Reply to PCIRR R&R 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/poDqckILrDRC

A track-changes manuscript is provided with the file: "PCIRR-S1-RNR2-Baron-Szymanska-2011-replication-extension-mainmanuscript-trackchanges.docx" (https://osf.io/69hgq)

Reply to Editor: Dr./Prof. Romain Espinosa

> **Thank you very much for submitting a revised version of your manuscript. Both referees find that you did a great job in revising the manuscript. Jonathan has no further comment. Amanda gives additional comments that I think can be addressed in a minor revision.**

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

> **Regarding Amanda's first comment, I find the small-telescopes analysis interesting. I do not know whether you can implement it because I am not sure you have effect sizes from the original paper (given in Table 5). If you can implement it, I think it would be worth mentioning that you plan to do this analysis. (Or some equivalence testing?)**
> **Regarding the second comment: I understand your point (overall replication), and I think Amanda is also right about what we can learn for charity giving (which hypotheses hold, which don't). In my understanding, you can discuss that in the discussion section ex-post.**
> **I let you decide what to do with the remaining comments. Amanda's work on the scale point order is exceptionally nice. I am sympathetic to the idea of sticking as close as possible to the paper you seek to replicate (because if the results do not replicate, we do not know why) but her results are interesting (and supportive of your design choice).**

We really appreciate the feedback. The review has challenged us to do better, and we have added analyses and explanations to address the points raised, and replied in detail to all points. We feel that the project is now much stronger as a result and for that we are grateful.

If PCIRR has any special way to thank reviewers, then we feel that Dr./Prof. Amanda Geiser's reviews deserve recognition. This is especially given the data collection and analyses testing our setup to check her assumptions and recommendation, this is truly extraordinary. Her insights are inline with our experience regarding such matters.

We note that in the process of addressing the way to aggregate insights and writing the mini meta-analytic code we realized minor gaps in our planned reporting in the methods section (regarding versions 1-4 in Study 4) and amended the tables and our R code accordingly.

In addition, we added planned reporting of several measures we considered exploratory that were included in the target's data collection but not reported, which we included for the sake of a comprehensive replication and will analyze in our design and will report in the manuscript.

Reply to Reviewer #1: Dr./Prof. Amanda Geiser

> **Thank you to the authors for responding to and addressing many of my initial comments. I appreciate your clarification that this is merely the first step in an incremental process of validating and then extending – this was a helpful framing for me as a reader.**

Thank you very much for the positive and constructive review.

We have never encountered a review quite like this one before. That you ran some of our experiments to try and assess the potential impact of an issue you raised to us seems like really going above and beyond the call of duty. We appreciate that very much and are very grateful.

> **I have reviewed your responses and updated proposal and have a few additional comments for this round:**
>
> **1. Terminology: I would recommend clarifying the terms you use throughout the paper, such as heuristics/biases vs. research questions. Maybe adding the relevant heuristic/bias to each row of the table on page 5 could be helpful too.**

Thank you for the suggestion. We added the names of the effects in the PCIRR study design table.

To clarify that heuristics refer to mental shortcuts, we simply added the word heuristics next to where we refer to mental shortcuts (added section is in parentheses)

> However, people might not donate according to these standards due [to heuristics,] mental shortcuts driven by cognitive constraints aiming to minimize use of cognitive resources, which end up going counter to the intended goal.

In their article, Baron and Szymanska seemed to use “heuristics” and “cognitive biases” interchangeably (p. 215):

> “For example, some donors make decisions that are simply not as well informed as they could be. But an even bigger threat to optimality is posed by overly simplistic decision rules, <u>also referred to as heuristics or cognitive biases</u>. Biases are systematic errors in thinking that are not necessarily a result of misinformation or ignorance, but rather are brought about by an overgeneralization of some decision rule that might be useful in one context but is ill-suited or even harmful when applied in another. Cognitive biases often lead to the systematic misallocations of funds and waste of resources.”

Though many in the literature follow that usage, there are those who make the differentiation and see the cognitive biases as the systematic errors that result from the use of heuristics (mental shortcuts), such as in the seminal paper by Tversky and Kahneman (1974, p. 1130):

> "This article has been concerned with cognitive biases that stem from the reliance on judgmental heuristics."

We therefore made a very slight adjustment to the following paragraph:

> Baron and Szymanska (2011) coined those as "non-utilitarian heuristics", and their research demonstrated five heuristics that result in biases, systematic deviations from the utilitarian model: 1) waste, 2) average cost, 3) diversification, 4) nationalism, and 5) forced charity.

Which we feel better supports that. We would rather not go in more detail into that debate, as to keep the manuscript focused on.

References:

- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124-1131.

> **2. Interpretations and impact (page 5): I appreciate that you added this section, but I have three concerns about it:**
>
> **a. First, you focus primarily on determining thresholds for a successful vs. unsuccessful replication, but you do not discuss what each outcome would mean for the psychology of charitable giving more broadly. I would recommend detailing, for each research question, how you would interpret the results if you do or do not replicate the original effect. As Jonathan Berman recommended in his previous review, if you fail to replicate one of the original findings then you can use Simonsohn's small-telescopes analysis to conclude that the original study was underpowered to detect the true effect in the first place. Then you have a basis to discuss which biases are likely real vs. not real, and how this contributes to what we know about the drivers of prosocial behavior.**

We believe it would be best to be careful and humble about what we might infer about the target book chapter and its findings from what we find. Us failing to find support for one or more of the effects given a well-powered sample does not necessarily inform us anything about whether the target was or was not powered enough to detect an effect in the first place, given the many other

factors at play. They reported finding support for their predictions, across four different studies with very similar designs testing the same hypotheses.

As for comparison of effects, as the editor noted, our analysis of the target article's findings is rather limited, given that their reporting was not complete, with many of the descriptives and stats needed to deduce effect sizes missing, and our planned analyses and expected reporting far more comprehensive and complete. We therefore consider whatever we find to be another more recent and precise data point that would allow us to adjust and update our priors regarding the replicability and robustness of the phenomenon.

In our reply to your point below, we do try and make it more specific what inferences we might draw regarding specific hypotheses.

> **b. Second, I think I disagree with lumping all of the individual research questions and studies into one bucket when discussing what would count as a successful vs. unsuccessful replication. For the purpose of declaring whether the overall replication effort was a success for the original paper, perhaps I agree that a threshold is useful. But I also think it is important to consider each research question (e.g., waste, past costs, overhead) in isolation, because they are quite different. I also find it a bit confusing what you mean when you say "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." My understanding is that there are many different research questions being tested in each study, so I wasn't sure how you would determine if an entire study replicates. Clarification here could be very helpful.**

Great feedback, thank you. It also helped us realize that our reference to "studies" was confusing, especially given that the target article ran four studies, not five, with five main hypotheses, some of which had sub-hypotheses.

In our "at least 80% of the studies (i.e., 4 or 5, out of 5)" we were referring to the five hypotheses, or effects, that we numbered in the manuscript and summarized in the design table. However, yes, you are right, in that each of those hypotheses is actually tested with a number of tests, one of these hypotheses, the diversification hypothesis, actually combined three different sub hypotheses, and we did not make it clear enough how we would conclude whether a hypothesis was supported. This is one of the trickiest parts of a replication of an entire complex article or book chapter that is not often discussed when we talk about replications: How does one conclude whether a replication was successful or not given a replication of an entire article with multiple hypotheses and multiple tests for each of those hypotheses? Thank you for encouraging us to do better in defining those.

We contemplated many different ways to approach this, but at the end, we chose to address this by (mini) meta-analyzing the effects within each hypothesis, first collapsing the within-effects, and then aggregating the different samples in the four between-subject studies. Our conclusion of whether a hypothesis is supported or not would depend on the (mini) meta-analytic summary.

We added the R code to compute that to the OSF (files BS2011-meta.Rmd/.html/.csv), and added the following in the "Evaluation criteria for replication findings" section:

> For each of the five hypotheses there are multiple data sources from different studies, and the diversification hypothesis also has three sub-hypotheses, summarized in Table 5. Each hypothesis was tested in 2-3 studies, and some of the studies with multiple versions. We therefore calculated the (mini) meta-analytic effects for each of the hypotheses, and will conclude support for a hypothesis if the confidence intervals of the effect do not overlap with the null.

> **c. And third, I would recommend fleshing out the theory impact a bit more, related to my first point above. I'd love to see a more detailed discussion of what each outcome would mean for each research question. This project could have several specific theoretical implications beyond showing that people donate inefficiently.**

The implications of these effects have been discussed by the target article and some of the follow-up literature. Our scope for this replication is narrow and already very complex, we wish to empirically test whether we can find support for the phenomenon using the target's methods. We do not wish this to be a new discussion of their hypotheses and literature.

That said, we see the value of a short discussion of what the findings mean and connect that with the literature, though we feel that this is best addressed in Stage 2. We therefore added the following planned discussion:

> Planned discussion for Stage 2: Following Dr./Prof. Amanda Geiser's suggestion, a brief discussion of the implications of the outcomes for each research question/hypothesis.

> **3. Diversification effect: I agree with Jonathan Berman's earlier comment that comparing donations to the less effective charity with a reference point of $0 is likely to result in some participants donating to the less effective charity no matter what, just due to inattention or misunderstanding. The equal-effectiveness control condition that you added is, I believe, insufficient as a comparison.**
> **I would recommend instead (or also) adding a condition where the many-projects charity is more effective (by the same amount as in the condition where the one-project charity is more effective). Then you can compare the amount allocated to the less-effective charity when (1) the less-effective charity is the one with one project vs. (2) the less-effective charity is the one with several projects. If the allocation to less-effective in (2) is greater than in (1), then you can conclude that people have a bias toward diversification.**

We understand the inclination to try and fix the target's methods and analyses to do a better job and go above and beyond what they did. We also see many weaknesses and things that we would ideally like to address, yet we feel it necessary to keep this investigation focused.

As in the previous replies, which you acknowledged in your opening note, we again ask for your understanding in keeping the scope and scale of this investigation focused on the replication. This is already a very complex undertaking, with many potential insights, which we hope would lead to follow-ups. Once we established what works and how, we would be able to further build on those foundations to try and test further.

In our reply to Prof./Dr. Jonathan Berman, in our previous revision, we acknowledged this as a limitation and an added planned discussion of this point with recommendations for future studies that would follow-up.

> **4. Scale point order: In my initial review, I suggested that you randomize the order in which Charity A and Charity B are presented for each scenario. I do understand that you want to stick as closely as possible to the original studies, and ultimately I defer to the authors' judgment.**

Thank you, we understand, and appreciate you deferring to us. Same as in our answer to your points above and in the previous revision, we feel it is best to try and keep things simple, and therefore opt to proceed with the current setup.

> **However, I wanted to discuss a few of my concerns with this choice and explain a bit more why it might matter:**
>
> **a. First, to clarify, I was not suggesting that you randomize the scale points within a given page/set of questions – instead, I was suggesting that you use the same order within a page of questions but merely randomize between-subjects which charity is A vs. B in the scenario itself (which naturally affects which side of the scale refers to which charity). This means there should be no issues with confusing or frustrating participants – both orders would be equally clear.**

This is a good clarification, we appreciate that.

Yet, please consider that adding any between-subject factor has implications for power and required sample size. This specific point is one of many factors that we may choose to vary in a survey, yet these are not the focus of the manuscript, each has implications for power and required sample size, and for complexity (and therefore interpretability and readability) of the manuscript. We feel that these are best left for future studies.

> **b. Second, I'm unsure whether it makes sense in this case to follow the exact methods of the original studies. On one hand, I understand the benefits of using the same scale orientations that the original studies used so that you can directly compare effect sizes between the original and replication studies. On the other hand, I believe that replications are valuable not only for determining if an original effect replicates in a technical sense, but also if the underlying phenomenon truly exists. Methodological choices in the original study that could obscure our understanding of the truth are important to consider and potentially address. For instance: say you noticed a confound in the original stimuli that might explain an effect. If this confound is indeed responsible for the effect, then replicating the original study with identical stimuli might not tell you anything about the existence of the phenomenon (even if it tells you that the effect is obtainable with these exact stimuli). However, I recognize now that you are thinking of this replication project as an initial step to establish the validity of the original effects before you or others build on them, so I'm sympathetic to your desire to keep the survey as-is.**

This is an ongoing debate about replications, and one that feels important to have, but goes beyond this specific replication. You raised one empirical concern, the editor and other reviewers raised others, and we have our own list of issues. Such is the nature of replications.

Our brief response would be that when we revisit classics, sometimes only a decade old, we notice many oversights and methodological concerns, yet at some point we need to make an important decision of what the scope and scale of the replication is. Our (possibly controversial) view based on the many replications we conducted is that doing a comprehensive high-quality replication is already a very complex and costly undertaking, and we do not think it should be up to the replicators to fix the classics, but rather to focus on repeating the classics and pointing out issues for the literature to address. By adding things that replications should do when revisiting the classics, we are making replications difficult to do and publish, trying to uphold standards for replications that the targets never needed to address.

Scale/survey construction issues like the possible need to randomize factors or reverse scales, or something like question framing or presentations modes, are ones best left for specific studies addressing those issues.

Finally, to summarize our view of the issue, we will conclude with the following: If effects, any effects, are so subtle, nuanced, and/or weak as to be reliant, driven by, or moderated by presentation, scale order, or framing, then this means that the effects are less robust and generalizable than claimed to be and we should revisit our general methods in the field. Our impression from our own work and the work of others is that these factors tend to have little, if any, impact on the findings, atleast in decision-making, and therefore adding these as factors, especially in a between-subject design, rarely warrants the investment and costs.

c. Finally, to see for myself whether charity order might impact your results, I collected data from 399 Prolific participants using four of your scenarios: waste/overhead (study 1), diversification with equal efficiency (studies 1-2), nationalism (study 1), and average cost (study 2). I used your exact stimuli and measures, but randomized which charity was A vs. B (and thus allocations in the reverse-order condition were reverse-coded). I found a few marginally significant order effects: For example, in the average cost scenario, allocations to the lower-average-cost charity were lower when it came second (as charity B, original order) rather than first (as charity A, reverse order), t=1.83, p=.069. And in the overhead scenario, allocations to the low-overhead charity were somewhat lower when it came second (as charity B, original order) rather than first (as charity A, reverse order), t=1.69, p=.092. For both of these scenarios, there was about a 5-percentage-point difference in allocations depending on order. In case it is helpful, I will include a table of descriptives below (and would be happy to share the raw data and survey if you'd like). My takeaway is that in

> **contrast to my initial intuition, there was if anything a slight bias towards the left side of the scale; and while the strength/significance of your findings might be affected by order, the direction is unlikely to change.**

Scenario	Order	Mean (SD) allocation (to charity preferred in original, 0-100)	Feeling (% choosing in line with original finding)	Effective (% choosing in line with original finding)	Some good (% choosing in line with original finding)
Diversification (preference for multiple over single)	original	46.58 (24.99)	20%	19%	20%
	reverse	43.25 (24.25)	16%	12%	12%
Nationalism (preference for local over foreign)	original	58.74 (20.55)	28%	20%	20%
	reverse	57.76 (23.42)	29%	22%	22%
Overhead (preference for low over high)	original	68.54 (27.82)	62%	64%	55%
	reverse	73.1 (25.96)	66%	69%	58%
Average cost (preference for low over high)	original	56.28 (25.67)	35%	38%	29%
	reverse	60.95 (25.41)	41%	43%	36%

This is nothing short of remarkable and outstanding. In all our years in academia we have never seen or heard of a review quite like this, and others we've shared this with agree. If PCIRR had a reviewer excellence badge, I would recommend the editor to submit this as a unique case-study.

Thank you very much for collecting the data and running that analysis, we appreciate that very much. Looking at the results, we feel it best not to overthink these findings, given the effects and p-values. To us these provide a demonstration that these factors tend to have little to no impact on the findings and is inline with our vast experience with such issues from similar replications. What you've done here is something we should all aim to learn from, keeping an open mind, thinking about and testing our assumptions. This is what we've been trying to do with our systematic replications of the judgment and decision-making literature.

The very least we feel we can do here to recognize this review and contribution is that if you upload all the materials, data and code to a public OSF folder with a DOI and provide us with a link and citation, then we would be very happy to cite this in the Stage 2 manuscript together with a link to your review.