**Reviewer:** Amanda Geiser (ageiser@berkeley.edu)

**Submission title:** Factors impacting effective altruism: Revisiting heuristics and biases in charity in a replication and extensions of Baron and Szymanska (2011)

**Authors:** Mannix Chan and Gilad Feldman

**Stage 1 Review**

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.

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.
* **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 stronglyrecommend 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](http://datacolada.org/4)). 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.
* **Ethics:** This research appears to pose minimal or no risk to participants and so it falls within ethical norms for the field.

1. **Additional areas for improvement**
2. **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).

1. **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.

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.

1. **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).