Reply to PCIRR decision letter #413:
Baron and Szymanska (2011) replication and extension

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

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

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

A track-changes manuscript is provided with the file: "PCIRR-S1-RNR-Baron-Szymanska-2011-replication-extension-mainmanuscript-trackchanges.docx" (https://osf.io/nts9p).

Reply to Editor: Dr./Prof. Romain Espinosa

> **Thank you very much for submitting your Stage-1 manuscript. I have read the paper with close attention and heard back from the reviewers. I really enjoyed reading it, thank you for submitting it.**

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

> **Still, I believe that the paper would benefit from a revision. I put below my comments and you'll find the feedback from the reviewers.**
>
> **First, I believe that the paper would greatly benefit from improving its structure. For instance:**
> **- Please give some numbering to the sections/subsections etc. to facilitate the reading and reviewing**
> **- Also, please make a clear distinction between the introduction, the presentation of the paper you replicate, the hypotheses you aim to test.**
> **- I do not know what « exploratory directions » are. It is uncommon to have a section in the beginning of the manuscript that you leave for Stage 2. If not necessary, I would recommend keeping only the section « exploratory analyses ».**

Thank you. In our revision we added numbering to the different factors, and also numbered the hypotheses, indicated in the headers. We also removed the section “Exploratory Directions”.

> **Besides, I have recently reviewed one RR and recommended another one from Gilad Feldman (co-author of the manuscript). If possible, it would be good to integrate the general remarks that the referees and I made on the other manuscripts. This includes for instance:**
> **- Justification of the p-value for multiple analyses. (Why not use standard methods like Holm-Bonferroni corrections and setting alpha=0.005 instead?)**

Thank you for raising these issues.

Each manuscript has its own challenges, and it is not always clear what the right strategy is, given that we are replicating a target article with well-defined and specific methodology. In some of the other papers we submitted that you mentioned that you reviewed, the need for multiple analyses was a bit more complex, given that we combined several studies into a single unified data collection in a within-subject design (all participants do all studies in random order), and per each of the studies added several extensions aimed to tap into the same phenomenon. This is not the case in this project, as here we are combining all the studies into a single data collection yet are running those in a between-subject design where participants are only doing one of the four studies from the target article (meant not to overwhelm the participants and to avoid repetition).

In each of the studies, there are specific questions that seem to tap into different phenomena, per each of the hypotheses. However, there are instances where several analyses are run from the same study to test a single hypothesis, such as in the case of Hypothesis 3 with 3 different tests in Study 2.

Our previous concern was that by setting to a stricter alpha we would be setting a more stringent threshold for us to conclude a successful replication than the one used by the target article, even if that was a methodological oversight on their part. This has been a topic of debate raised in some of the pushback against replication conclusions in the past. That said, we do understand that when examining so many phenomena using the same methodology with NHST we are setting ourselves to a higher likelihood of "capitalizing on chance", and your requesting this as an editor as part of the peer review process makes it easy for us to implement the adjustment.

As a first step, we updated all of our ggstatsplot to also output Bayesian analyses, as to supplement the NHST reporting. We added the subsection "Supplementary Bayesian reporting" to our methods section:

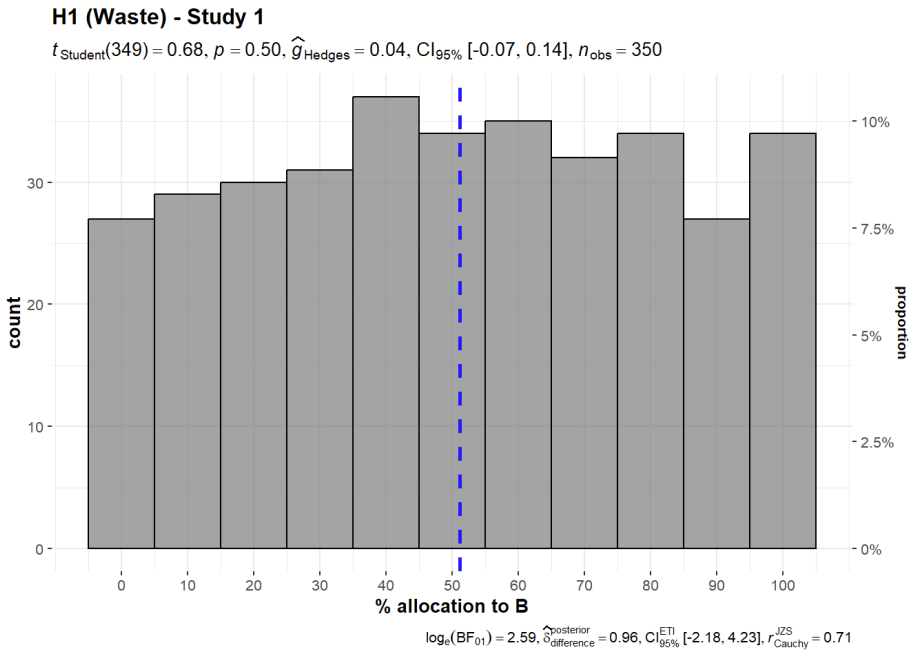
> Our main analysis is using Null Hypothesis Significance Testing, which is focused on rejecting the null hypothesis, yet to address the possibility that we may fail to find support for rejecting the null, we complement all our analyses with Bayesian analyses

> using a prior of BF = 0.707. Bayes factor will be reported in our figures, using the R package ggstatsplot.

We only updated the Bayes analysis in our Rmarkdown code, but have chosen not to replace the figures in the simulated results section. Instead, we added the following note at the beginning of the results section:

> [Note: All ggstatsplot plots in the Rmarkdown code and outputs now also contain Bayes results based. These will be updated to replace the plots before following data collection in Stage 2.]

This is an example of what a ggstatsplot with Bayesian reporting looks like (see bottom right of the image):



As a second step, given that some of the analysis run several analyses for the same research question in the same Study we adjusted our alpha to .005 and added a section explaining our decision:

> The tests for some of the hypotheses involve several analyses on similar dependent variables in the same Study, such as having three analyses in Study 2 to test Hypothesis 3. Following on Editor Dr./Prof. Romain Espinosa suggestion to compensate for multiple analyses, we adjust our target alpha to .005 throughout. We will report raw p-values. For ANOVA analyses we will report Holm corrections for multiple analyses and will report

> both raw and corrected p-values, yet our criteria for signal will use the corrected p-values against the .005 alpha threshold.
>
> We will complement our NHST reporting with Bayesian analyses reporting although we note that our replication success criteria will follow the NHST signal and directionality per the LeBel et al. (2019) criteria.

And added a planned discussion of this point in our discussion section:

> Based on the Dr./Prof. Romain Espinosa's comment on multiple analyses, we will discuss our adjusted alpha threshold with advantages and disadvantages, with references to supplementary Bayes analyses, and suggestions for replication work.

> **- Giving more details in your Study-Design Table about the analysis plan and/or the hypothesis. For instance, the Design Table of Kroll et al. that Gilad Feldman co-authored is very good.**

[Corresponding author note: A clarification, I have not co-authored with Kroll et al. so I was not sure which project this is referring to, but looking at the PCIRR website I am assuming it's referring to this project: https://rr.peercommunityin.org/articles/rec?id=344 and this preprint https://osf.io/6a9gs?view_only=a1fc6796bb92424aad28ff10c11fe595 and the table at the end of that preprint]

We worked to revise our hypotheses to make those clearer and more in line with the methodology.

We provided a very detailed results section with data analysis on a simulated dataset, spelling out each and every analysis, and providing accompanying detailed code showing how we plan to analyze each and every step. The design table is too small and crowded to be able to spell out each and every analysis we are doing. Instead we simply added "(See results section)" to the column "Analysis plan" in the design table.

We also further clarified the interpretation column.

> **- There are many missing elements in the current design table. What is H0 exactly? What is the significance threshold? What would you do if you reject some of the hypotheses and not others? I think that you might replicate some biases but not some others. Given that you indicated a joint interpretation of the results, what will you do?**

Great comment about the possibility of mixed findings. We added the following to our "Evaluation criteria for replication findings":

> We pre-register our overall strategy to conclude a successful replication if at least 80% of the studies (i.e., 4 or 5, out of 5) showed a signal in the same direction as the original study by Baron and Szymanska (2011), a failed replication if only one or no studies (out of 5) showed a signal in the same direction as the original, and any mixed findings with lower than 80% and above 20% (i.e., 2 or 3, out of 5) to be a mixed results replication.

We also updated this in the design table.

Your question regarding H0 has also helped us realize that we may have not made the hypotheses clear enough in our Table 2 and the design table. We have now modified Table 2 to include clear directional hypotheses that include the phrase "People prefer/tend to…", and we adjusted our questions and predictions to more accurately reflect the methods used to test them.

Please see updated Table 2 and PCIRR design table.

> **- Please, discuss the potential risks for floor and ceiling effects. (I believe there is no risk for one-sample t-test but there might be for two-sample t-tests.)**

All of the dependent variables in this project are allocations between two charities (most referring to fixed percent or amounts like all/equally, etc.). We therefore do not see any potential ceiling or floor effects, but it is possible we are misunderstanding your point or overlooking something. We would be happy to revise if given clear editorial guidelines, preferably with an example that indicates the issue, and hopefully with a suggestion for how to address it.

> **Other comments:**
>
> **1) In Roth et al (2015) which is cited, it is written: " [...] Arkes and Blumer (1985: 124) define the sunk-cost effect as "a greater tendency to continue an endeavor once an investment in money, effort, or time has been made." "**
>
> **--> In your experiment, I am not sure that you explore the sunk costs effect. For me, your experiment does not capture how one "continues" investing in a charity based on previous decisions. So, it is not about sunk costs. It might be an issue in the original study (or I might be mistaken here).**

That is a good catch, thank you! We agree. The classic sunk cost effect might actually go counter to the point that Baron and Szymanska (2011) were making, in that people prefer not to invest in charities with "baggage" that already had sunk costs. The "sunk cost effect" tends to also be about one's own escalation of commitment, but here it's about costs incurred by the charity the person is evaluating. We removed the references to sunk costs altogether.

> **2) When reading the paper, I am not fully convinced that this is the best structure. When you introduce the hypotheses, you mention the studies but you haven't presented them so far. So it is a bit misleading for the readers. You might be willing to present the studies first and then describe the associated hypotheses and how they can be tested. (It would be much more natural for me.)**

Thank you, we appreciate the suggestion. We agree. We combined presenting the studies, hypotheses, and empirical tests in one section.

> **3) I have a question regarding your tests. In the appendix (page 4), section about the SESOI, it seems that you're running two-sided tests. I think that you have directional predictions, so I wonder why you run two-sided t-tests.**

This is a good point and a repeating point of debate. The target article had clear predictions but conducted two-tail tests, and there is some sense in that - just in case predictions are not met and effects go in the opposite direction, supporting the idea that people do in fact donate effectively. We feel this is reasonable given that their predictions seemed exploratory and some may seem somewhat counter-intuitive.

In this specific case, the SESOI you are referring to has to do with the power analysis, in which case the two-tail estimations lead to larger sample size which has higher likelihood of detecting the effect (in whichever direction).

> **4) Regarding your assumption checks: you explain that you'll check for normality and homogeneity of variance. Please indicate how you will control for the validity of these assumptions precisely. (How do you define heavy violation?) Please indicate which non-parametric tests you will use in case the assumptions do not hold.**

We note that we will only check for these in case we fail to find support for the replication's hypotheses. We previously wrote: "we will conduct the data analysis again but using non-parametric versions of the same tests"; to clarify exactly which tests we will be using in this scenario, we added the phrase

> (with Wilcoxon tests being used in replacement of one-sample and paired t-tests and the Kruskal–Wallis test in replacement of the one-way ANOVA test)

to the "assumption checks" section.

> **Regarding the reviewers' comments, I would like you to consider them. More specifically:**
>
> **- I agree with Amanda Geiser on the fact that the interpretation of the results should be discussed more, especially if the effect sizes (ES) are smaller than expected.**

We agree with them on how the results could be interpreted and have replied to the point below; in short, the range of possible results is too wide to say anything meaningful as a whole before the results, but we will discuss them in detail once we obtain them. We replied in greater detail below.

> **- I also agree with increasing sample size to ensure that you can capture smaller ES than the original paper reports. For instance, the second reviewer, Jonathan Berman, suggests to multiply the original sample size by 2.5.**

We took the approach of aiming for a SESOI of a weak effect, which covers that. Our current sample size proposal is above the number set out by Simonsohn (2015). We updated the manuscript to reflect this accordingly (per Dr. Berman's suggestion) (in the "Power and SESOI analyses" subsection of the "Method" section):

> "This required a sample size of 327 (and multiplied by 4 = 1308), for a larger total sample size of 1400 participants, accounting for possible exclusions due to incomplete data. Simonsohn (2015) suggests an approach of multiplying the initial sample size by

> 2.5, which would call for a total sample size of 320 × 4 = 800 participants; the total sample size of this replication is larger than this figure."
>
> **- I believe that it might be difficult to implement, but if you can indeed include incentive-compatible donation decisions as Amanda Geiser suggests, it would be indeed excellent. (But it is not necessary for the replication.**

We agree that these would be excellent future directions, yet this goes beyond the scope and scale of what is intended in this project. We added this to the list of planned future directions. Please also see our detailed response below.

> **- Jonathan's comment regarding the diversification effect is particularly interesting.**

We agree that this point seems to be a major limitation to the original study. However, due to replication closeness concerns, we chose not to change the text from the original study to account for this, but we will be adding a discussion section on this point regarding this if no support is found for Hypothesis 3. A more detailed response can be found below.

Reply to Reviewer #1: Dr./Prof. Jonathan Berman

> **I have now read the proposal for the replication and extension of Baron & Szymanska (2011). My view is that these are worthwhile studies to replicate since they test some essential claims in the literature on charitable giving. (I have a strong prior that most of these studies will successfully replicate, though I have been wrong in the past about this).**

Thank you very much for the positive encouraging opening note, and for your time and effort in reviewing our proposal.

> **1) I question the usefulness of the overhead funding extension. First, my understanding is that the manipulation the authors provide suggest that it is actually more cost effective to donate to the charity with the overhead that has been already paid. That is, the donor can save $5 to donate when the overhead has already been paid. This, from a utilitarian perspective, is more efficient, and therefore presents a confound with overhead. The authors could change it such that for B, a contribution of $100 buys the package—and another donor has already covered the $5 overhead costs—but the problem here is that B now has a waste confound which is already covered previously in the proposal. There may be a way around these confounds that the authors may want to think about.**

Excellent feedback. Thank you for raising this point about confounding factors. Much appreciated!

We agree. We changed the phrasing of the overhead funding extension based on your suggestion. Original framing:

> **A** and **B** provide the same basic food and health package for poor children. For **A**, a contribution of $100 buys this package, of which $5 goes to overhead costs. For **B**, a contribution of $95 buys this package; another donor has already covered the $5 overhead cost.

Which has now been changed to:

> **A** and **B** both help thousands of children. Both charities spend 50% of the donations they receive on administrative costs. For each $100 contribution to **A**, $50 will go to helping children and $50 will be used to cover administrative costs. For each $100 contribution to **B**, all $100 will go to helping children; another donor will cover the corresponding $100 administrative cost of this contribution.

This modification makes it such that both:

1. the amount contributed by the donor in each scenario ($100); and
2. the cost-effectiveness of each charity (50%)

are the same for both charities, which means that there should be no confounds for both efficiency and waste. We hope that this addresses your concerns.

We note that the parameters are based on the designs by Gneezy et al. (2014) and Camerer et al. (2018) (e.g., 50% overhead).

> **Second, my understanding of the initial results is that they found an overhead effect in a between-subjects design and this is a within-subjects design. As a result, this makes the overhead differences more salient which likely will cause individuals to alter how they behave relative to the real world. Although such an investigation is informative, it essentially is testing a boundary condition of the previously found effect.**

Yes, we agree, there are differences. We are not claiming a direct replication, but rather a conceptual replication using a different methodology. We can see how switching from a between-subject design to a within-subject design may change things. Studies such as those by Caviola et al. (2014) have shown that the salience of the comparison can affect participant behavior.

If we find support for an effect in our extension, then that would act as evidence in support of the findings of Gneezy et al. (2014) and Camerer et al. (2018) as a form of conceptual replication; however, failure to find an effect can also raise interesting questions related to the possible confounding element of the joint-evaluation condition.

We added a planned discussion of this point for Stage 2 in the Discussion section:

> Following on Dr./Prof. Jonathan Berman's comment we will discuss the differences between Gneezy et al. (2014) and Camerer et al. (2018) and our conceptual replication implementation in our extension, discussing between versus within designs and other elements

> **2) Regarding sample size. I would suggest the authors consider adopting Simonsohn (2015) as a guide to determine a sample size for replication, which suggests simply multiplying the initial sample size by 2.5. This way, if the authors find a null effect they can say that the replication indicates that the initial studies did not have sufficient power to detect an effect. If the authors' current proposal is above that number, then I suggest they note that their sample size is above the number set out by Simonsohn (2015)**
>
> **Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. Psychological science, 26(5), 559-569.**

Thank you for this suggestion. Please see our reply above to the editor on this point.

Our current sample size proposal is above the number one would aim for if following Simonsohn (2015). We added the following to the Methods section:

> Simonsohn (2015) suggests an approach of multiplying the initial sample size by 2.5, which would call for a total sample size of 320 × 4 = 800 participants; the total sample size of this replication is larger than this figure.

> **3) Related to the above, one big caveat of all these studies is that they are all within-subjects designs. People may therefore be reacting in a way they wouldn't do so if this were between-subjects. This ought to be mentioned as a main factor affecting the interpretation of these results.**

Thank you for this suggestion. As this proposal is a replication; we aim to follow the original study's methods as closely as possible; however, this is an important methodological issue, and we thus plan to discuss the possible impact of this design on the results in Stage 2. We added a planned discussion to the Discussion section:

> Planned discussion for Stage 2: Following on Dr./Prof. Jonathan Berman's comment: Studies are within-subject, as per the target article we aim to replicate; Yet, we will discuss advantages and disadvantages of that approach and comparisons between a within and a between subject design. People may react in ways they wouldn't if the experiments were conducted using a between-subject design . We will also note this as directions for future research.

4) Unlike the other studies, Hypotheses 3 (Diversification effect) uses an allocation of 0 as a standard, such that to act in accordance with utilitarianism, individuals should give all their money to one rather than more than one charity. However, it seems that even a few inattentive participants who just indicate randomly could show such an effect. This seems potentially problematic with the original study and problematic here too. Either a different test or a control condition would be appropriate to include.

Yes, this is a good point.

We generally aim to follow the methods of the target article. We agree that this could be considered a limitation of the target article's methodology, and therefore our replication. We added this point to the planned discussion in the Discussion section:

> Planned discussion for Stage 2:Following on Dr./Prof. Jonathan Berman's comment: There seems to be a methodological weakness in the the test of Hypothesis 3, such that few participants may impact results.

Additionally, in the original manuscript, the "equal efficiency" questions under Hypothesis 3 is partly meant to act as a control condition to prevent the number of groups benefiting from each charity from acting as a confounding factor, but we did not make this sufficiently clear to the reader. We edited the manuscript to make this point clearer.

Additional comments and references:

Page 11. There is a typo involving the apostrophe in Qualtrics

Apologies, we tried to find what typo you were referring to, but failed, your comment is too vague for us to be able to identify the issue and correct it.

Reply to Reviewer #2: Dr./Prof. Amanda Geiser

> **Thank you to the authors and editors for the opportunity to review this registered report submission. In it, the authors outline their plan to replicate four studies from Baron & Symanska (2011) on heuristics and biases in charitable giving. Generally, I am thrilled that the authors have undertaken this replication project and am eager to learn what they find. The findings reported in Baron & Symanska (2011) are foundational to our understanding of the factors that impede effective charitable giving. For instance, they are the most frequently cited empirical evidence I have seen for the idea that donors prefer giving to local over foreign charities. Moreover, because this paper was published just before the replication crisis in psychology, and thus before preregistration was standard practice, it will be extremely valuable to learn which of its conclusions hold true in a larger, preregistered replication.**
>
> **Below are my comments. First, I will address the list of considerations recommended by PCI RR for inclusion in Stage 1 reviews. Second, I will outline what I see as some additional areas for improvement. I would be happy to review the Stage 2 proposal when it is submitted.**

Thank you very much for the supportive opening note and for your time and effort in reviewing our proposal.

1. Basic considerations for PCI RR

• Research questions and hypotheses: The research questions and hypotheses for each study are clearly defined. Because this is a replication project and the questions/hypotheses are largely drawn directly from the original paper, they are appropriate for this registered report.
• Materials: The authors state that they will make all of their materials available on OSF, which will make it easy for others to replicate their work.

• Possible interpretations: One area that I believe is lacking is a more thorough consideration of possible interpretations of different outcomes. The studies in Baron & Symanska (2011) are not necessarily the type that would yield interesting results no matter how results turn out. Rather, the results will be much more notable if Baron & Symanska (2011)'s studies successfully replicate than if they do not. If the original effects do not replicate, it could very possibly be because, for example, the effects are smaller than expected.

Thank you for the comment.

We plan to discuss the implications of our research more in-depth once we analyze the results in Stage 2. As we have seven different hypotheses that we are attempting to research, it is hard to say anything concrete about all the possible combinations of the successes and failures (128 combinations!) beyond the obvious until we are sure which combination it is, as most of the hypotheses have different theoretical backgrounds.

However, if you're referring to the overall summary of whether we successfully replicated their findings overall we added the following:

> We pre-register our overall strategy to conclude a successful replication if at least 80% of the studies (i.e., 4 or 5, out of 5) showed a signal in the same direction as the original study by Baron and Szymanska (2011), a failed replication if only one or no studies (out of 5) showed a signal in the same direction as the original, and any mixed findings with lower than 80% and above 20% (i.e., 2 or 3, out of 5) to be a mixed results replication.

> • **Sample size: Based on the authors' power analysis, they concluded that the largest minimum sample size required per study is 178 (and they end up with a planned total sample size of 1,400 across all four studies). I strongly recommend that the authors considerably increase their sample size and do not rely solely on original effect sizes to determine the necessary sample size. Given that the original studies were not preregistered, we do not know whether they involved selective reporting that would inflate the reported effect sizes (see this blog post). Thus, the original effect sizes should not be treated as meaningful benchmarks, and at the very least not the sole benchmarks. If I were conducting this project myself, and ignoring the original effect sizes completely, I would likely recruit closer to 800 participants per study (3,200 total). Given the low time and monetary costs of recruiting participants via CloudResearch, and given the importance of this work, I believe it is worthwhile to err on the side of recruiting a larger sample.**

We understand the concern about the target's effects and the possible benefits in trying for larger samples, but at the end it all comes down to analyzing what's needed and balancing that with reasonable budget concerns.

We agree that a large enough sample size is an important consideration. Our initial power analyses, based on the 178 minimum sample size per study when taking into account the original effect sizes, would mean that we should recruit 800 participants; this is coincidentally also the number of participants that the Simonsohn (2015) benchmark would recommend recruiting to prevent our replication from being underpowered (being 2.5x the original sample size, i.e. 320×2.5=800; see above). However, as we do agree that we should not solely be relying on original effect sizes to determine original effect sizes, especially since there is a clear risk of selective reporting, we further conducted a SESOI analysis that we believe addresses this very concern.

In our revision, we tried to make things clearer in the "Power and SESOI analyses" to explain that we do not rely solely on original effect sizes as a benchmark:

> The original article recruited about 80 participants per study, for a total of approximately 320 participants overall. To prevent the replication from becoming underpowered, we conducted effect size calculations and power analyses based on the information and statistics reported by the target article. Based on the original article's effect sizes, we found that the largest minimum sample size required in one study was 178 (for Hypothesis 5: Study 4, version 3). We ran the four studies separately, with participants

evenly distributed, and therefore multiplied the minimum number of participants by four (178 × 4) resulting in 712.

However, to account for the possibility that the target's effects were an overestimation, for possible exclusion of participants, and to allow for additional analyses, we conducted a SESOI analysis aiming for the ability to detect a Cohen's d of 0.2 (power = 95%, alpha = 0.05) with one-sample and paired samples t-tests, commonly considered weak effects (Xiao et al., 2023). This required a sample size of 327 (and multiplied by 4 = 1308), for a larger total sample size of 1400 participants, accounting for possible exclusions due to incomplete data. Simonsohn (2015) suggests an approach of multiplying the initial sample size by 2.5, which would call for a total sample size of 320 × 4 = 800 participants; the total sample size of this replication is larger than this figure.

- **Ethics: This research appears to pose minimal or no risk to participants and so it falls within ethical norms for the field.**

2. Additional areas for improvement

A. Aims

One area for improvement is clarifying the key aims of the project (aside from replicating the studies from one specific paper). A set of direct replications of Baron & Symanska (2011)'s findings could be worthwhile on its own, but these are not the only biases impeding effective giving. I would recommend either clarifying specifically why the researchers are focusing exclusively on this one paper, or broadening the investigation (e.g., as described above).

B. Scope and generalizability

The scope of the current proposal is quite narrow, both because it focuses on a few simple findings from a single paper and because these findings are not obviously generalizable (based on the stimuli and DVs) to the real-world contexts they are meant to represent.

First, the current framing of the project focuses specifically on four simple studies reported in Baron & Symanska (2011). However, it could be even more valuable to broaden the replication effort to include other foundational findings on impediments to effective altruism. One such finding that comes to mind is Caviola et al. (2014)'s research examining how people value overhead costs and impact in joint versus separate

> **evaluation contexts. They suggest that one reason for overhead aversion is the ease of evaluating proportions (e.g., overhead ratios) as opposed to absolute numbers (e.g., numbers of lives saved). Another such finding is unit asking (Hsee et al., 2013), the finding that donations are more scope-sensitive to the number of recipients when donors start by considering their willingness to donate for one recipient. More generally, Caviola, Schubert, & Greene (2021) provide an excellent review of the obstacles to effective altruism, which could provide a set of findings to replicate (as well as a framework that separates knowledge-based vs. preference-based mistakes). Many of the obstacles discussed in this paper have only limited empirical support and would be useful to either replicate or extend.**

We appreciate the enthusiasm for widening the scope of our study, and we do hope to have the opportunity to follow up with further replications and extensions in the future.

One reason we chose to replicate this paper, as mentioned in the manuscript, is because the design of the studies allowed for the straightforward inclusion of multiple extensions, so in the future we hope to be able to test more hypotheses (i.e. the reputation and overhead extensions) without creating an unnecessarily complex proposal that would be hard for readers and reviewers to parse. For the given study, we prefer the incremental step-by-step approach of validating and then extending, and so we first want to test the baseline methodology with a couple of extensions, before embarking on broader scope and scale.

We also note that some of the follow-up literature is far more complex, and not all questions can be easily answered using the target's methodology. We also see much value in revisiting Hsee et al. (2013) yet that on its own is a large scale project. For example, the "unit asking" study mentioned is completely different from the studies presented in our target in both questionnaire structure (e.g. discrete choices vs. continuous value input) and data analysis (e.g. one-way and paired tests vs. factorial ANOVAs, among other things).

We agree that this is just a first step, and we hope that through this replication and extensions other work will follow. We think that the target studies strike a decent balance between covering several important and potentially impactful effects in the (in)effective altruism literature, while keeping the design and interpretation of the study focused and simple enough.

To clarify this point in our manuscript, we have edited the "choice of target for replication" section such that the second point reads as follows:

> Second, the design of the studies in this paper allowed for the straightforward inclusion of extensions to allow for additional tests and insights on impediments to effective

> altruism. The formatting of the studies' questionnaires lent itself well to the inclusion of selected extensions examining other factors that may preclude the effectiveness of participant choice in donation allocations, which we achieved by inserting items of a similar format to the original studies without heavily increasing the complexity of the replication, thus striking a balance between coverage of different impediments to effective altruism and the complexity of the study and the associated data analyses.

Additionally, we added an item that indicates we also plan to add a section to the "future directions" section further elaborating on this point, mentioning that the studies presented within are not representative of all impediments to effective charity as a whole, while mentioning the aforementioned reasons and using the studies mentioned as examples:

> Planned discussion for Stage 2: Results not representative of all possible impediments to effective altruism, we'll discuss suggestions for future replications and work referring to recommendations made by reviewer Dr./Prof. Amanda Geiser for Hsee et al. (2013) and Caviola, Schubert, & Greene (2021).

> **Second, in addition to the direct and near-direct replications proposed here, I would love to see the findings conceptually replicated using more realistic stimuli (e.g., real charities) and/or incentive-compatible donation decisions as the dependent variable. I realize that this would go beyond the scope of a close/direct replication, but it would benefit the project by showing that these effects (if replicable) emerge in real-world contexts and influence consequential behaviors.**

Thank you very much for this suggestion. We agree that real-life follow-ups are needed, yet we see those to be beyond the scope of the current investigation focusing on first replicating and establishing the lab demonstration of classic effects. We added this item to the list of planned directions for future research:

> Planned discussion for Stage 2: Following Dr./Prof. Amanda Geiser's suggestion, discussing the generalizability of results and expanding on these in future studies in real-life and the field.

A. Methods

For all studies, I would suggest randomizing the direction of the scale endpoints and which charity is A vs. B if this is not already being done. This is particularly important for any analyses that compare participants' responses to the midpoint of a scale. Online participants often respond using more of the right side of the scale, which could bias any results that are contingent on the absolute level of the DV (particularly for the waste/overhead, past/sunk costs, and forced charity findings, which as described in your methods could result merely from a right-side bias if they are ordered as described).

Thank you for the suggestion. We decided to keep the order of the questionnaire answers for several reasons.

Firstly, the target article also had the items ordered from left to right in the same way. We aim to follow the target's methods as closely as possible, and as they did not do that, and so changing that factor may lead to all sorts of unexpected deviations that would make it more difficult to interpret.

Secondly, we feel it important to keep the responding to be as simple as possible as to ensure clarity and elicit quality responding. Participants respond to the same question text multiple times, and so randomizing the scale endpoints may cause unnecessary confusion and potentially result in frustration for participants. In addition, if the order is randomized per question, they would have to see which end of the scale is A and which end is B every time they move onto the next item, which may cause incorrect responses as participants may assume that the endpoints are the same for every question. It is hard to anticipate how participants would respond, as for example some participants may become frustrated enough to start randomly selecting items instead, which runs the risk of adding noise or bias to the data.