**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 2**

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. 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.
2. **Interpretations and impact (page 5):** I appreciate that you added this section, but I have three concerns about it:
	1. 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.
	2. 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.
	3. 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.
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.
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. However, I wanted to discuss a few of my concerns with this choice and explain a bit more why it might matter:
	1. 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.
	2. 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.
	3. 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% |