a simple solution that addresses these problems? Also of great value is that it allows reviewers and the editor to help the authors catch oversights or mistakes or suggest additional analyses and approaches to make the planned study even stronger or more informative of the tested research question and hypotheses When everyone is left to guess what the authors meant and visualize their own approach, we are missing the added value of a community aligned towards the best pre-registration for that investigation.

The one possible pushback as to why not to do it that you raised is that it adds additional work for the reviewers. Yet, as with common misconceptions about Registered Reports allegedly adding more work for authors and reviewers, what this ignores is the overall time spent for both Stage 1 and Stage 2, and the costs of possible mistakes and misalignments. Reviewers will need to go over the analyses in Stage 2, but when there is a validated Stage 1 results section written up, the task of going over the results section in Stage 2 becomes fairly straightforward, focused on examining the change between the results based on simulated data and the results based on the real actual data. Instead of reviewers in Stage 2 going back and forth between the Stage 1 pre-registration's vague description of a planned analysis to try and figure out how to compare those, the comparison against the simulated results section becomes rather simple with no guess-work or individual interpretation. Then, you can also more easily distinguish the

confirmatory planned analysis from the adjustments and the exploratory unplanned analyses. Win-win for all.

> **Minor comments:**
>
> **.5. This sentence was confusing to read, please consider rephrasing for clarity. « At the time of writing (January 2024), there were 301 Google Scholar citations of the article and have broadly inspired many important follow-up theoretical and empirical articles, such as on the psychology of (in)effective altruism (Berman et al., 2018; Caviola et al., 2020; Caviola et al., 2021), related to the now influential Effective Altruism movement (MacAskill, 2015). »**

Response: Thank you for the feedback.

Action: We edited the statement for clarity:

> The article has had an impact on scholarly research in the area of moral psychology and altruism. At the time of writing (January 2024), there were 301 Google Scholar citations of the article. Barasch et al. (2014) has inspired many important follow-up theoretical and empirical articles, such as a literature on the psychology of (in)effective altruism (Berman et al., 2018; Caviola et al., 2020; Caviola et al., 2021), which is linked with the now influential Effective Altruism movement (MacAskill, 2015).

> **.6. This fact « Despite its impact, to our knowledge, there are currently no published direct replications of their study. » is already mentioned in text before, please consider deleting.**

Response: Thank you for pointing that out.

Action: Deleted "...and the absence of direct replications" rather than the quoted statement.

Reply to Reviewer #2: Dr./Prof. Angela Sutan

This report is very carefully designed.

.1.
1A. The scientific validity of the research question(s) - OK
1B. The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses) - OK
1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable) - OK
1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses - OK
1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s). - OK
.2. I only have one remark. The authors specifically ask feedback on one improvement:
"We note that when reconstructing the materials we noticed that the conditions in Study 6, at least as described in the target article, were not entirely equivalent, and seemed to conflate expectations and outcome. For example, expectations for reputation seemed to conflate whether the donation was private or public. We categorize this as a possible weakness in the experimental design and decided to deviate and make an adjustment to the target's stimuli to focus solely on manipulation of expectations.
[Note to reviewers: We would appreciate feedback on our assessment of this issue and this adjustment, and are open to changing it given a well-justified argument and/or clear editorial guidelines.]"
However, there is not clear information about the comparision of the stimuli side by side, there is only a description and interpretation, so it's difficult to give any feedback.

Thank you for the support and endorsement.