Reply to PCIRR decision letter reviews #647: Olivola and Shafir (2013) replication and extension

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response to each item. We also provide a summary table of changes. Please note that the editor's and reviewers' comments are in bold with our reply underneath in normal script.

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

A track-changes manuscript is provided with the file: "PCIRR-S1-RNR-Olivola-Shafir-2013-replication-main-manuscript-track-changes.docx" (https://osf.io/bjngy)

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: Clarified the terminology and procedure. R2: Amended references.
Introduction	R1: Added explicit explanation of hypotheses; details of power analysis Rmarkdown R2: Explained matching scheme and the details of study 3; Added Xygalatas et al. (2013) as reference; Clarified Study 3 hypotheses..
Methods	R1: Explained checks; Updated the "Power and sensitivity analyses" subsection; Added attention checks details; Double checked and re-published the survey.
Results	Overhauled the entire section extensively to include details about all tests.

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. Rima-Maria Rahal

> **I have now received two reviews of your submission on a replication project addressing Olivola and Shafir (2013). In line with my own reading of your manuscript, the reviewers highlight important strengths of your outlined approach, but also note some areas for further improvement. In line with these suggestions, I would like to invite you to revise the manuscript.**
>
> **The suggestions focus largely on clarifications needed regarding the power analysis as well as specific statements of how analyses inform the conclusions drawn about the hypotheses, with some additional requests to clarify the terminology and procedure. These issues fall within the normal scope of a Stage 1 evaluation and can be addressed in a comprehensive round of revisions.**

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

Reply to Reviewer #1: Dr./Prof. Vanessa Clemens

> **Thank you for the opportunity to review this stage 1 registered report titled "Do pain and effort increase prosocial contributions?: Revisiting the Martyrdom Effect with a Replication and extensions Registered Report of Olivola and Shafir (2013)".**
>
> **This pre-registered report is a planned replication of a classical work on the Martyrdom Effect from Olivola and Shafir (2013). Specifically, the researchers aim to replicate and extend findings of study 3, 4 and 5 from Olivola and Shafir (2013). For this replication, they will test the hypotheses proposed in the original article and plan to investigate whether the prospect of enduring pain and exerting effort for a prosocial cause promotes charitable giving. They are closely following the target's article design while combining study 4 and study 5's design and collecting data for this combined study and study 3 in one session.**
>
> **Overall, the report is very well-written, easy to follow, and well-structured. It transparently describes the methods, the procedure, and the analytic strategy. Still, I discovered some small inconsistencies and a lack of clarity which I will describe below.**

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

> **1A. The scientific validity of the research question(s).**
> **The researchers aim to replicate and extend findings of Olivola and Shafir's (2013) main studies 3, 4 and 5 on the Martyrdom Effect and test an alternative account to explain findings attributed to the Martyrdom Effect. Specifically, they investigate how the willingness to participate, donate and perceived meaning is impacted by effort and pain involved in fundraising activities and secondly how the cause and fundraising type (effortful vs. easy) interact in impacting the willingness to participate. Based on the original article and the outline described in the registered report the research question is considered to be scientifically justifiable.**
>
> **1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.**
> **In line with the original article, the authors propose and test hypotheses with regards to the Martyrdom Effect. They test whether greater anticipated effort and costs is correlated with greater prosociality (e.g.,**

> **donations / willingness to participate in fundraiser activities) and further investigate a moderating effect of cause. Additionally, they aim to test one hypothesis based on the attribute substitution strategy (Kahneman & Frederick, 2002) as an alternative theoretical account that could explain the empirical findings described as evidence for the Martyrdom Effect. It is appreciated, that the authors reframed the original hypothesis for this part of the project to avoid relying on conventional null hypothesis significance testing. To my understanding, the hypotheses are coherent with the theoretical reasoning and are described clearly.**
>
> **However, I would ask the authors to state more explicitly in the main text of the manuscript that the hypotheses described in the original article are the exact hypotheses tested in this replication. This could for instance be added in the section in which the authors describe the hypothesis from the original article (p. 12).**

Response: Thank you. We appreciate the suggestion and revised accordingly.

Action: We added the following to the "Olivola and Shafir (2013): Hypotheses and findings" section -

> We aimed to stay as faithful as possible to the hypotheses in Olivola and Shafir (2013), and when those were not clearly stated - we provided an approximation based on their designs. In Study 3 the target used a null hypothesis using Null Hypothesis Significance Testing, which we reformulated by adding the alternative hypothesis, and by focusing on the contrast between the two conditions with a combined hypothesis. Similarly, in Study 5 we reformulated the hypotheses to more clearly state the main effects for each condition and then the expected interaction between the two.

> **1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).**
>
> **Overall, the planned study design is a close replication of the original studies with minor deviations (e.g., combining study design of 4 and 5 and having one unified data collection for study 3, 4 and 5). The authors sufficiently describe their reasoning behind these small adjustments. The suggested randomization and counterbalancing and a potential analysis of order effects seems reasonable. It is also appreciated that the authors added Manipulation Checks regarding the importance and the effort of the**

charitable cause which enables a check for internal validity.

I would suggest stating explicitly in section Manipulation Checks (extension) that the authors of the original study did not include any Manipulation Checks.

We originally included this in Table 6, and have now added the following sentence to the "Manipulation checks (extension)" section:

> “We note that this is a deviation as Olivola and Shafir (2013) did not include manipulation checks (see Table 6).”

1C-2. The authors based their sampling plan on a power analysis conducted with GPower. While the authors describe some parameters of their power analysis, it would be helpful if the authors could provide more information regarding the type of statistical test and design they used when running the power analysis to increase transparency.

Thank you, that is valuable feedback.

We revised to be clearer about the statistical tests and all the analyses we did in detail in the supplementary in a section called “Sensitivity analyses”, and also summarized those in the main manuscript in the “Power and sensitivity analyses” of the Method section.

We revised to the following:

> "We summarized the calculated effects for Studies 3 to 5 in Table 2. The smallest effect size for the impact of effort on donations was d = .41 in Study 5, requiring 130 per condition (260 overall). In our unified design in Studies 4 and 5 we had nine conditions, and therefore 1170 overall. We added a buffer of an extra 180 participants (20 participants per condition), therefore aiming for a total sample of 1350 participants (or 150 per condition).
>
> We conducted a series of sensitivity analyses, detailed in the “Sensitivity analyses” subsection of the supplementary materials. We aimed for each condition pair to be powered for the smallest between-subject design t-test contrast (d = .41), yet given that we have three cause effort conditions, two effortful conditions compared to one easy, and two human suffering cause conditions compared to one human enjoyment, the contrast (without the buffer) is 260 effortful compared to 130 easy, or if cause is collapsed then 520 effortful compared to 260 easy in the combination of the human suffering conditions. Our sensitivity analyses show that when collapsing these, we should be powered to detect much weaker effects of d = 0.35 and d = 0.25, respectively. Sensitivity analyses further

show that 130 per condition is powered to detect effects of f = 0.16 for the 2x2 interaction in the replication of Study 5, f = 0.12 for the 3x3 interaction in our extension of the unified design of Studies 4 and 5, w = 0.22 for the chi-square test in the replication of Study 4, w = 0.18 for the chi-square test in the replication of Study 5, and w = 0.14 for the unified design of Studies 4 and 5.

[Note: We will update the sensitivity analyses in Stage 2 to the exact final number of participants after data collection]"

1C-3. The authors state that to detect the smallest effect size for impact of effort on donations, N = 113 participants are needed per condition. They then multiply this number by 9 (= number of conditions) which results in a sample of N = 1017. Finally, they state that to account for possible exclusions of 10% based on their previous experience and their integrated design they will collect data from 1350 participants. The final sample size is therefore roughly 30% higher than the sample size based on the smallest effect of interest from the original study. However, in the section outliers and exclusion (p. 34) the authors explicitly state that they will include all data of those who successfully completed the study (i.e., they specify no exclusion criteria). Therefore, I would like the authors to clarify the required sample size of 1350 participants.

Additionally, the authors mention a "small-telescope" approach in table 6 to explain their power analysis. This approach is not mentioned and explained in the main section of the manuscript. I would suggest to also explain this in the main section of the manuscript.
[Note: The authors mentioned that sometimes their calculations of their effects deviated from the effects mentioned in the original study. Unfortunately, I was not able to identify why these effects deviate.]

Thank you for catching both issues, and for trying to help with the calculations, much appreciated!

Indeed, we do not plan on conducting any exclusions. We removed the mention of a 10% exclusion rate from our sample size calculation, the reference to the "small-telescope" approach from Table 6, yet kept the additional participants as an additional planned buffer, as sometimes the target recruited sample in Prolific does not equal the final sample of those who finish the study when we analyze the data.

We updated the "Power and sensitivity analyses" subsection of the "Method" section:

> Given that in the unified design in Studies 4 and 5 we had nine conditions, 1170 overall. We added a buffer of an extra 180 participants (20 participants per condition), aiming for a total sample of 1350 participants, or 150 per condition.

We will indeed analyze all data collected.

Please also see our reply regarding revisions and clarifications on our sensitivity analyses above.

> **1C-4. The analysis plan proposed by the authors closely follows the approach of the original article. The authors describe which type of statistical analysis will be used to test the respective hypothesis. I would kindly ask the authors to provide more details on the planned statistical analysis e.g., specifying the respective dependent and independent variables and providing information regarding which result would be considered as a finding in line with / or against the proposed hypotheses.**
> **[Note: The authors refer to the RMarkdown for a summary of tests used to test the hypotheses. However, I would kindly ask the authors to already specify the planned test for the hypothesis in the main text of the manuscript (e.g., to test Hx, we conduct a XY with Y as a dependent variable and X as an independent variable).]**

Thank you for the valuable feedback.

We did a major revision to our planned data analysis and results section to be very specific about all the tests that we will be conducting using our Qualtrics simulated data. It should now be clear what tests are conducted on which conditions and to answer what part of the study.

Please see our overhauled and detailed Results section.

> **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.**
>
> **Overall, the authors provide sufficient methodological detail to be reproducible and ensure protection against research bias. The method section is clearly structured (e.g., through the structuring with headlines and the provision of complex information in tables) and accessible to readers.**
>
> **However, as I already outlined below, the authors should provide more information about the parameters of the power analysis and the planned statistical analysis to test their hypotheses. In the section procedure, the**

> **authors describe that "three of four questions also served as attention checks". I would kindly ask the researchers to state which questions were considered as an attention check and specify how these attention checks will be used (e.g., participants then cannot take part in the study if they fail the attention checks?).**

We amended that section to include more information about these questions:

> Participants indicated their consent, with four questions confirming their eligibility, understanding, and agreement with study terms, which they must answer with a "yes" and required responses in order to proceed to the study. Three of the four questions also served as attention checks, with the options order being rotated (yes, no, not sure) indicating confirmation of: (1) paying close attention to details and answering subsequent questions carefully, (2) agreement to having to answer attention and comprehension checks, and (3) being a native English speaker born, raised, and currently located in the US. Failing any of the three attention questions meant that the participants did not indicate consent and therefore could not embark on the study. These were followed by writing a statement indicating that they understand and agree and terms, which participants had to write correctly in order to proceed, with as many attempts as needed. Upon completion of these steps, participants proceeded to begin the survey with Study 3 and Studies 4 and 5 combined, presented in random order.

Therefore, yes, failing those consent questions (and choice rotation attention checks) means that the participants have not indicated consent and can no longer take part and complete the survey and must therefore return the assigned task to Prolific.

> **1D-2. The researchers also provided us with a Qualtrics link to test the survey. Overall, the survey looks very good. However, I noticed that one had to indicate the donation / pricing for all 20 conditions of distance. I would kindly ask the authors to clarify whether this was an issue in the programming of the study or whether the factor distance is a within-subject factor and not a between-subject factor as indicated in the design of the study.**

Thank you very much for catching that. Much appreciated!

This may have been an issue with the programming of the study, perhaps something to do with having to "Publish" changes made (which is what we do just before the pre-registration and the survey is expected goes "live"). Our oversight. This seems to be a public preview issue, as the simulated data provided in our previous submission shows the randomization was for 1 and not 20 conditions.

Indeed, participants should have been assigned to one of the donation conditions, and should not be completing all 20 conditions.

We have now re-published the survey and double checked to confirm that this is indeed the case in the preview of our survey.

> **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).**
>
> **The authors have considered sufficient outcome-neutral conditions by including both positive controls and other quality checks. Namely, they added two manipulation checks for the combined studies 4 and 5 to test whether the manipulations indeed have effects on effort and cause importance. Additionally, the authors added attentiveness checks in the beginning of the study.**
>
> **I hope the authors consider my comments to be helpful in preparing the final version of the registered report.**

Thank you for the positive and constructive feedback.

Reply to Reviewer #2: Dr./Prof. Liesbeth Mann

1. Scientific validity of research question

.1.1. This is an important and relevant proposal. I think the authors provide good reasons for conducting this replication. The research question is clear and valid. As it concerns a replication, it is already clear that this RQ is testable. Additionally, having more certainty about the Martyrdom Effect and whether it indeed exists in different contexts is valuable because it has important practical implications, as the authors also indicate by saying that the findings by Olivola and Shafir (2013) "have been impactful and offered a new perspective on altruistic behavior and charitable giving."

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

.1.2. However, to me the theoretical background and rationale for choosing this specific study is not clear enough yet. The background part of the report should, in my opinion, be written for an audience that is not directly familiar with the studies by Olivola and Shafir (2013). So someone should be able to understand the current theoretical background without having (just) read the paper by Olivola and Shafir (2013). To give an example, the first sentence of the background mentions a "donation matching scheme". What is this exactly? This should be clarified.

We appreciate the feedback and so revised the opening paragraph to explain a bit about matching schemes and provide a relevant citation:

> The Martyrdom Effect is the phenomenon that in charitable giving decisions with a donation matching scheme, individuals are willing to donate more when the donation involves personal effort or sacrifice. In such donation matching schemes, every donation made is matched by a donation of the same amount by another donor, effectively doubling a donor's contribution, aiming to increase the motivation of donors to donate and donate more (e.g., Caviola & Greene, 2023). Olivola and Shafir (2013) demonstrated that individuals were more inclined to donate more in fundraisers (with a donation matching scheme) when the fundraiser activities required effort, such as running long distances or fasting, compared to more leisurely activities like attending a picnic. Their findings have been impactful and offered a new perspective on altruistic behavior and charitable giving.

Actually, regarding donation schemes, Olivola and Shafir (2013) do not mention it, but this is something we just happened to notice when we were working on reconstructing the survey. This is why we had the following in the discussion section:

> [Planned discussion for Stage 2: The target article scenarios involved matching schemes, which seem like an important factor not clearly discussed in the target's introduction of the theory, and it is unclear whether these findings will generalize to similar fundraisers and causes with no such matching scheme. We will discuss matching schemes in relation to the theory and methodology in the target article.]

> **.1.3. The Martyrdom effect itself is clearly explained. However, the third paragraph under "Martyrdom Effect" is unclear to me. What was exactly tested in experiment 3 by Olivola and Shafir (2013) and why?**

Thank you, we revised that section to include more details about the idea, the setup, the findings, and the conclusions drawn:

> Olivola and Shafir (2013) also argued that the Martyrdom Effect is not simply a cue for value, as suggested by attribute substitution strategy (Kahneman & Frederick, 2002). The concept of attribute substitution refers to a cognitive process in which individuals substitute a complex, less accessible attribute, like the value of a donation, with a more readily available and simpler attribute, like the amount of pain or effort. This would suggest that the reason why people donate more in more effortful tasks is because level of effort serves as a proxy for value (importance, impact, etc.). Olivola and Shafir (2013) aimed to contrast the two explanations in their Experiment 3, examining whether it is the mere presence of effort or the extent of the effort involved. They experimentally varied the hypothetical running distances, with one group of participants indicating how much they would donate to run that distance, and another group of participants indicating how much money they would ask for to run that distance (unrelated to charity). They found no support for an association between running distance and donations, yet support for an association between running distance and pricing. They concluded this as evidence of the "Martyrdom Effect" as different from the attribute substitution explanation and other theories arguing for a U-shaped relationship between pain–effort and behavior (e.g., goal setting; Locke & Latham, 2006; Trope & Fishbach, 2000). We note that the target article framed the effect as having no association for donations (null effect) and having an association for pricing, to avoid testing the null with Null Hypothesis Significance Testing, we reframing this to a joint hypothesis of a comparison of the two associations, predicting a stronger association for pricing than for donations.

.1.4. Also the choice to replicate these particular studies (Study 3, 4, and 5) explained under "Choice of study for replication: Olivola and Shafir (2013)" is not yet sufficiently clear to me. Why these studies specifically? This could be justified better, especially the explanation on page 10 is unclear to me, this part seems more like a method section to me, so would it be possible to explain this in more theoretical terminology?

Thank you for the feedback. Apologies but it is not clear what using a "theoretical terminology" would mean here, given that the reasons are technical and straightforward. The aspects tested by Studies 1a/b are covered by the more complex and informative Studies 4 and 5 (with our extensions), and Study 2 was a far more complex and costly investigation that built on the phenomenon tested in Studies 4 and 5 using hypothetical scenarios.

We took this comment to mean that we should try and explain our reasoning better, so we moved the discussion of the studies we chose not to replicate till after we discussed the studies that we do aim to replicate, and then revised aiming to make that reasoning clearer:

> "We focused our investigation on the studies examining the willingness to donate and donation amount in relation to the pain and effort involved in fundraising: Studies 3, 4, and 5.
>
> We chose to combine Studies 4 and 5 into a single unified design for a more direct and comprehensive investigation contrasting the two types of painful-effortful fundraisers used in each of the studies compared to the one easy event that was used by both studies. We also manipulated the three types of causes used separately in the two studies, examining joint impact on both willingness to participate, donation amount, perceived meaningfulness, perceived impact, and manipulation checks of perceived effort and importance.
>
> Study 3 was designed to address an alternative explanation to the Martyrdom Effect, in that effort (running distance) had a weaker association with donation amount compared to price (amount of money asked for participating), indicating that it is not about the level of effort, but rather effort itself. We reframed the hypotheses from a null hypothesis, and simplified and updated the methodology to allow for a clearer comparison between the two conditions, on both item (distance) and participant level.
>
> We chose not to replicate Studies 1a and 1b as their designs were mostly covered by the designs of the more complex and informative Studies 4 and 5 (contrasting effortful versus easy activities) and our extensions (which address baseline preferences when both effortful and easy presented together). We chose not to replicate Study 2 as it involved real money and actual pain, where participants in a public goods game made larger

contributions when doing so was expected to be painful. We felt that it would be better to first successfully revisit the baseline demonstration studies using hypothetical scenarios in Studies 4 and 5, before embarking on the more complex and costly replication of Study 2."

.1.5. By the way, I think this study could also be relevant to mention in the theoretical background as it clearly relates to the current topic: Xygalatas, D., Mitkidis, P., Fischer, R., Reddish, P., Skewes, J., Geertz, A. W., ... & Bulbulia, J. (2013). Extreme rituals promote prosociality. Psychological science, 24(8), 1602-1605.

Smaller point: not all references are included in the literature list.

Thank you for suggesting Xygalatas et al. (2013). We added the following:

Xygalatas et al. (2013) demonstrated a related effect showing that those taking part in high-ordeal rituals made more prosocial donations compared to those who participated in low-ordeal rituals.

2. Logic rationale and plausibility of hypotheses:

Clear hypotheses for Study 4 and 5, but less clear for Study 3, I think this can be formulated more precisely.

Thank you, we reframed to the following:

1 (null)	In fundraisers requiring effort, level of effort is not associated with donation amount (original null hypothesis deduced from target article)
1 (alternative)	In fundraisers requiring effort, level of effort is positively associated with donation amount. (reframed from null hypothesis)
2	In activities requiring effort, level of effort is positively associated with the price people demand to participate in that activity.
1 and 2 combined	The positive association between level effort and price is stronger than the association between level of effort and donation amount.

3. **Soundness and feasibility of methodology and analysis pipeline:**

This is all clear. Because the current study concerns a replication, this is to a large extent similar as in the original article. Procedure, manipulations and data analysis are all explained well. However, I am not sure whether collapsing the three studies in one single design doesn't create unwanted bias (e.g., one study affecting the next) even though they are randomized. It does diverge from the original study so I was happy to see that this is discussed and taken into account under "Order effects". Sample seems large enough to detect small effects so the power of these replications should be good.

4. **Sufficient clarity and methodological detail for replication: Yes, this is all clear.**

Thank you for all your feedback. We are grateful for your time and expertise in reviewing our manuscript.