Numbing or Sensitization? Replications and Extensions of Fetherstonhaugh et al. (1997)'s "Insensitivity to the Value of Human Life"

*Ignazio Ziano
Department of Marketing, Grenoble Ecole de Management,
F-38000 Grenoble (France)
ORCID: 0000-0002-4957-3614
ignazio.ziano@grenoble-em.com

*Qinyu Xiao
Department of Psychology, University of Hong Kong, Hong Kong SAR
xqy1020@connect.hku.hk
ORCID: 0000-0002-9824-9247

*Siu Kit Yeung
Department of Psychology, University of Hong Kong, Hong Kong SAR
u3517520@connect.hku.hk
ORCID: 0000-0002-5835-0981

*Cho Yan Joan Wong, *Mei Yee Sena Cheung,
*Joey Lo, *Melody Yan,
*Gregorius Ivan Narendra, *Kandy Li Wing Kwan, *Rachel Ching Sum Chow, *Chak Yam Man
Department of Psychology, University of Hong Kong, Hong Kong SAR
wcy411@hku.hk, u35390181@hku.hk / nsl19980716@gmail.com,
joey0919@hku.hk/ joeylochungyi@gmail.com, u35382213@hku.hk / yanhcmelody@gmail.com,
u3534477@hku.hk / narendraivan@gmail.com, kandywk@hku.hk / liwingkwan8@gmail.com,
u3537870@connect.hku.hk / rachelchow555@gmail.com, u3544200@connect.hku.hk /
jackyman7151997@gmail.com

^*Gilad Feldman
Department of Psychology, University of Hong Kong, Hong Kong SAR
ORCID: 0000-0003-2812-6599
gfeldman@hku.hk / giladfel@gmail.com

In press at Journal of Experimental Social Psychology

Accepted for publication on August 18, 2021

*Contributed equally, joint first author
^Corresponding author: Gilad Feldman, gfeldman@hku.hk

Author bios

Ignazio Ziano is an assistant professor at the Department of Marketing, Grenoble Ecole de Management. His research focuses on consumer behaviour and judgment and decision-making.

Siu Kit Yeung is a Master of Philosophy student at the University of Hong Kong Department of Psychology. His research focuses on Judgment and Decision-Making, as well as Meta/Open-Science.

Qinyu Xiao is an M.Phil. candidate at the Department of Psychology, University of Hong Kong. His research focuses on moral judgment, choice architecture, and social cognition.

Cho Yan Joan Wong, Mei Yee Sena Cheung, Joey Lo, Melody Yan, Gregorius Ivan Narendra, Kandy Li Wing Kwan, Rachel Ching Sum Chow, and Chak Yam Man were undergraduate students at the University of Hong Kong psychology department during academic year 2019-20.

Gilad Feldman is an assistant professor with the University of Hong Kong psychology department. His research focuses on judgment and decision-making.

Declaration of Conflict of Interest

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial disclosure/funding

The research was supported by the European Association for Social Psychology seedcorn grant awarded to the corresponding author.

Authorship declaration

Cho Yan Joan Wong, Mei Yee Sena Cheung, Joey Lo, Melody Yan, Gregorius Ivan Narendra, Kandy Li Wing Kwan, Rachel Ching Sum Chow, and Chak Yam Man completed pre-registrations and drafted reports for one of the four studies presented here, in groups of two.

Gilad led the replication efforts, supervised each step in the projects, conducted the pre-registrations, and ran data collections. Kit and Qinyu, supervised by Ignazio, followed up on initial work by the other coauthors to verify analyses and conclusions, and performed new ones, and completed the methods and results sections of the manuscript submission draft (Qinyu did Studies 1a and 1b and Kit did Studies 2a and 2b). Ignazio analyzed Study 1c, verified by Qinyu and Kit. Ignazio, Kit, Qinyu, and Gilad jointly designed Study 1c and finalized the manuscript for submission.

Acknowledgments

We thank Ioannis Evangelidis and Bram Duyx for their useful and kind feedback on previous drafts.

Availability of data and material / Code availability

Materials, preregistrations, data, and analyses are available at https://osf.io/786jg/

Corresponding author

Gilad Feldman, Department of Psychology, University of Hong Kong, Hong Kong SAR; gfeldman@hku.hk

Contribution

In the table below, we employ CRediT (Contributor Roles Taxonomy) to identify the contribution and roles played by the contributors in the current replication effort. Please refer to the url (https://www.casrai.org/credit.html) on details and definitions of each of the roles listed below.

Role	Ignazio Ziano	Siu Kit Yeung	Qinyu Xiao	Gilad Feldman	Cho Yan Joan Wong Mei Yee Sena Cheung Joey Lo Melody Yan Gregorius Ivan Narendra Kandy Li Wing Kwan Rachel Ching Sum Chow Chak Yam Man
Conceptualization				X	
Pre-registration				X	X
Data curation				X	
Formal analysis	X	X	X		X
Funding acquisition				X	
Investigation	X	X	X	X	X
Methodology		X		X	X
Pre-registration peer review / verification	X	X	X	X	X
Data analysis peer review / verification	X	X	X		X
Project administration				X	
Resources				X	
Software	X	X	X	X	X
Supervision	X			X	
Validation	X	X	X		
Visualization	X	X	X		
Writing-original draft	X	X	X		
Writing-review and editing	X	X	X	X	

Abstract

Is it better to save 4,500 lives out of 11,000 or 4,500 lives out of 250,000? Fetherstonhaugh et al. (1997) showed that people prefer the former: to save lives if they are a higher proportion of the total, a phenomenon they termed “psychophysical numbing”. We attempted to replicate Studies 1 and 2 of Fetherstonhaugh et al. (1997) (5 data collections, total N = 4799, MTurk and Prolific, USA and UK), and added several extensions (e.g., donation amounts, procedural differences, and individual-level ideology and knowledge). We found mixed support, with two successful replications of Study 2 that indeed showed psychophysical numbing (original: $\eta^2{}_\mathrm{p}$ = 0.55, 90% CI [0.45, 0.62], Study 2a: $\eta^2{}_\mathrm{p}$ = 0.62, 90% CI [0.58, 0.66], Study 2b: $\eta^2{}_\mathrm{p}$ = 0.24, 90% CI [0.21, 0.27], all in same direction), yet also three unsuccessful replications of Study 1 showing instead an opposite psychophysical *sensitization*, a preference for saving a *smaller* proportion of lives (original effect size: $\eta^2{}_\mathrm{p}$ = 0.14, 90% CI [0.02, 0.28], replications: Study 1a: $\eta^2{}_p$ = 0.06, 90% CI [0.02, 0.10], Study 1b: $\eta^2{}_p$ = 0.21, 90% CI [0.17, 0.26]; Study 1c: $\eta^2{}_\mathrm{p}$ = 0.13, 90% CI [0.08, 0.17], all in the opposite direction). We discuss theoretical implications and potential drivers of psychophysical numbing and sensitization, including evaluation mode, comparison procedure, ideology, knowledge, and prioritizing of one’s own country, and practical implications for research on perceptions of charity, aid effectiveness, and donations. Materials, preregistrations, data, and analyses are available at https://osf.io/786jg/.

Keywords: replication, psychophysical numbing, decision-making, life-saving, disaster charitable donations, open science

Numbing or sensitization?

Replications and extensions of Fetherstonhaugh et al. (1997)'s

"Insensitivity to the value of human life"

Is it better to save 4,500 lives out of 11,000 or 4,500 lives out of 250,000? If people care solely about the total amount of lives saved, then they should be indifferent to these two options, yet it is possible that decision-makers who face similar dilemmas on how to best allocate a fixed amount of aid are affected by contextual information such as target population size. Fetherstonhaugh et al. (1997) showed that people exhibit *psychophysical numbing* in evaluating the benefits of life-saving aid. They found that people judge an intervention to be more beneficial if it saves a higher percentage of total lives, even though the absolute number of lives saved is the same. People judged an intervention that saves 4,500 lives out of 11,000 more beneficial than an intervention that saves 4,500 lives out of 250,000 (for instance, for lives of refugees in a camp). Their findings go counter to the possibility of psychophysical *sensitization*, the preference for saving a *smaller* proportion of lives (4,500 lives out of *250,000* rather than out of 11,000).

This work touches on a fundamental aspect of our existence: the value of a human life. If either psychophysical numbing or sensitization exist, there are important psychological, philosophical, and political implications to consider. On the psychological side, it would be clear that the same amount of lives has a different value to people depending on how it is presented. Further, this would mean that even preferences about lives, rather than being fixed and describable by a utility function, are fluctuating and potentially constructed on-the-fly. This poses philosophical questions: how can one act morally in a high-stakes situation – for instance, following a Kantian categorical imperative[1]– if how they perceive the values of others is swayed by its presentation? Further, political responses to genocides and disasters, in terms of military intervention or international aid, may be impeded by psychophysical numbing (Slovic 2007) because the very large, impersonal numbers of lives in danger may not be strong enough to motivate individual citizens and political decision-makers to action.

Present investigation: Five replications and extensions

We conducted pre-registered replications and extensions of Fetherstonhaugh et al. (1997) Study 1 and Study 2, with participants from both the USA and the UK. We successfully replicated Study 2 twice in Studies 2a and 2b, finding evidence in favor of psychophysical numbing. However, we also report three failures to replicate Study 1 (Studies 1a, 1b, and 1c), finding opposite results to those reported in the target article, that is, a preference for sending aid to a larger refugee camp compared to a smaller one. This finding supporting psychophysical sensitization is in direct contrast with the notion of psychophysical numbing, and represents an unusual instance of a failed replication (signals in the opposite direction; following the common LeBel et al. 2019 classification for replication evaluation).

We also tested several theoretically important extensions, to examine moderators and consequences of psychophysical numbing. In Study 1a, we found no evidence that a closer target country (i.e., Haiti rather than Rwanda for USA participants) moderated the results of Study 1. In Study 1b, we found that knowing that other countries are considering the same aid moderated the tendency to prefer aiding the larger camp. In Study 1c, we found that people exhibited psychophysical sensitization even when they had to compare the two interventions

[1] "Act according to the maxim that you would wish all other rational people to follow, as if it were a universal law" (Kant, 1785).

directly (rather than indirectly as in the original paper), and we found no evidence that hypothesized individual-level factors (e.g., political ideology, knowledge of refugee crises, and preference for domestic vs. foreign aid) were associated with the extent of psychophysical sensitization. Lastly, in Studies 2a-2b we found that people reported higher total hypothetical donations if the camp was smaller, suggesting that donation intentions are also influenced by psychophysical numbing.

Psychophysical numbing

What are the roots of psychophysical numbing? Fetherstonhaugh et al. (1997) suggested that Weber-Fechner's laws and the value function are at play. Weber-Fechner laws show that the ability to detect a stimulus quickly decreases as the magnitude of the stimulus increases (Weber, 1834), following a logarithmic function (Fechner, 1860) or a power law (Stevens, 1975). This means that if the initial stimulus is small, a small increase can produce a noticeable change. If the initial stimulus is large, a larger increase is needed in order to produce a comparable, noticeable change. Fetherstonhaugh et al. (1997) reasoned that this applies to number of lives to be saved. In particular, they argued that, since the value function is concave for gains and convex for losses (Prospect Theory; Kahneman & Tversky, 1979), the subjective value of a higher percentage of lives saved – keeping the actual amount of lives saved constant – should be higher. This means that an intervention of reducing the number of casualties from 11,000 to 6,500 may be perceived as more worthwhile than an intervention of reducing the number of casualties from 250,000 to 245,500, despite the same number of lives being saved, as saving 4,500 out of 250,000 may be perceived as just "a drop in the bucket" (Fetherstonhaugh et al., 1997). The Weber-Fechner principles of stimulus detection suggest that when stimulus magnitude is already large as reflected by the large number of lives involved, people may become insensitive to the sheer number of people who are in danger.

Together these suggest that people may prefer saving 4,500 lives out of 11,000 rather than 4,500 out of 250,000, resulting in psychophysical numbing.

Psychophysical sensitization

However, there are arguments for the possibility of an opposite effect - psychophysical sensitization. The first argument is based on the inferences that people make based on different amounts of lives to be saved. People can make surprising inferences about attributes expressed in numerical forms, which may vary with the magnitude of such attributes. This includes numerical measures of response speed (Ziano & Wang, 2021), product attributes expressed in quantities (Evangelidis & van Osselaer, 2017), and the percentage of people that make the same consumer choices (Ziano & Pandelaere, 2018). People may also make inferences about the total number of people in danger in a specific situation. For instance, they may interpret a higher number of lives to be saved as indicating a worse situation overall. While the overall number of people which will be saved stays the same (for instance, 4,500 refugees in camp), people may infer that those in the larger 250,000 people camp are in a worse state and more urgent danger than those in the 11,000 people camp. For instance, those in the larger camp may be perceived as experience worse suffering because the disaster that struck them was of a larger magnitude, or perhaps because necessities such as food and water are scarcer even among those that have access to them, compared with the smaller camp. Therefore, they may wish to direct their help towards the larger camp because they believe that the same number of lives saved are spared a worse fate while they are still alive. Keeping all else equal, this inferential process may also result in psychophysical sensitization.

The second argument for the possibility of psychophysical sensitization is based on miscalculations and the attention-drawing power of large numbers. The reasoning behind a value function explanation of psychophysical numbing follows an assumption that people quickly calculate a percentage by making a division (for 4,500 out of 11,000 lives, it is about

41%; for 4,500 out of 250,000 it is about 2%). However, most people are imprecise when calculating percentages or making divisions, showing predictable biases towards values of 50% when guessing percentages (Landy et al., 2017) or relying on familiar attribute values when making divisions (Ziano & Villanova, 2021). The calculation in this case also involves large numbers, that are hard to understand and cognitively taxing for most people (Landy et al., 2013). It is therefore possible that people do not make these operations and instead focus on the number that seems larger. This may lead to a preference for aid in a situation that involves larger numbers, thereby giving way to psychophysical *sensitization*.

Choice of replication target

We chose to replicate Fetherstonhaugh et al. (1997) because of two factors: its large impact on subsequent research and, to the best of our knowledge, the absence of direct replications. We discuss these two factors in detail below.

Fetherstonhaugh et al. (1997) includes three total studies. We chose to replicate Studies 1 and 2 because they were the most germane to the argument that people exhibit psychophysical numbing when it comes to evaluating aid programs that can save lives, and use very similar scenarios. Study 3, on the other hand, focuses on the evaluation of medical treatments and changed some procedural details.

Fetherstonhaugh et al. (1997) is a highly cited paper, with 444 citations on Google Scholar at the moment of writing (July 2021). In addition, several important seminal articles on the psychology of aid, risk and of number perception came out directly of Fetherstonhaugh et al. (1997). For instance, Friedrich et al. (1999) shows that considerations of psychophysical numbing extend to public policy and consumer domain, and are very hard to debias; Slovic (2007) argues that psychophysical numbing makes it very hard to fully communicate the scale of crimes against humanities such as genocides. Further, many influential articles on cost-effective aid (Caviola, Schubert, Teperman, et al., 2020), and charity and aid structure and communication (Caviola et al., 2014; Gneezy et al., 2014) are directly inspired by the results of Fetherstonhaugh et al. (1997), as are marketing papers focused on the best way to raise money for charitable causes (Sudhir et al., 2016). Psychophysical numbing has also been proposed as the root cause of the identifiable victim effect (Jenni & Loewenstein, 1997), that is, the tendency to offer greater aid to a specific, identified person under hardship than to a larger group with the same problem, another effect with great consequences on how to effectively communicate and advertise charity appeal (Caviola et al., 2014).

In recent years, lower than expected replication rates in several large-scale projects have led to increased calls for scholars to assess the reliability of social science findings (Camerer et al., 2018; Klein et al., 2014; Shah et al., 2019), raising awareness regarding common yet questionable research practices in social sciences (Brodeur et al., 2016, 2020; John et al., 2012). Replication is at the heart of science, fundamental to knowledge accumulation and scientific evidence, and important in examining findings' reliability, robustness, and generalizability (LeBel et al., 2017). No single isolated experiment is sufficient in demonstrating any natural phenomenon. Given that replications should be informed by a cost-benefit analysis (Coles et al., 2018), it makes sense that a high-impact paper such as Fetherstonhaugh et al. (1997) should be replicated.

Direct or Conceptual Replication?

Very close replications (also referred to as "direct replications") strive to remain similar to the original studies in designs and study materials (LeBel et al., 2018). This type of replication has important theoretical value because it attempts to minimize the number of factors (e.g., wording, stimuli, procedure, context) to which deviations from the original results can be attributed in a conceptual replication (Simons, 2014; Zwaan et al., 2017). This allows for the highest chance of re-assessing original results (LeBel et al., 2017). Direct, independent replications have *a priori* unique theoretical value independently of their results, as they assuage potential issues of publication bias (Smaldino & McElreath, 2016; Zwaan et al., 2017). Large majorities of academic psychologists believe that more direct replications should be published and that they are theoretically and practically important (Agnoli et al., 2020). Since we are not aware of any prior direct replications of Fetherstonhaugh et al. (1997), we attempted direct replications of Study 1 and Study 2 of Fetherstonhaugh et al. (1997).

Overview of the Present Research

We conducted five replications in total, three of Study 1 (two on MTurk with U.S. participants and one on Prolific with U.K. participants) and two of Study 2 (one on MTurk and one on Prolific). For each replication, we also tested extensions and explored correlated factors, which we summarize below.

Extensions

Study 1a extension: country closeness.

The literature has suggested that physical distance decreases emotional bond and perceived moral obligation (Trope & Liberman, 2010; Williams & Bargh, 2008). We therefore added a condition with an additional, closer country in need of aid. This served to test whether psychophysical numbing can be moderated by the tendency of U.S. participants to help a country that is geographically closer to them. It is important to test this extension because it would shed light on the psychological roots of psychophysical numbing by investigating whether it is connected to psychological distance.

Study 1b extension: aid from other countries.

Does psychophysical numbing increase or decrease following the news that others are considering similar interventions? The literature offers contradicting predictions. Diffusion of responsibility is the phenomenon for which people are less likely to act prosocially if they are in company (Latané et al., 2006). However, seeing others considering the same options can increase conformity (Asch, 1951; Goldstein et al., 2008). We therefore tested whether knowing that a site in need of aid could also be aided by other parties would decrease the impact of psychophysical numbing. It is important to test this extension because it would shed light on the psychological roots of psychophysical numbing by investigating whether it can be reduced or augmented by social loafing and diffusion of responsibility.

Study 1c, first extension: direct comparison between camps.

Study 1 in the original paper tested psychophysical numbing in an indirect fashion. The original authors used an experimental design in which participants compared foreign aid to a larger or smaller camp with an unemployment or transportation program in their own country. What happens when we directly ask people which is their preference between a program that saves 4,500 lives out of 11,000 and a program that saves 4,500 lives out of 250,000? Answering this question would provide stronger evidence for either psychophysical numbing or psychophysical sensitization.

Study 1c, second extension: political ideology, familiarity, aid importance.

There may be several factors that influence the extent of psychophysical numbing or sensitization. For instance, given that Study 1 involved a trade-off between aid in one's own country and aid to a foreign country, it is possible that U.S.A.-specific political ideology (liberal vs. conservative) may influence the results, as in the U.S.A. (where the original studies were conducted) conservatives seem less likely than liberals to favor international aid, especially after 2016 (Kull, 2017). Further, familiarity with refugee crises, the estimation of the amount of people involved, and the importance one attributes to a specific aid program may also correlate with psychophysical numbing and supply important insights into it.

Study 2a-2b, extensions: donation intention.

People often donate to inefficient charities (Fiennes, 2017). In studies 2a and 2b, we tested whether psychophysical numbing affects donation intentions (Zagefka & James, 2015). This extension helps with assessing whether psychophysical numbing affects donation behavior, and if yes, to what extent, and therefore is an important and meaningful addition to the replication. We also varied the way in which donation was elicited (either as amount given out of $5 in Study 2a or as a percentage of the earnings for task completion in Study 2b) in order to test whether psychophysical numbing can affect donation intentions across different operationalizations.

Participants

The original article recruited "undergraduate volunteers" (*n* = 54) in Study 1, most likely U.S. American, and University of Oregon students (*n* = 162) who were paid $4 in Study 2. In the present research, we recruited US residents from MTurk and UK residents from Prolific. We used MTurk because of the convenience it provides in reaching a large enough sample size in a short time. MTurk participants produce very similar results to U.S. representative samples in experimental psychology (Coppock, 2017; Coppock et al., 2018; Mullinix et al., 2015) and economics (Snowberg & Yariv, 2018). Further, there are several examples of replication of judgment and decision-making and social psychology studies such as Fetherstonhaugh et al. (1997) originally conducted with U.S. American undergraduate students which were successfully replicated with MTurk, even decades later. For instance, overestimation of others' willingness-to-pay (Frederick, 2012) was successfully replicated on MTurk (Jung, Moon, & Nelson, 2019, study 3), and Ziano et al. (2020) successfully replicated the above-average effect shown in Alicke (1985).

In order to increase the generalizability of both the original study and of the present replications, we conducted the replications of both studies also on Prolific, a platform that, similarly to MTurk, allows participants to complete surveys for payment. Prolific participants are mostly UK residents, although participants of other nationalities can also subscribe. Prolific is widely used as a convenience sample for social science, and has a participant base which is demographically similar, but more attentive, compared to MTurk (Kothe & Ling, 2019; Palan & Schitter, 2018). An ongoing mass replication effort successfully replicated a large number of judgment and decision making studies using MTurk and Prolific, with results consistent with student samples and each other (Collaborative Open-science Research, 2020; Chandrashekar et al., 2019; Chen et al., 2020; Ziano et al., 2020). Overall, this supports the notion that both MTurk and Prolific are viable samples for a replication of the impactful findings in Fetherstonhaugh et al. (1997).

Open science, pre-registrations, and disclosures

Materials, preregistrations, data, and analyses are available at https://osf.io/786jg/ . (Pre-registrations: Study 1a: https://osf.io/saqc5/, Study 1b: https://osf.io/ume56/, Study 1c: https://osf.io/ju2fe/; Study 2a: https://osf.io/enc48/, Study 2b: https://osf.io/jtk93/)

All studies, participants, measures, manipulations, and exclusions conducted for this investigation are reported, all inferential tests not explicitly marked "exploratory" were pre-registered with power analyses, and data collection was completed before hypothesis testing. All *t*-tests were two-tailed and α was set at .05.

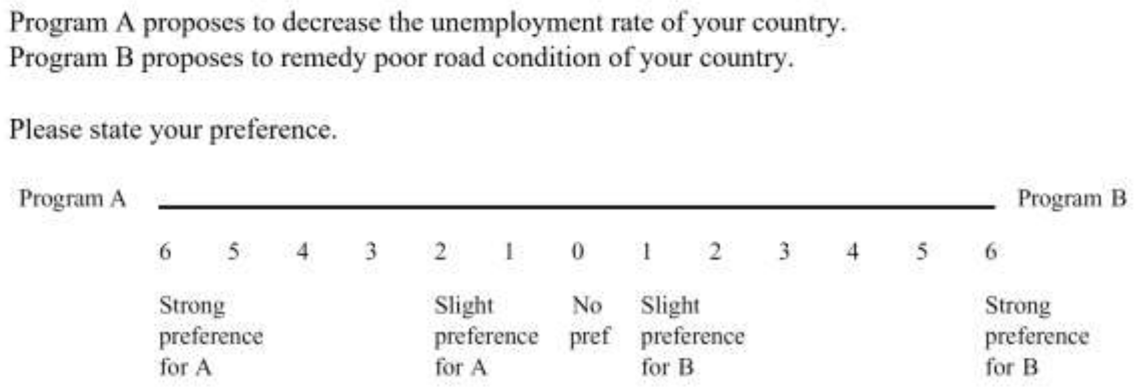
Studies 1a, 1b, and 1c

Studies 1a, 1b, and 1c were three separate direct replications of Study 1 in Fetherstonhaugh et al. (1997). In the original study, 54 undergraduate student volunteers compared four government-funded programs in pairs. These programs were said to cost about the same. Program A and B proposed to decrease local jobless rate (the employment program) and to remedy poor road conditions (the transportation program), respectively, whereas Program C and D proposed to offer clean water to a camp of 250,000 (Program C) or 11,000

(Program D) refugees that suffered from cholera in Rwanda. It was said explicitly, however, that these two Rwanda programs would save the same number of refugees (which was 4,500), despite the camp size difference. The participants evaluated five out of the six possible pairs of these programs, excluding the one that compared C and D directly. On a 13-point scale, they indicated which program they preferred within each pair (a sample question is shown in Figure 1). After they evaluated all five pairs, participants answered a few questions designed to verify whether they knew the same number of lives would be saved by the two Rwanda programs.

Figure 1

A sample question in the comparison task, which compared Program A and B



Program A proposes to decrease the unemployment rate of your country.
Program B proposes to remedy poor road condition of your country.

Please state your preference.

Program A ________________________________ Program B

6	5	4	3	2	1	0	1	2	3	4	5	6
Strong preference for A				Slight preference for A		No pref	Slight preference for B					Strong preference for B

Original results

Preferences for the refugee programs were always coded as positive values (e.g., if participants indicated a relative preference of 3 for a comparison program, the preference for the refugee program would be coded as -3; refer to Figure 1 above). The preference ratings were subject to a two (*domestic comparison program*: A – unemployment - or B - transportation) by two (*camp size*: small or large) repeated-measures ANOVA, which revealed a main effect of camp size, $F(1, 52) = 8.24$, $p < .01$, $\eta^2_p = 0.14$, 90% CI [0.02, 0.28]. The Rwanda program that offered water to a smaller refugee camp ($M = 0.45$) was preferred over its counterpart that offered water to a larger camp (*M* = -0.20) regardless of the domestic program that they were compared with (*d* = 0.40, 95% CI [0.11, 0.68]). One of the verification questions asked the participants whether it was better to save lives in the larger or the smaller refugee camp, and the same ANOVA was conducted only on those who indicated no preferences (*n* = 22). There was also weak support for the effect of camp size, *F*(1, 21) = 3.92, *p* = .06, η^2_p = 0.16, 90% CI [0.00, 0.37]. These participants, who indicated that saving lives in the larger camp was neither worse nor better than saving the same number of lives in the smaller camp, still preferred the smaller camp program (*M* = 0.93) over the larger-camp one (*M* = 0.41; *d* = 0.43, 95% CI [-0.02, 0.87]). As another remarkable piece of evidence for psychophysical numbing, 44% of the participants indicated that they preferred to save lives in the smaller camp, whereas only 14% preferred to save lives in the larger one.

Study 1a – Very Close Replication on MTurk

Methods

Participants.

We recruited 757 U.S. participants from MTurk using CloudResearch/Turkprime (Litman et al., 2017) – the largest number that we could collect given our budget. Participants were paid $1 for this task. The sample after applying the pre-registered exclusions consisted of 386 participants (M_{age} = 38.92, *SD* = 11.37; 178 males, 205 females, and 3 others/preferred not to disclose gender). This sample size (*n* = 386) allowed us to detect the original effect size *d* = 0.40 at over 99.9% power and *d* = 0.14 at 80% power. Our exclusion criteria are detailed in the supplementary, which also reports the full-sample results and a comparison between the samples. We report the results based on the sample after exclusion here.

Replication.

Participants provided their consent at the beginning of the survey and were informed of the general structure of the task. They were instructed to imagine themselves as a governmental official of a small developing country, who was to evaluate *six* government-funded programs that cost about the same (Table 1). As in the original study, the programs proposed to improve local employment rate (Program A), to remedy local road conditions (B), and to provide clean water to save 4,500 lives in either a large camp of 250,000 refugees (C) or a small camp of 11,000 refugees (D) in Rwanda. We classified this replication as "very close replication", with reference to the criteria by LeBel et al. (2018), as most of the factors were very similar compared to the original study.

Extension.

Extending the original study, two additional programs proposed to provide clean water to save the same number of lives as the Rwanda programs, but in refugee camps in Haiti. Program E and F proposed to offer water to refugee camps of the same sizes as in Program C and D, respectively, and would also save 4,500 lives. We added these Haiti programs to test whether U.S. participants would prefer to help a country that is geographically closer to them. Information about all six programs was presented in the beginning (i.e., before they evaluated any programs; Table 1). Though this may raise the concern that more numerical information was given as compared with the original study (2 in the original vs. 4 in this replication), we deemed psychophysical numbing should be robust to this small variation. Participants were given five comprehension questions after the task and program descriptions. They had to answer these questions correctly before they could proceed.

Following the comprehension questions, participants proceeded to compare the programs in pairs. They first compared the original four programs in pairs, i.e., Program A to D. Six pairs could be made from four programs, but the one that pairs the two Rwanda programs was omitted to avoid direct comparison. Hence, each participant made five pairwise comparisons. For each pair of the programs, they indicated which they preferred on a 13-point scale (see Figure 1 above). The order of the five pairs was randomized.

Table 1

Study 1a: The six government-funded programs

Program	Description
A	Program A addresses the employment problem in your country to help decrease the unemployment rate.
B	Program B addresses the transportation problem in your country and proposes to remedy poor road conditions.
C	Program C proposes to provide clean water to save the lives of 4,500 refugees suffering from cholera in Rwanda. It offers water to a camp of 250,000 refugees.
D	Program D proposes to provide clean water to save the lives of 4,500 refugees suffering from cholera in Rwanda. It offers water to a camp of 11,000 refugees.
E	Program E proposes to provide clean water to save the lives of 4,500 refugees suffering from cholera in Haiti. It offers water to a camp of 250,000 refugees.
F	Program F proposes to provide clean water to save the lives of 4,500 refugees suffering from cholera in Haiti. It offers water to a camp of 11,000 refugees.

Note. Participants evaluated these programs in pairs as illustrated in Figure 1. Programs A to D were used in the original study as well as in Study 1b.

After participants finished comparing the five pairs of programs, they answered a comprehension check question asking where the refugee camps were in the programs mentioned up to this point (the correct answer was Rwanda). Responses to this question were used for exclusion. Then, participants compared each of the Haiti programs (i.e., E and F) to the employment and the transportation programs (thus four pairs in total). Again, the pairs came in random orders. Participants then indicated whether they thought the same number of refugees would be saved by the refugee programs (yes or no), which size of a camp should receive funding (regardless of location; large, small, or neither), and which refugee camp location they preferred to fund (regardless of size; Rwanda or Haiti). They also provided brief reasons for each of the latter two questions. They answered a few funneling and demographic questions in the end and were debriefed and paid. Table 2 presents a comparison between the design of this replication and that of the original study.

Table 2

Comparing the designs of the replications and the original study

Phase	Original	Study 1a	Study 1b	Study 1c
Main study	Participants compared Programs A to D in pairs (i.e., A vs. B, A vs. C, A vs. D, B vs. C, B vs. D; C vs. D was omitted) in one of two random orders.	Participants first compared Programs A to D in pairs (like in the original study). The order of pairs was randomized for each participant.	Participants first compared Program A to D in pairs (like in the original study). A vs. B was always the first pair. The order in which A and B served as the comparison program was randomized, so was the order of the two refugee programs. But the two refugee programs were always compared with the same domestic program consecutively.	Participants first compared Programs A to D in pairs (like in the original study). The order of pairs was randomized for each participant.
Extension	N/A	Participants then compared Program A, B, E, and F in pairs (A vs. E, A vs. F, B vs. E, and B vs. F). Because there were four pairs (no more A vs. B), there were 24 possible orders. Programs E and F considered Haiti rather than Rwanda as a target country.	Participants repeated the task after they learnt that other 15 countries were also considering similar refugee programs.	Participants directly compared aid to the smaller and the larger camp, and completed measures of political ideology and familiarity with refugee crises at the moment, an estimation of the number of refugee in five countries, rated the importance of each aid program and were asked to allocate $100 million across them.

Note. The comparisons were done on a 13-point scale as shown in Figure 1.

Results

Replication.

Preferences for the transportation and employment programs were always coded as negative values, whereas those for the refugee programs were coded as positive values. Therefore, higher ratings indicate higher preferences for the refugee programs. Table 3 presents the descriptive statistics of the ratings and Figure 2 depicts the estimated marginal means.

We conducted a three-way repeated-measures ANOVA, taking comparison program (employment vs. transportation), camp size (large vs. small), and location of refugee camps (Rwanda vs. Haiti) as the within-subject factors. We found no support for either the three-way or the two-way interactions. We found support for a main effect of camp size, $F(1, 385) = 22.63$, $p < 0.001$, $\eta^2_p = 0.06$, 90% CI [0.02, 0.10], and for a main effect of comparison program, $F(1, 385) = 70.45$, $p < .001$, $\eta^2_p = 0.16$, 90% CI [0.10, 0.21]. Contrary to the original findings and to the notion of psychophysical numbing, participants preferred to fund a program in a larger refugee camp ($M = 0.25$) than in a smaller refugee program ($M = -0.03$), regardless of which program the refugee program was compared to and the location of the refugee program, $t(385) = 4.76$, $p < 001$, Cohen's $d = 0.24$, 95% CI [0.14, 0.34]. Also, participants preferred the refugee programs less when the comparison program was the employment program than when it was the transportation program, $t(385) = -8.39$, $p < .001$, $d = -0.43$, 95% CI [-0.53, -0.32]. An overall preference for the employment program over the transportation program was also evidenced by the result of a one-sample *t*-test conducted on the preference ratings when the two were compared ($M = 2.26$), $t(385) = 12.16$, $p < .001$, $d = 0.62$, 95% CI [0.51, 0.73].

Extension.

We found no support for a main effect of camp location in the three-way repeated-measures ANOVA, $F(1, 385) = 0.27$, $p = .602$, $\eta^2_p < .001$, 90% CI [0.00, 0.00]. Therefore, we found no indication that participants preferred to fund refugee programs that were geographically closer, in Haiti, compared to refugee programs geographically more distant, in Rwanda.

Table 3

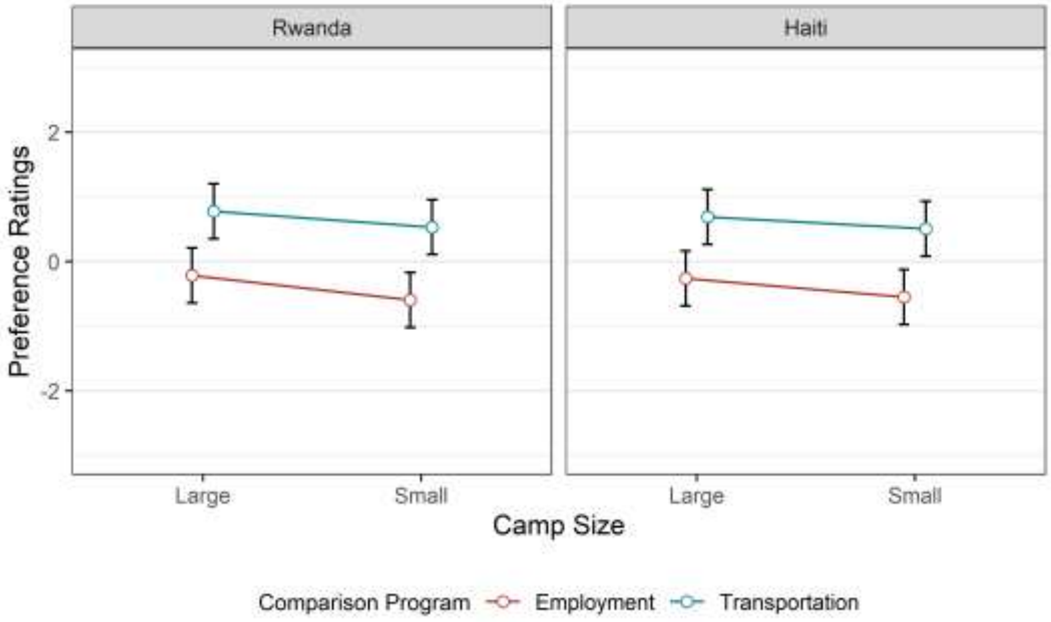
Study 1a: Descriptive statistics

		Refugee camp size			
Refugee program	Comparison	Large		Small	
location	program	Mean	*SD*	Mean	*SD*
Rwanda	Employment	-0.22	4.27	-0.60	4.26
	Transportation	0.78	4.23	0.53	4.14
Haiti	Employment	-0.26	4.36	-0.55	4.27
	Transportation	0.69	4.24	0.51	4.20

Note. Higher values indicate higher preferences for the refugee programs as compared with the employment or transportation programs.

Figure 2

Estimated marginal means of preference ratings in Study 1a. Error bars represent 95% CI of the estimates. Higher values indicate a preference for the refugee aid compared to the domestic program



Subgroup analyses.

As in the original study, we confined our analyses to those who indicated the same number of lives would be saved regardless of whether the refugee program would be carried out in a larger or smaller camp (*n* = 203). Conducting a three-way repeated-measures ANOVA again we only found support for a main effect of camp size, *F*(1, 202) = 13.56, *p* < .001, η^2_p = 0.06, 90% CI [0.02, 0.12], and a main effect of comparison program, *F*(1, 202) = 41.57, *p* < .001, η^2_p = 0.17, 90% CI [0.10, 0.25]. We found no support for other main effects or interactions.

Again, contrary to the original findings and to the notion of psychophysical numbing, people preferred to fund a program in a larger refugee camp than in a smaller one, *t*(202) = 3.68, *p* < .001, *d* = 0.26, 95% CI [0.12, 0.40], and refugee programs were less preferred when compared with the employment program than with the transportation program, *t*(202) = -6.45, *p* < .001, *d* = -0.45, 95% CI [-0.60, -0.31]. We also confined the ANOVA to those who preferred to fund neither the larger size nor the smaller size camp (*n* = 54). In this ANOVA we found no support for main effect or interactions (camp size: *F*(1, 53) = 0.04, *p* = .835, η^2_p < .001; equivalent to *t*(53) = -0.21, *p* = .835, *d* = -0.03, 95% CI [-0.30, 0.24]). As a side note, 279 (72.3%) participants in our study preferred to fund the larger camp whereas 53 (13.7%) preferred the smaller camp.

Full sample results.

The full sample results were generally consistent with the results for the sample after exclusion, and with the conclusion that participants preferred the larger camp than the smaller camp, despite our high exclusion rate. There were two minor differences. First, the camp size main effect was slightly smaller in the full sample, and we found some support for an interaction between camp size and comparison program (*d* = 0.24 for the sample after exclusion vs. *d* = 0.17 and *d* = 0.09 for the full sample, employment program and transportation program respectively). Second, when the ANOVA was confined to those who indicated that they preferred neither the larger camp program nor the smaller camp program, the full sample exhibited a preference for programs in Rwanda over those in Haiti (*t*(119) = 2.09, *p* = .039, *d* = 0.19, 95% CI [0.01, 0.37]), which was contrary to our initial prediction. Yet an effect size this small could not be reliably detected by the small sample size of the sub-group. Overall, our results appear to be robust to exclusions.

Discussion

Replication.

The camp size main effect as revealed in our replication Study 1a was in the opposite direction to the original effect, i.e., an inconsistent and opposite signal in LeBel et al.’s (2019) terminologies (Table 5). Overall, participants preferred to fund the larger refugee camp than the smaller refugee camp, which is an instance of *reverse* psychophysical numbing, or psychophysical *sensitization*. This result held when the analysis was confined to those who explicitly indicated that the same number of lives would be saved by both types of refugee programs. When the analysis was confined to those who indicated that they preferred neither the larger camp nor the smaller camp (n = 54), no evidence was found suggesting that camp size made a difference. Still, 14% and 44% of the participants in the original study preferred to save lives in the larger and the smaller camps, respectively; the corresponding percentages were 72% and 14% in our replication, which were again contrary to the prediction of psychophysical numbing.

Extension.

In addition, we found no evidence for the hypothesized propinquity effect. Participants did not seem to prefer the Haitian programs over the Rwandan programs, despite Haiti being geographically closer to the U.S.

Study 1b – Very Close Replication on Prolific

Methods

Participants.

We initially recruited 750 British participants on Prolific Academic who were paid £0.90 for this task. The sample after exclusion consisted of 723 participants (M_{age} = 40.48, *SD* = 13.09; 299 males, 424 females; refer to the supplementary for details about exclusion). This sample size (n = 723) allowed us to detect d = 0.40 at 99.9+% power and d = 0.10 at 80% power. As in Study 1a, we report results based on the sample after exclusion here and the full sample results in the supplementary. The results did not differ substantially.

Replication.

Participants provided consent and were informed of the general structure of the task in the beginning. As in Study 1a, they imagined that they were a governmental official of a small developing country, who was to evaluate the four programs in the original study (i.e., Program A to D in Table 1). They answered four comprehension questions after the task description. The questions must be answered correctly for them to proceed. The pairwise comparison task on the programs was the same as in Study 1a (see Figure 1). Yet the order in which the pairs were presented was randomized in a slightly different manner compared to the original study. Participants always compared Program A and B, i.e., the employment and the transportation programs, first. They then compared each of these two domestic programs with the Rwanda programs, such that the Rwanda programs were compared with the same domestic program consecutively but in randomized orders. The order in which the domestic programs served as the comparison program was also randomized. For instance, participants might compare the Rwanda programs with the employment program first (either starting with the larger camp program or the smaller camp program) or with the transportation program first (again, they could start with either of the two Rwanda programs). After these five pairwise comparisons, participants answered whether they thought the same number of refugees would be saved by the Rwanda programs (yes or no). They also indicated in which camp it was better to save 4,500 lives: the smaller camp, the larger one, or it was the same. They provided reasons for both questions. We classified the replication as "very close" (LeBel et al. 2018).

Extension.

Following these two questions, as an extension, participants read that 15 other developing countries are also consider funding similar Rwandan camp relief programs. After a comprehension question, which must be answered correctly to proceed, participants reevaluated Programs A to D by comparing them in pairs (A vs. B and C vs. D were left out; hence there were four pairs in total). The order of the pairs was randomized in the same way as in the first round of evaluation. After the second round of evaluation, participants completed a funneling section and answered a few demographic questions. They were debriefed in the end.

Results

Replication.

As in Study 1a, preferences for the domestic programs were coded as negative values and those for the Rwanda programs were coded as positive values. Table 4 presents the descriptive statistics of the preference ratings and Figure 3 depicts the estimated marginal means. We conducted a three-way repeated-measures ANOVA on the ratings, taking comparison program (employment or transportation), camp size (large or small), and round of evaluation (prior to or after hearing that 15 other countries are also considering funding the Rwanda programs) as the within-subjects factors. We found support for two two-way interactions (camp size × comparison program, $F(1, 722) = 10.43$, $p = .001$, $\eta^2_p = 0.01$, 90% CI [0.00, 0.03]; camp size × round of evaluation, $F(1, 722) = 65.28$, $p < .001$, $\eta^2_p = 0.08$, 90%

CI [0.05, 0.12]); we found no support for the remaining one two-way interaction and the three-way interaction .

We first explored the simple main effect of camp size on each level of comparison program (i.e., which domestic program served as the comparison program), collapsing round of evaluation. Our analysis found that, contrary to the original results and to the notion of psychophysical numbing, participants had higher preferences for a clean water program in the larger refugee camp (*M* = 0.35) than in the smaller camp (*M* = -0.38) when these programs were compared with the employment program, *t*(1189) = 13.67, *p* < .001, *d* = 0.40, 95% CI [0.34, 0.46]. The same was true when the refugee programs were compared with the transportation program (*M* = 1.68 for the larger camp program, *M* = 1.13 for the smaller camp program; *t*(1189) = 10.32, *p* < .001), though the effect size estimate was relatively smaller, *d* = 0.30, 95% CI [0.24, 0.36].

Extension.

As we found support for the interaction between camp size and round of evaluation, *F*(1, 722) = 65.28, *p* < .001, η^2_p = 0.08, 90% CI [0.05, 0.12]), we then explored the simple main effect of evaluation round on each level of camp size, collapsing across comparison programs. Our analysis revealed that participants' preferences for the larger camp program decreased after they learnt that other countries were considering similar programs (before: *M* = 1.27; after: *M* = 0.76), *t*(985) = -6.25, *p* < .001, *d* = -0.20, 95% CI [-0.26, -0.14]. However, their preferences for the smaller camp program did seem to change (before: *M* = 0.37; after: *M* = 0.38), *t*(985) = 0.19, *p* = .853, *d* = 0.01, 95% CI [-0.06, 0.07].

Table 4

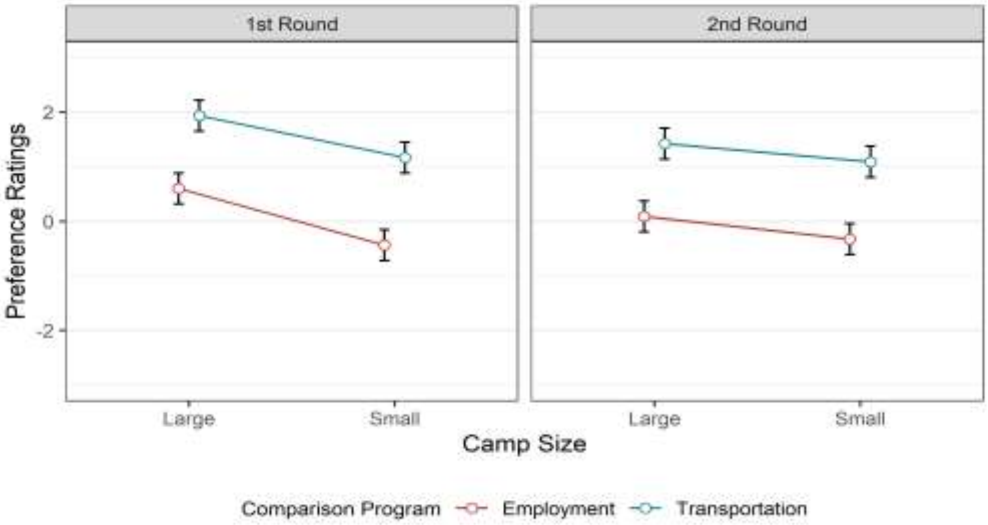
Study 1b: Descriptive statistics

Potential diffusion of responsibility	Comparison program	Refugee camp size			
		Large		Small	
		Mean	SD	Mean	SD
No (i.e., 1st round evaluation)	Employment	0.60	3.99	-0.44	3.91
	Transportation	1.94	3.90	1.17	3.88
Yes (i.e., 2nd round evaluation)	Employment	0.09	4.00	-0.33	3.92
	Transportation	1.43	3.79	1.09	3.74

Note. Higher values indicate higher preferences for the refugee programs as compared with the employment or transportation programs.

Figure 3

Study 1b: Estimated marginal means of preference ratings.



Note. Error bars represent 95% CI of the estimates. Higher values indicate a preference for the refugee aid compared to the domestic program

Subgroup analyses.

We again confined the analysis to those who indicated that it would be the same to save 4,500 lives regardless of the refugee camp size (N = 556). In a three-way repeated-measures ANOVA we again found support for an interaction between comparison program and camp size, $F(1, 555) = 10.91$, $p = .001$, $\eta^2_{\mathrm{p}} = 0.02$, 90% CI [0.005, 0.042], and an interaction between camp size and round of evaluation, $F(1, 555) = 50.11$, $p < .001$, $\eta^2_p =$ 0.08, 90% CI [0.05, 0.12]. Our follow-up analysis on the simple main effect of camp size on each level of comparison program revealed that participants preferred the larger camp program over the smaller camp program regardless of whether these programs were compared to the employment program, $t(914) = 13.18$, $p < .001$, $d = 0.44$, 95% CI [0.37, 0.50], or the transportation program, $t(914) = 9.76$, $p < .001$, $d = 0.32$, 95% CI [0.26, 0.39]. The simple main effect of evaluation round on each level of camp size exhibited a similar pattern as in the analysis that included all participants: after learning that other countries were considering similar programs, participants' preferences for the larger camp program decreased, $t(743) = -5.55$, $p < .001$, $d = -0.20$, 95% CI [-0.28, -0.13], which was not the case for the smaller camp program, $t(743) = -0.09$, $p = .930$, $d = -0.003$, 95% CI [-0.075, 0.069]. As a final note, 104 (14%) participants said that saving lives in the smaller camp was better, whereas 63 (9%) participants opted for the larger camp program.

Discussion

Replication.

While the original results were not successfully replicated, Study 1a results were successfully replicated. The main effect of camp size was in the same direction as in Study 1a but in the opposite direction as in the original study, again failing to provide evidence for psychophysical numbing yet showing support for psychophysical sensitization. In LeBel et al.'s (2019) terminologies, we again found inconsistent and opposite signals (see Table 5).

The interaction between comparison program and camp size was unexpected, as it did not emerge either in Study 1a or in the original study. Although this interaction made the main effect of camp size not directly interpretable, participants consistently preferred the larger camp program regardless of which program the refugee programs were compared with. This remained the case even after we confined our analysis to those who said saving lives in the larger camp and in the smaller camp were the same.

Extension.

We found some evidence for diffusion of responsibility. Participants' preferences for the larger camp program decreased after they learnt that other countries were also considering funding similar programs. However, this was not the case for the smaller camp program, which was not expected.

Study 1c: Very Close Replication and Extension to Direct Comparison, Political Ideology, and Familiarity with Refugee Crises

This study had three objectives. First, we tested whether the results of Studies 1a and 1b would replicate in the context of an ongoing refugee crisis in South Sudan. Second, we extended the studies by asking participants to directly compare the aid programs directed to camps of different sizes. Third, we explored potential factors associated with psychophysical numbing or sensitization.

Methods

Participants.

The final sample after exclusions involved 437 MTurk participants (218 males, 214 females, 2 non-binary, 3 preferred not to disclose their gender; $M_{\text{age}} = 42.43$, $SD = 13.52$) paid $1.00 for this task. This sample size implies over 99% power to detect a Cohen's $d_z = 0.40$

(original effect size) with a two-tailed paired-samples *t*-test and 80% power to detect a Cohen's $d_z = 0.13$ (a small effect; cf. Lovakov & Agadullina, 2021) with the same test.[2]

Procedure.

Participants were presented with a scenario identical to the ones used in Studies 1a and 1b, except that the country in need of aid was changed from Rwanda to South Sudan. We chose South Sudan because the country was home to one of the largest humanitarian crises and numbers of refugees at the time of the study (United Nations High Commissioner for Refugees, 2021) and it was in a region of the world close to Rwanda (East Africa), used in the original study. Considering that this was the only element that changed compared to the original, study, we classify this replication as a close replication. Participants read the following introduction to the scenario:

> Imagine that you are a government official of a small, developing country and you are asked to evaluate four government programs (Program A, B, C, and D). All four programs **cost about the same** and all are being considered for funding. Details of each program are explained in the following:
>
> **Program A** addresses the employment problem in your country to help decrease the unemployment rate, while **Program B** addresses the transportation problem in your country and proposed to remedy poor road conditions.

[2] We received complete responses from 451 participants recruited on the Amazon Mechanical Turk (224 males, 222 females, 2 non-binary, 3 preferred not to disclose their gender; $M_{\text{age}} = 42.33$, $SD = 13.49$), who took part in exchange for $1.00. Participants who took part in the other studies in this paper were barred from participating in this study. No participant failed an attention check at the end of the survey (i.e., replying "yes" to the question "Have you ever been on Jupiter"). Four participants indicated that they were not serious about the survey (< 4 on a 5-point scale) and hence were excluded. These were our pre-registered exclusion criteria. We went further to exclude eight participants who failed a comprehension check question and two participants who indicated suboptimal understanding of the English used in the survey (< 6 on a 7-point scale). These two criteria were not pre-registered for this study but were pre-registered and applied in the other studies in this paper.

On the other hand**, <u>Program C and D</u>** proposes to provide clean water **to save the lives of 4,500 refugees** suffering from cholera in **<u>South Sudan</u>**. The two programs only differ in the size of refugee camps: **<u>Program C</u>** will offer water to a camp of **250,000 refugees** and **<u>Program D</u>** will offer water to a camp of **11,000 refugees**.

You are to evaluate the programs in pairs and state your preferred program.

After participants read this introduction, they were asked to reply to two comprehension checks ("How many government programs are presented for evaluation in this scenario?" [3, 4, or 5]; and "Do all programs cost the same?" [yes or no]), which they had to answer correctly ("4" and "Yes," respectively) to proceed with the rest of the survey.

Then, participants were asked to evaluate the four programs in pairs (five pairs in total, excluding the one that consists of Program C and D, i.e., the two refugee programs) on a 13-point scale (in the example of Program A vs. Program B: −6 [strong preference for A], 0 [no preference], 6 [strong preference for B]; higher scores indicated stronger preferences for refugee programs, if there was one in the pair under evaluation), as in Studies 1a and 1b. The pairs were presented in randomized order.

Extension, direct comparison.

As an extension, participants were asked to also compare the two refugee programs (Program C and Program D) directly, on the same 13-point scale used to compare the other pairs of programs. This comparison was not included in Studies 1a and 1b, nor in the original study. Participants made this comparison only after they evaluated the other five pairs, and therefore the replication part of this study was not be affected by this extension. Participants were also asked whether in their understanding, the two refugee programs would save the same number of refugees (yes-or-no question), and which camp they would like to send aid to (the smaller camp or the larger camp), to which they replied by choosing the smaller or the larger camp in a dichotomous answer. They explained their answer to the latter question in an open-response question.

Extension, ideology and familiarity.

For exploratory purposes, we measured political ideology (with a 7-point Likert item with anchors at 0 = *very liberal*, 3 = *moderate*, 6 = *very conservative*) and participants' familiarity with refugee crises in Afghanistan, Syria, Venezuela, South Sudan, and Rwanda, respectively (7-point Likert items: 0 = *not at all familiar*, 3 = *somewhat familiar*, 6 = *very familiar*). In addition, we asked participants how many refugees they believe there were in the same five countries (each in a text box in which we allowed only numeric inputs). We chose these countries because, according to the UNHCR, they were home to the largest refugee camps in the world at the time of the study (United Nations High Commissioner for Refugees, 2021). We also asked participants how important they believed it was to address unemployment and transportation problems in their home country, and to address the refugee crisis in South Sudan (each on a 7-point item: 0 = *not at all important*, 3 = *somewhat important*, 6 = *very important*), and if they had a $100 million budget, how they would allocate it across the three programs (unemployment and transportation program in their own country and the refugee crisis in South Sudan). Finally, participants completed a funnelling section and reported their demographics at the end of the survey.

Results

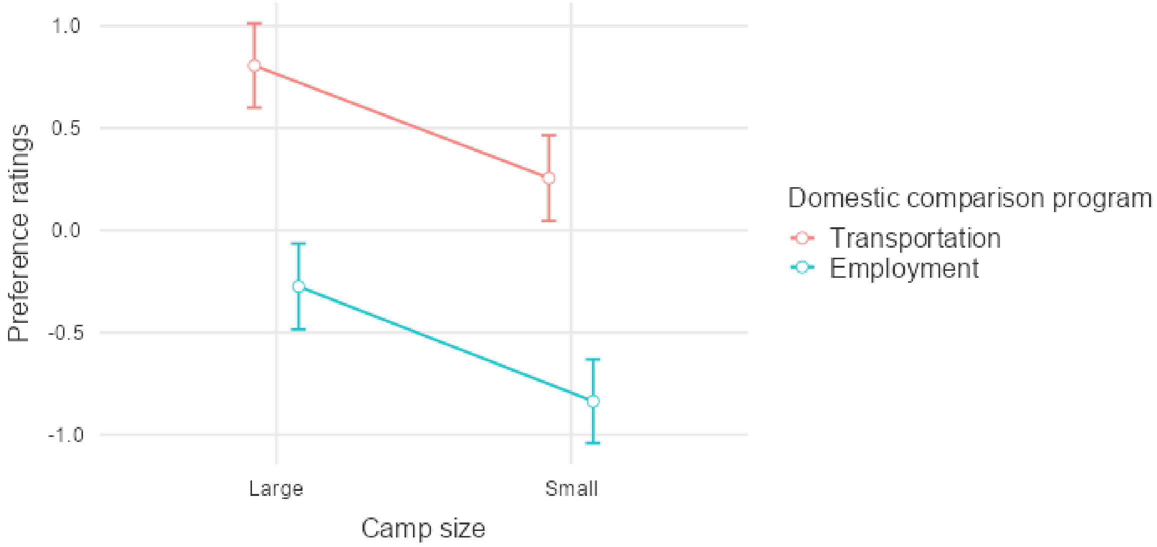
Replication.

A two-way within-subjects ANOVA on preference ratings with domestic program (transportation or unemployment) and camp size (small or large) as factors revealed a large main effect of domestic program for comparison, $F(1, 436) = 60.30$, $p < .001$, $\eta^2_p = 0.12$, 90% CI [0.08, 0.17], such that refugee programs were preferred more when compared with transportation programs (vs. employment programs); a large main effect of camp size, $F(1, 436) = 42.44$, $p < .001$, $\eta^2_p = 0.09$, 90% CI [0.05, 0.13], indicating a stronger preference for the program providing aid to the larger refugee camp. We found no evidence for a two-way interaction between camp size and domestic program, $F(1, 436) = 0.07$, $p = .79$, $\eta^2_p < 0.001$. 90% CI [0.00, 0.00]. Overall, participants preferred to save 4,500 lives out of 250,000 (i.e., provide aid to the larger refugee camp; $M = 0.28$, $SE = 0.20$) than 4,500 lives out of 11,000 (i.e., provide aid to the smaller refugee camp; $M = -0.26$, $SE = 0.20$). These results replicated those in Studies 1a and 1b, corroborating the existence of psychophysical sensitization in this paradigm.

The same pattern of results emerged when we limited the analysis to those who explicitly indicated that the same number of refugees would be saved by the two aid programs ($n = 309$). We conducted an ANOVA and found evidence for a main effect of domestic program for comparison, $F(1, 308) = 48.52$, $p < .001$, $\eta^2_p = 0.14$, 90% CI [0.08, 0.20], a main effect of camp size, $F(1, 308) = 20.63$, $p < .001$, $\eta^2_p = 0.06$, 90% CI [0.03, 0.11], and no evidence for a two-way interaction between camp size and domestic program, $F(1, 308) = 0.70$, $p = .405$, $\eta^2_p = 0.002$, 90% CI [0.00, 0.02]. Again, consistent with psychophysical sensitization, participants preferred to save 4,500 lives out of 250,000 ($M = 0.33$, $SE = 0.23$) than 4,500 lives out of 11,000 ($M = -0.08$, $SE = 0.23$).

Figure 5

Study 1c: Estimated marginal means of preference ratings.



Note. Error bars represent one standard error around the mean. Higher values indicate a preference for the refugee aid program compared to the domestic program

Table 5

Study 1c: Descriptive statistics

Comparison program	Refugee camp size	
	Large	Small
	M (*SD*)	*M* (*SD*)
Employment	-0.25 (4.46)	-0.81 (4.33)
Transportation	0.81 (4.35)	0.29 (4.42)

Note. Higher values indicate higher preferences for the refugee programs as compared with the employment or transportation programs.

Extension: Direct comparison.

To test psychophysical numbing in a more direct fashion, we asked participants for their preferences between the two refugee programs (providing aid to save 4,500 lives in a camp of 11,000 vs. in a camp of 250,000) on the same 13-point scale that they used for evaluating the other program pairs. A higher rating indicated a stronger preference towards aiding the smaller camp. A one-sample *t*-test against the no-preference midpoint of 0 showed evidence for a large effect in favor of the larger camp (*M* = −3.13, *SD* = 3.28), *t*(436) = −19.93, *p* < .001, *d* = −0.95, 95% CI [−1.07, −0.84]. Limiting the analysis to those who explicitly indicated that the two programs would save the same number of people yielded very similar results: *M* = −2.81, *SD* = 3.27, *t*(308) = −15.10, *p* < .001, *d* = −0.86, 95% CI [−0.99, −0.73]. Therefore, this extension provides support for psychophysical sensitization but not for psychophysical numbing.

Apart from indicating relative preferences, participants also made a choice between these two refugee programs. In total, 372 out of 437 (85%) participants thought that the program providing aid to the larger refugee camp should be funded, z = 14.64, *p* < .001 (binomial test against 50%). Similar results were found when the analysis was limited to those who indicated that the same number of people would be saved by the two programs (247 out of 309, or 80%, chose to fund the larger refugee camp, z = 10.47, $p < .001$).[3]

Then, we proceeded to explore whether psychophysical numbing/sensitization is associated with a number of individual differences. These tests were not preregistered, although in the preregistration we did mention that we were going to explore the correlation between these factors and the extent of psychophysical numbing. For reasons of brevity, the full analyses are reported in the Supplementary Materials.

Extension: Ideology and aid importance.

We explored the correlations between the one-item relative preference measure for the two refugee programs and self-reported political ideology, several measures of importance of addressing the different causes (i.e., unemployment, transportation, and refugee crisis in South Sudan), and the imaginary budgets allocated to each of these causes out of $100 million. We could only find evidence in favor of the correlation between relative preference and the budget allocated to addressing unemployment problem, $r(435) = -.10$, 95% CI [-.004, -.19], $p = .042$; we could not find any evidence in support of an association with the other measures and the relative preference, and all effects were very small (all Pearson's *r*s < . 08; cf. Lovakov & Agadullina, 2021).

Extension: Familiarity and refugee number estimation.

We also explored whether the relative preference between the two refugee programs was associated with familiarity with refugee crises in Afghanistan, Syria, Venezuela, South Sudan, and Rwanda. Overall, participants reported that they were not very familiar with any of the refugee crises (with perhaps the exception of Syria, $M = 2.65$; *M*s = 1.44~1.77 for the other four countries; 3 meant "somewhat familiar"). We could not find support for the

[3] These tests were not preregistered.

hypothesis that familiarity with refugee crises in these countries were correlated with relative preference between the two refugee programs (lowest p = .15; largest r = −.07). Similarly, we found no evidence in support of the notion that estimation of the number of refugees was correlated with the relative preference between the two refugee programs (lowest p = .19; largest r = −.06). Since the distributions of estimated refugee numbers were positively skewed (skewness ranged from 5.81 to 20.90), we examined the same correlations after excluding estimates of refugee numbers that were over the 80th percentile (per estimates of each country). We did not pre-register these analyses, and the cut-off was determined in a very conservative fashion, with the objective of eliminating many of the most extreme responses. After this exclusion, all correlations were small (largest r = −.11, 95% CI [-0.005, - 0.21]) and we could only find support for one correlation (Rwanda: p = .041).

Discussion

We successfully replicated our results in Studies 1a and 1b in this study with a different refugee context but failed to replicate the results of the original study. Again, we found evidence for psychophysical sensitization but not for numbing. With an extension, we directly probed participants' relative preferences between the two refugee programs, and our evidence strongly supports psychophysical sensitization. Further, the exploratory measures that we collected allowed us to rule out a number of alternative explanations. We found no support that political ideology, knowledge about refugee crises, estimation of refugee numbers, and perceived importance of aid programs correlated with participants' relative preference between the two refugee programs, with very small effects.

Studies 2a and 2b

In their Study 2, Fetherstonhaugh et al. (1997) extended the investigation of psychophysical numbing in several respects, using a scenario in which aid was going to be supplied via plane. In addition to the camp size manipulation investigated in Study 1 (either 250,000 refugees or 11,000 refugees), in Study 2 they also varied the amount of prior aid (camps had either received lots of assistance already or received limited assistance), as well as water-purifying plane reliability (either 100%, completely reliable or 60%, limited reliability), in a fully within-subjects experimental design. Fetherstonhaugh et al. (1997) asked participants to evaluate the benefits of sending planes and asked them whether they would choose sending the planes or not. As hypothesized, Fetherstonhaugh et al. (1997) found that people generally preferred programs which would send benefits that could save 1,500 people in a camp with 11,000 refugees rather than benefits that could save 1,500 people out 250,000 refugees. Further, they found that participants also rated programs that almost satisfied the needs as more beneficial than programs that are further away from satisficing the needs, and participants preferred programs with perfectly reliable planes.

Original Results.

Fetherstonhaugh et al. (1997) conducted a 2 × 2 × 2 full within-subjects ANOVA, with camp size, prior aid, and plane reliability as factors. They found support for a main effect of camp size, $F(1, 132) = 160.50$, p <.001, $\eta^2_p = 0.55$, 90% CI [0.45, 0.62]. Respondents believed sending planes to the small camp size, helping 4,500 out of 11,000 people (M = 6.46) was more beneficial than sending planes to the large camp, helping 4,500 out of 250,000 people (M = 4.54). Further, Fetherstonhaugh et al. (1997) found support for a main effect of prior aid, $F(1, 132) = 15.35$, p <.001. Participants perceived that sending planes to those camps which were already satisfied a certain portion of water need (M = 5.73) as more beneficial than those only satisfied a small portion of water need (*M* = 5.27). They also found support for a similar main effect for plane reliability, *F*(1, 132) = 12.01, *p* < .001, in which respondents tended to think that sending 100%-reliable planes (*M* = 5.67) was more beneficial than sending 60%-reliable ones (*M* = 5.33). The findings supported all three hypotheses on benefit ratings.

Participants also indicated whether they supported the binary decision of sending a water-purifying plane. Fetherstonhaugh et al. (1997) found support for a main effect of camp size, *F*(1, 130) = 105.40, *p* < .001, in which participants chose sending planes to smaller camps more often (93%) than larger camps (59%). They also found support for a main effect of plane reliability, *F*(1, 130) = 4.61, *p* = .034, in which participants chose sending plane more often when the plane was 100% reliable (78%), than when it was 60% reliable (74%). However, they failed to find support for a main effect of prior aid, *F*(1, 130) = 0.47, *p* = .50, as the percentage difference in sending planes decisions was minimal (lower prior aid: 75%, higher prior aid: 77%).

We provided a summary of the original findings in Supplementary Materials. See Table 6 for a comparison between original Study 2 and the Study 2 replications.

Table 6

Comparison between Original Study 2 and our Study 2 Replications and Extensions

	Original Study 2	Study 2a	Study 2b
Manipulation	Participants underwent all 8 conditions, which vary in camp size, prior aid, and plane reliability, all within-subject factors.	We randomized participants into two groups (reliability: 100% vs. 60%). Each participant underwent 4 of 8 conditions, which vary in camp size, prior aid, as within-subject factors, and plane reliability, as the between-subject factor See Supplementary Table 20.	We randomized participant into 1 of the 8 conditions, which vary of camp size, prior aid, and plane reliability, all between-subject factors. See Supplementary Table 20.
Replication Dependent Variables	Benefit Ratings Yes/No Decision of Plane Sending	Benefit Ratings Yes/No Decision of Plane Sending See Supplementary Table 17.	Benefit Ratings Yes/No Decision of Plane Sending See Supplementary Table 18.
Extension Dependent Variables	N/A	Hypothetical donation out of $5 See Supplementary Table 17.	Hypothetical donation as percentage of earnings See Supplementary Table 18.

Study 2a – Very Close Replication on MTurk

Method

Participants.

We recruited 821 participants from Amazon Mechanical Turk through TurkPrime/CloudResearch (Litman et al., 2017). We excluded 322 participants based on the pre-registered exclusion criteria detailed in the supplementary materials, leading to a final sample of 499 participants (M_{age} = 38.99, SD_{age} = 12.09; 258 males, 238 females, 3 other). This sample size can detect d = 0.37, the weakest statistically significant effect in the original article, with 99.9+% power, and *d* = 0.12, the weakest not statistically significant effect in the original article, with 76% power. Participants were paid $1 for this task.

Replication.

Firstly, participants read a scenario regarding the Rwandan refugee crisis. We followed a 2 X 2 X 2 mixed-design, in which participants underwent conditions that varied in two within-subject variables: camp sizes (11,000 and 250,000), amounts of prior water aid (low and high), and one between-subject variable – plane reliability (60% and 100%). In the original study, all conditions were within-subjects. We reminded them the same absolute number of lives (1,500) would be saved in each scenario regardless of camp size. The order of display of conditions was counterbalanced. After each scenario, the participants answered comprehension check questions to assess if they correctly understood that the same number of lives would be saved in each camp regardless of their camp sizes. We classified this replication as “very close”, with reference to the criteria by LeBel et al. (2018).

Extension.

Then, participants evaluated the benefits of sending a plane, answered a yes/no question on sending a plane, and, as an extension, indicated the amount they are willing to donate if they were awarded $5. After that, they answered a few questions in the funneling section and demographic section, followed by a debriefing section at the end.

Results

We present the descriptive statistics of each condition in Tables 7 and 8. We summarized the statistical tests of the hypotheses in the Supplementary Materials.

Benefit Ratings.

We conducted a 2 × 2 × 2 mixed-design ANOVA, with camp size and prior aid as within-subject independent variables, and plane reliability as the between-subject independent variable. We found support for a main effect of camp size, $F(1, 498) = 821.82$, $p < .001$, η^2_p = 0.62, 90% CI [0.58, 0.66]. Respondents perceived it was more beneficial to send planes to smaller camps ($M = 6.78$, $SD = 1.14$) than larger camps ($M = 4.02$, SD = 2.22). This successfully replicated the results of the original study. We also found support for a main effect of prior aid, $F(1, 498) = 43.10$, $p <.001$, $\eta^2_p = 0.08$, 90% CI [0.05, 0.12]. Respondents perceived it was more beneficial to send planes to camps with higher level ($M = 5.60$, SD = 1.58) than lower level of prior aid ($M = 5.19$, $SD = 1.56$), successfully replicating the original finding. Similarly, there was support for a main effect of plane reliability, $F(1, 497) = 5.70$, p = .029, $\eta^2_p = 0.01$, 90% CI [0.00, 0.03]. Respondents believed it was more beneficial to send planes with 100% reliability ($M = 5.54$, $SD = 1.40$) compared to planes with 60% reliability ($M = 5.24$, $SD = 1.37$), replicating the original finding successfully. Test statistics are provided in the Supplementary Materials.

Sending Water-Purifying Plane Decision.

We initially pre-registered to conduct an ANOVA, as the original article. However, during analysis stage we realized that yes/no binary responses such as the ones in this design are better analyzed with a binomial logistic regression. We present 2 X 2 X 2 ANOVA and t-test results in Supplementary Materials. We conducted a binomial logistic regression and found support for an association between camp size and plane sending decision, in which participants were more likely to send planes to small camps (95%) than to large camps (56%), $X^2(1) = 471.69$, $p < .001$, parameter estimate = 2.77 [2.46, 3.10]. This successfully replicated the original finding.

Consistent with the original finding, we failed to find support for the association between prior aid and plane sending decision, with no support for differences in sending plane percentage between the higher prior aid condition (75%) and the lower prior aid condition (75%), $X^2(1) = 0.37$, $p = .542$, parameter estimate = 0.10 [-0.22, 0.42].

Moreover, we failed to find support for the association between plane reliability and plane sending reliability, in which there is a minimal difference in sending plane percentage between 60%-reliability condition (74%) and 100%-reliability condition (76%), $X^2(1) = 0.05$, $p = .817$, parameter estimate = 0.04 [-0.28, 0.36]. We failed to replicate the effect of reliability on plane sending decision. There was no support for an interaction. We presented the descriptive statistics in Table 7.

Extension.

We conducted a 2 × 2 × 2 mixed-design ANOVA to measure respondents' reported donation amounts. We found support for a main effect of camp size on donation amount, $F(1, 498) = 55.01$, $p < .001$, $\eta^2_p = 0.10$, 90% CI [0.06, 0.14]. Respondents indicated higher donations to small camps ($M = 1.89$, $SD = 1.78$) than large camps ($M = 1.62$, $SD = 1.74$), in line with psychophysical numbing.

We also found support for a main effect in an opposite direction for prior aid, $F(1, 498) = 13.26$, $p < .001$, $\eta^2_p = 0.03$, 90% CI [0.01, 0.05]. Participants preferred donating to camps with lower prior aid ($M = 1.81$, $SD = 1.79$) than to camps with higher prior aid ($M = 1.70$, $SD = 1.71$).

Finally, we found no support for a main effect of system reliability on mean donation, $F(1, 497) = 1.23$, $p = .269$, $\eta^2_{\text{p}} = 0.00$, 90% CI [0.00, 0.01]. The 100% reliability condition ($M = 1.68$, $SD = 1.65$) and 60% reliability condition ($M = 1.85$, $SD = 1.78$) were very similar. See Table 11 for t-test results and Table 9 for descriptive statistics for amount of donation.

Full-sample results were very similar to post-exclusion results, regarding both main effects and interactions. Full results of Study 2a are in Tables 49, 51, 53-53 in supplementary.

Table 7

Study 2a: Descriptive statistics for Mean Beneficial Ratings

	Low Prior Aid		High Prior Aid	
	100% Plane Reliability	60% Plane Reliability	100% Plane Reliability	60% Plane Reliability
Small Camp	6.63 [1.41] (259)	6.38 [1.33] (240)	7.19 [1.28] (259)	6.90 [1.24] (240)
Large Camp	4.03 [2.46] (259)	3.68 [2.35] (240)	4.31 [2.57] (259)	4.01 [2.57] (240)

Note. Descriptives are in the format of *M* (*SD*) [*n*]

Table 8

Study 2a: Descriptive statistics for plane sending decisions (yes/no)

	Low Prior Aid		High Prior Aid	
	100% Plane Reliability	60% Plane Reliability	100% Plane Reliability	60% Plane Reliability
Small Camp	243 [94%] (259)	229 [95%] (240)	246 [95%] (259)	231 [96%] (240)
Large Camp	151 [58%] (259)	127 [53%] (240)	151 [58%] (259)	125 [52%] (240)

Note. Reporting format is - No. of "Yes" answers [Percentage of "Yes" answers] *(N)*

Table 9

Study 2a: Descriptive statistics for donations

	Low Prior Aid		High Prior Aid	
	100% Plane Reliability	60% Plane Reliability	100% Plane Reliability	60% Plane Reliability
Small Camp	1.84 [1.77] (259)	1.95 [1.85] (240)	1.80 [1.75] (259)	1.99 [1.86] (240)
Large Camp	1.66 [1.80] (259)	1.82 [1.94] (240)	1.40 [1.70] (259)	1.63 [1.84] (240)

Note. Descriptives are in the format of *M* (*SD*) [*n*]

Discussion

Replication.

Overall, Study 2a was a successful replication. Five out of six findings were consistent with the findings in the original study. The lone exception was the effect of plane reliability on plane sending decisions, which failed to replicate. Most importantly, however, we could replicate the effect of psychophysical numbing, as participants evaluated sending planes to smaller camps to be more beneficial than sending planes to larger camps. This is not in line with the findings of Studies 1a and 1b, and we will return to discuss this issue in the General Discussion. Similarly, participants perceived sending planes to camps with more prior aid, as more beneficial than sending planes to camps with less prior aid. It seems that participants preferred interventions in later stages when the threat is close to being contained. Regarding the decision of sending planes or not, consistent with original study, we found support for an effect of camp size, but not for an effect of prior aid. Moreover, the rated benefit for higher reliability plane was higher than lower-reliability plane, which is expected. However, we could not find support for a difference in the plane sending decision between the higher reliability and the lower-reliability condition. Overall, psychophysical numbing affected both perceived benefits and decision, but other manipulations yielded inconsistent results with different.

Extension.

Participants were willing to donate more to the smaller camps size than to the larger camps, finding an effect of psychophysical numbing on donation intention. Surprisingly, we found that participants were more willing to donate to camp with lower prior aid than higher prior aid. This is in the opposite direction to our expectations. This may be because participants believe that camps with limited prior aid are more in need of donation, contradicting the above finding that participants were more likely to perceive sending planes to camp with higher prior aid as more beneficial. It appears the influence of prior aid information on intention of donation differs substantially from perceived benefits of sending planes. Finally, we failed to find support for the hypothesis that participants would be willing to donate more when the plane had higher reliability.

Study 2b - Very Close Replication on Prolific

Method

Participants.

We recruited 2020 participants on Prolific. We excluded 414 participants based on a pre-registered exclusion criteria, leading to the final analyzed sample of 1606 participants (M_{age} = 38.31, SD_{age} = 13.03; 565 males, 1036 females, 5 other). This sample size can detect *d* = 0.37, the weakest statistically significant effect in the original article, with 99.9+% power, and *d* = 0.12, the weakest not statistically significant effect in the original article, with 92% power. Participants were paid £0.35 for this task.

Replication.

As in Study 2a, participants first read a scenario describing a Rwandan refugee crisis. Refugees were suffering from water-borne disease, and would need purified water to survive. One country was considering sending a purification plane to the refugees. There were 8 scenarios, as a results of a 2 (camp sizes: 11,000 and 250,000) X 2 (levels of prior pure water aid received: low and high) X 2 (levels of reliability of the planes: 60% and 100%) experimental design. However, it is important to note that we changed the within-subject design in the original article to a between-subject design, in which we randomly assigned participants to evaluate only one of the eight scenarios. After reading the scenario, participants answered four comprehension questions to ensure that they read and understood the scenario. As in Study 2a and the original target study, after reading the scenario, participants evaluated the benefits of sending a plane and answered a yes/no question on sending a plane with the presented benefits. We classified the replication we conducted in Study 2b as a "close replication", with reference to the criteria by LeBel et al. (2018) (see supplementary Tables 12 -13).

Extension.

As an extension, participants also answered a question on donation. The donation question was different from that of Study 2a, where we asked participants how much they would donate out of $5. In Study 2b, we asked the participants what percentage of their earnings for the present task they would be willing to donate. After these questions, participants answered a few questions in the funneling section and demographic section, then read the debriefing statement.

Results

We presented the descriptive statistics for benefit ratings in Table 10, and plane sending decision in Table 11.

Replication

Benefits rating.

We conducted a 2 × 2 × 2 between-subject ANOVA on benefit ratings regarding psychophysical numbing hypothesis. For camp size, we found support for a main effect, $F(1, 1598) = 511.60$, $p < .001$, $\eta^2_p = 0.24$, 90% CI [0.21, 0.27]. Note that despite the 90% CI not overlapping with the original one, we still consider this a successful replication, as $\eta^2 p$ is calculated in a different, more stringent way in a between-subjects ANOVA like the one we ran compared to a within-subjects ANOVA like the one the original authors ran (though we note this interpretation and justification was not pre-registered). The benefit ratings were higher for small camps (M = 7.26, SD = 1.30) compared to large camps (M = 5.34, SD =

2.13), which is consistent with the original findings and with the notion of psychophysical numbing. For the level of prior aid, we found support for a main effect, *F*(1, 1598) = 9.10, *p* = .003, η^2_p = 0.01, 90% CI [0.00, 0.01]. An independent-sample t-test indicated that, the benefit ratings were higher for higher prior aid condition (*M* = 6.41, *SD* = 1.91), compared to lower prior aid condition (*M* = 6.21, *SD* = 2.10), which successfully replicated the original finding. We conducted an ANOVA and found support for a main effect of reliability, *F*(1, 1598) = 88.37, *p* < .001, η^2_p = 0.05, 90% CI [0.04, 0.07]. Participants rated benefits as higher for system with 100% reliability (*M* = 6.72, *SD* = 1.97) compared to system with 60% reliability (*M* = 5.89, *SD* = 1.95). This is consistent with the original finding. Our results supported all three hypotheses, as we replicated all three main effects for benefit ratings successfully.

Table 10

Study 2B: Descriptive statistics of benefit ratings

	100% reliability		60% reliability	
	High Prior-aid	Low Prior-aid	High Prior-aid	Low Prior-aid
Small Camp	7.42 (1.41) [208]	8.04 (1.49) [215]	6.07 (1.07) [189]	6.43 (1.39) [203]
Large Camp	5.95 (2.00) [199]	5.33 (3.54) [175]	7.07 (2.25) [203]	4.84 (1.96) [191]

Note. Descriptives are in the format of *M* (*SD*) [*n*]

Sending Water-Purifying Plane Decision.

In a binomial logistic regression [4]we found support for an association between camp size and plane sending decision, in which the percentage of plane sending for the small camp (97%) is higher compared to the large camp (74%), $X^2(1)$ = 184.43, *p* < .001, parameter estimate = 2.49 [2.04, 3.01]. This successfully replicated the original finding. In a binomial

[4] The original article analyzed these decision with an ANOVA, and we initially planned to conduct an ANOVA. However, yes/no binary responses are more suited to a binomial logistic regression. We moved the presentation of the results of the 2 X 2 X 2 ANOVA and t-test results to the Supplementary Materials.

logistic regression we failed to find support for the association between prior aid and plane sending decision, as difference in percentage of plane sending between low aid condition (87%) and high aid condition (84%) was very small, $X^2(1) = 0.06$, $p = .806$, parameter estimate = 0.06 [-0.42, 0.56]. This is consistent with the original finding. In a binomial logistic regression we failed to find support for the association between plane reliability and plane sending decision, in which the difference of plane sending between 100% reliability condition (87%) and 60% reliability condition (84%) was very small, $X^2(1) = 1.15$, $p = .284$, parameter estimate = 0.26 [-0.23, 0.75]. This failed to replicate the original finding. We found substantial support for 1 out of 3 of our hypotheses. We successfully replicated the findings for prior aid and camp size, but not for plane reliability. Additionally, we found support for an interaction between prior aid and plane reliability. For condition with 60% plane reliability, the difference in plane sending decision % between higher prior aid and lower prior aid is larger, compared to that of 100% plane reliability, $X^2 = 4.43$, $p = .035$, parameter estimate = 0.99 [0.07, 2.01]. We found no support for other interactions.

Table 11

Study 2B: Descriptive statistics of Yes/No plane sending decisions

	100% reliability		60% reliability	
	High Prior-aid	Low Prior-aid	High Prior-aid	Low Prior-aid
Small Camp	202 [97.1%] (208)	211 [98.1%] (215)	186 [98.4%] (189)	190 [93.6%] (203)
Large Camp	144 [72.4%] (199)	136 [77.7%) (175)	141 [69.5%] (203)	146 [76.4%] (191)

Note. Reporting format is - No. of "Yes" answers [Percentage of "Yes" answers] *(N)*

Extension.

We conducted a 2 × 2 × 2 full between-subject ANOVA, with camp size, prior aid and plane reliability as between-subjects factors. For camp size, we found support for a main effect, *F*(1, 1598) = 4.59, *p* = .032, η^2_p = 0.00, 90% CI [0.00, 0.01]. An independent-sample t-test indicated that, the donation, in percentage of earnings, under the small camp size condition (*M* = 46.87, *SD* = 42.17) was larger than the donation amount under the large camp size condition (*M* = 42.43, *SD* = 42.24). This indicated that, even as a percentage of their compensation, participants were influenced by psychophysical numbing. However, we found no support for a main effect of prior aid, *F*(1, 1598) = 0.90, *p* = .344, η^2_p = 0.00, 90% CI [0.00, 0.00]. An independent sample t-test indicated not support for differences between low aid condition (*M* = 43.79, *SD* = 42.18) and high aid condition (*M* = 45.59, *SD* = 42.32). Similarly, for system reliability, we found no support for a main effect, *F*(1, 1598) = 2.84, *p* = .092, η^2_p = 0.00, 90% CI [0.00, 0.01]. An independent-sample t-test indicated no support for differences between 100% reliability condition (*M* = 43.02, *SD* = 42.26) and 60% reliability (*M* = 46.42, *SD* = 42.20). In summary, we found support for only one of our hypotheses. Descriptive statistics are presented in Table 12.

Full sample results.

Results for the full sample were very similar, with a few minor differences. Before exclusion, we detected difference between 100% reliability and 60% reliability, with binomial regression, yet we failed to detect support for the effect after exclusion. In addition, we did not find support for the interaction between prior aid and plane reliability for plane decision before exclusion, but we found support for the interaction after exclusion. See Supplementary Tables 50, 52, 56-58, for full-sample results of Study 2b.

Table 12

Study 2B: Descriptive statistics of donation percentage of earnings

	100% reliability		60% reliability	
	High Prior-aid	Low Prior-aid	High Prior-aid	Low Prior-aid
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Small Camp	44.02 (7.07)	48.18 (41.88)	46.07 (42.79)	49.37 (41.20)
Large Camp	42.61 (42.51)	36.07(41.25)	46.61 (42.65)	43.67 (42.21)

Discussion

Replication.

Overall, Study 2b was a successful replication. Most importantly, the effects of camp size on benefits ratings was successfully replicated, in line with psychophysical numbing. Further, plane sending decision, the effect of prior aid on rated benefits, and the lack of effect of prior aid on plane sending decision successfully replicated. However, the effect of reliability on plane sending decision failed to replicate. It is important to recognize that despite changing the original study from within-subject design to between-subject design, most of the findings are consistent with the original study. For benefits ratings, our findings of the effect of camp size and prior aid, the two independent variables of interest, are in line with the psychophysical numbing hypothesis and original study findings. The perceived benefit of sending planes to smaller camps was higher than the perceived benefits of sending planes to larger camps. Similarly, the perceived benefit of sending planes to camps that have satisfied more need, i.e. with more prior aid, was higher than the perceived benefits of sending planes to camps that have satisfied less need, i.e. with lower prior aid. This is consistent with the idea that there appeared to be general preference for interventions in later stages, when compared to earlier stages (Kivetz et al., 2006). For hypothetical decisions, we found support for the effect of camp size but not for the effect of prior aid. Both these replication findings are consistent with the original findings. For system reliability, we found support that the rated benefit for higher reliability plane was higher than lower-reliability plane. However, we could not find support for an effect of reliability, which is surprising.

Extension.

For our extension on donation, we found support for the hypothesis that there would be more donation for small camp size, but not hypotheses on prior aid and plane reliability. All three effects were weak, although support for the effect of camp size, again, in line with psychophysical numbing.

General Discussion

In the present paper, using much larger samples on MTurk (USA) and Prolific (UK) we attempted to replicate Study 1 and 2 of Fetherstonhaugh et al. (1997), and added several extensions. Below, we discuss the replications (see a quantitative summary in Table 13) and extensions, with a particular focus on possible reasons for our findings. Then, we discuss theoretical and practical implications, and provide indication for future research.

Table 13

Quantitative comparison between original and replication results of camp size on preference for aid (Study 1a, 1b, 1c) and benefit ratings (Study 2a and 2b)

Study	Original effect size [90% CI]	Replication effect size [90% CI]	Replication classification following LeBel et al. (2019)
Study 1a	η^2_p = 0.14 [0.02, 0.28]	η^2_p = 0.06 [0.02, 0.10]	Signal-inconsistent (opposite)
Study 1b	η^2_p = 0.14 [0.02, 0.28]	η^2_p = 0.21 [0.17, 0.26]	Signal-inconsistent (opposite)
Study 1c	η^2_p = 0.14 [0.02, 0.28]	η^2_p = 0.12 [0.08, 0.17]	Signal-inconsistent (opposite)
Study 2a	η^2_p = 0.55 [0.45, 0.62]	η^2_p = 0.62 [0.58, 0.66]	Signal-consistent
Study 2b	η^2_p = 0.55 [0.45, 0.62]	η^2_p = 0.24 [0.21, 0.27]*	Signal-consistent

Note. We calculated 90% CIs because η^2_p cannot be smaller than zero, unlike other effect size measures.

*Despite the 90% CI not overlapping with the original one, we still consider this a successful replication, as η^2_p is calculated in a different, more stringent way in a between-subjects ANOVA like the one we ran compared to a within-subjects ANOVA like the one the original authors ran.

Replications

Three times, we were unable to successfully replicate the findings in the original's Study 1, and in fact found results going in the opposite direction. Twice, we successfully replicated Study 2. How to reconcile these contrasting results? What does this mean for psychophysical numbing? The answer is not so straightforward. Considering the very strong results of the both the successful replications of Study 2 and the consistent failures in replications of Study 1 with opposite findings, we believe the phenomenon of psychophysical numbing crucially depends on the paradigm of choice. It is possible to find *reverse* psychophysical numbing (or *psychophysical sensitization*) with the paradigm that Fetherstonhaugh et al. (1997) used in Study 1, and find psychophysical numbing using the paradigm Fetherstonhaugh et al (1997) used in Study 2. Below, we discuss some factors that may help explain this discrepancy.

Inconsistent findings: Discussion of possible factors

Evaluation mode.

In our Studies 1a-1b, faithful to the original design, we used a within-subjects design and observed evidence for psychophysical *sensitization*, whereas in Study 2b we switched to a between-subjects design, unlike the original, and obtained evidence for psychophysical *numbing.* A possible explanation, suggested by the different designs we employed, is that there are so-called "preference reversals" when the small and large camps are presented in a within-subjects design (joint evaluation), compared to when they are presented in a between-subjects design (separate evaluation). Against this notion, however, is the fact that we could find evidence for psychophysical numbing in Study 2a despite manipulating camp size within-subject, and the design used by the original authors. In both the original studies 1 and 2, Fetherstonhaugh et al. (1997) manipulated the camp size using within-subjects designs, presenting all scenarios to all participants, and obtained evidence supporting psychophysical numbing. It is therefore hard to reconcile completely the original results with our replication results.

The notion that manipulating camp size either between- or within-subjects can reverse the effects observed in Fetherstonhaugh et al. (1997) would connect the research on psychophysical numbing to the research on evaluation mode (Hsee et al., 1999), and would have several interesting theoretical implications for the understanding of how people reason about proportions and their sensitivity to the value of human life in different contexts, in addition to practical implications regarding charity and aid messaging. Testing this hypothesis is outside of the scope of this paper, but in our mind it is worthy of future research.

Time.

Other explanations, based on the passage of time and the different samples between the original and the present replications, are less convincing. While the original study was conducted in or before 1997, proposing that a change in norms or culture in the twenty-two years between the original study and the present replications as a cause of the failed replications of Study 1 can hardly explain why, on the other hand, results of Study 2 were convincingly and successfully replicated twice. Note that such an explanation would significantly reduce the theoretical importance of psychophysical numbing as a phenomenon. If the passage of time is enough to reduce its strength, psychophysical numbing could hardly be proposed as an explanation for why genocides are still happening (Slovic, 2007) or why identifiable victims are more effective than large groups in eliciting charity donations and aid (Butts et al., 2019; Jenni & Loewenstein, 1997; Lee & Feeley, 2016). Further, other studies in judgment and decision making have been successfully replicated after a similar time-lag (e.g., Ziano, Wang, et al., 2020).

Participants.

Along the same lines, it is also hard to argue that Study 1 of the original paper failed to replicate because of the different participants (U.S. American undergraduates in the original study, MTurk and Prolific participants in the present study), for two reasons. First, Study 1 failed to replicate in two different samples, both MTurk and Prolific, therefore requiring college students to be different from *both* MTurkers and Prolific participants. Second, Study 2 was successfully replicated in both the MTurk and the Prolific sample, therefore making this explanation unviable, as Study 2 was also conducted on U.S. American undergraduates in the original paper. Again, note that invoking differences in samples to explain a failed replication undermines the contribution of the original findings by reducing its generalizability. If a finding cannot be replicated in a different sample, its importance for understanding human psychology and behaviour is reduced. Finally, other studies in judgment and decision making conducted on U.S. American undergraduates in the 1990s have been successfully replicated on online samples such as MTurk (e.g., Ziano, Tsun, Lei, & Kamath, 2020), casting further doubt on this explanation.

Other plausible explanations for the differences between the failed replication of Study 1 and the successful replication of Study 2 regard the role of political ideology, familiarity with refugee crises, and the absence of a direct comparison between the large and the small camp in Study 1.

Political ideology and aid importance.

Examining Study 1a and 1b participants' explanations of their decisions there were indications of a "my country first" theme. This decision strategy seemed to prioritize the investment of funds for the in-group over helping foreign countries, at times emphasizing the in-group-outgroup factor over the second factor contrasting infrastructure versus lives factor. For example, some responses from Study 1a were "Help USA first," "I would want to take care of my own country." Therefore, it is possible that shifting attitudes in the US and UK towards focusing inwards, as indicated by the policies of the elected politicians in both countries since 2016, and the indirect comparisons contrasting against in-group favouring policies in Study 1, may have somehow affected the findings. If there is any truth to this idea, we should find that considering domestic problems more important than refugee aid should be correlated with psychophysical sensitization. Following the same reasoning, we should find a correlation between psychophysical numbing and political ideology, as U.S. American conservatives are more likely to favour the ingroup in terms of international aid compared to U.S. American liberals (Kull, 2017).

However, in Study 1c, we did not find support in favour of the hypothesis that political ideology, absolute aid importance (as measured on Likert-type items), and relative aid importance (as measure by the imagined distribution of $100 million between unemployment and transportation programs in one's own country and refugee aid in a foreign country) were associated with psychophysical sensitization, with very small effect sizes. We found support for only one association with the extent of psychophysical sensitization - the budget allocated to addressing unemployment problem, albeit the effect was small and the p-values associated with the effect were just below the alpha level we set in advance (5%): *r*(435) = −.10, 95% CI [-.004, -.19], *p* = .042. It is unclear why this measure, and not the other ones, may correlate with psychophysical numbing. Considering the large number of tests we conducted, it is possible that this is an instance of a false positive finding. A Bayesian look at the p-value also suggests quite weak evidence in favor of the alternative hypothesis for this particular correlation (Sellke et al., 2001), because it is very close to the pre-specified alpha level of 5%. Therefore, we find it unlikely that a more inward-looking ideological turn in the wake of the political events of 2016 is a viable explanation of our results.

Self-reported knowledge and refugee number estimation.

One could argue that perhaps participants did not know much about the refugee crisis in Rwanda when we conducted the study, and that a more current refugee crises could yield different results. In Study 1a we tested an additional location (Haiti) and In Study 1c we changed the refugee crisis to yet another context (South Sudan), and again we found evidence in favour of psychophysical sensitization and against psychophysical numbing. Further, we found no evidence that the self-reported familiarity with several of the major refugee crises happening in 2021, and the estimation of how many refugees are involved in a crisis were correlated with the extent of psychophysical sensitization, with very small effect sizes. We only detected an association in one measure, yet it was very small and associated with a p-value very close to the alpha level we set in advance (5%), providing weak evidence for the alternative hypothesis. We struggled to find a reason to explain why psychophysical sensitization should only correlate with the refugee estimation in this country, after outliers were excluded, and not with all the other measures we included here. Since we ran a lot of statistical tests, it is possible that this one is a false positive. Overall, participants reported that they did not know much about any of the five largest refugee crises happening in 2021, and we did not find evidence in favour of the hypothesis that knowledge or estimation of the size of a refugee crisis is associated with psychophysical sensitization. These results are hard to reconcile with the notion that these factors contribute to psychological sensitization, and are even harder to reconcile with the notion that changes in their levels across time resulted in psychophysical sensitization in the present paper but in psychophysical numbing in the original work

Direct comparison.

In Study 2a and 2b there was a direct comparison between the programs meant to save lives and all programs were about aid to a foreign country, yet in Study 1a and 1b there was no such direct comparison, and the comparison pitted refugee aid against programs investing in own country infrastructure (transportation) and economy (unemployment) in an indirect fashion[5]. This is a possible explanation of the contradicting results of our replications. In Study 1c, we introduced a direct measure of preference for aid for the smaller rather than the larger camp, and once more we found strong evidence in favour of psychophysical sensitization. Therefore, we find it unlikely that the design of Study 1, which focuses on indirect comparisons, may be responsible for the results of Study 1a, 1b, or 1c.

Presentation format, and inclusion and salience of key information

Several other features differed across the two studies. We already mentioned the difference between the studies in that Study 1 involved a comparison of aid programs to domestic infrastructure projects, whereas Study 2 only compared aid programs. In addition, Study 2 introduced several other factors in describing the aid programs (prior aid, needs met, etc.). One possible direction for future research is to include neutral conditions to investigate whether removing any of these factors has any impact on aid decision making.

There were also differences in the format of presentation, and the salience of the key information. Study 2 presented some of the information in percentages, and there were differences between the studies in whether both the number of lives saved -and- the size of the camp were presented together and were salient. Future research may follow to investigate whether these presentation factors and information salience may affect aid preferences.

Extensions

We conducted several extensions, which we review and discuss below.

[5] However, in Study 1 and in our replications (Study 1a, 1b, and 1c), a three-options or a dichotomous aid choice was included. While in the original Study 1 this choice favored the smaller camp, in our replications, participants tended to favor aid to the larger camp.

Study 1a extension.

The extension of Study 1a was unsuccessful. Perhaps Haiti did not engender sufficient feelings of physical and psychological closeness compared to Rwanda. The interpretation of this extension is complicated by the fact that we could not replicate the original findings, but rather obtained opposite results. Nonetheless, it is important that this variant of the original study be tested, as we now know of this null effect.

Study 1b extension.

The extension of Study 1b was successful. Knowing that other countries are considering the same type of aid moderated the impact of camp size on our dependent variable. This is consistent with findings in diffusion of responsibility (Wiesenthal et al., 1983). Again, it is important that this variant of the original study was tested, although the interpretation of this extension is complicated by the fact that we could not replicate the original findings. We therefore recommend caution in interpreting these results.

Study 1c extensions.

In Study 1c, we found support for psychophysical sensitization in a direct comparison of the two aid programs. Political ideology, knowledge, estimation of the number of refugees, and absolute and relative importance of refugee aid (compared with domestic programs) showed very small correlations with psychophysical sensitization. Overall, we could not find support for the hypotheses underlying these associations.

Studies 2a-2b extensions.

The extensions of Studies 2a and 2b found similar and important results, despite the different question format (sum versus percentage). We found that hypothetical donations follow psychophysical numbing, such that participants were more willing to donate to smaller camps than larger camps. Future research may wish to investigate the effects of psychophysical numbing on actual donations, which may differ from hypothetical donation (Lee & Feeley, 2016).

Theoretical and practical implications

These results are important for researchers in social psychology. The fact that we could replicate psychophysical numbing in Studies 2a-2b has importance for research on people's lay perceptions of charity effectiveness (Karlan & Wood, 2017). Efforts to correct people's erroneous perceptions of and charity and aid communications (Caviola, Schubert, & Nemirow, 2020; Caviola, Schubert, Teperman, et al., 2020) should continue including efforts to correct people's beliefs and attitudes towards charity effectiveness, or at least devise methods in which these incorrect attitudes can be overcome (e.g., Gneezy et al., 2014). However, the fact that we found reverse psychophysical numbing in Studies 1a-1c is both worrying and exciting for the very same area of research. On the one hand, it seems that psychophysical numbing depends on how it is elicited and the paradigm of choice can produce opposite results. On the other hand, this opens potentially fruitful avenues for subsequent research building on the present work. In fact, we encourage future research to study where psychophysical numbing happens, where it does not, where one can find psychophysical sensitization, and what causes psychophysical sensitization. Psychophysical numbing was predicated on diminishing sensitivity to additional number of lives above a certain threshold. It is clear that psychophysical sensitivity cannot have the same theoretical rationale. Overall, this implies that psychophysical numbing may depend on the experimental paradigm of choice, although the exact details that yield such changes are yet to be identified. We therefore encourage researcher wishing to extend the literature on this topic to take into account the present results before they proceed with their investigation, as different procedures may yield vastly different – and puzzling – results compared to expectations.

Constraints on Generality

Participants.

We recruited U.S. American and British participants. Research using different samples may obtain different results. Particularly interesting would be a cross-cultural examination of psychophysical numbing, for instance by investigating participants in non-WEIRD cultures (Henrich et al., 2010), perhaps through multi-lab/multi-country collaborations such as the Psychological Science Accelerator (Moshontz et al., 2018), to test whether psychophysical numbing and sensitization extend to non-WEIRD cultures and whether our results are robust to cultural variations. Given our proposed explanation for the differences between the replication of Studies 1 and 2, it might be especially relevant to examine Study 1 in countries less focused inwards and adopting more universal values.

Materials.

We used materials from Fetherstonhaugh et al. (1997). Materials that dramatically changes the context used in the original study (that is, a refugee crisis in a developing country) may yield different results. In fact, we encourage future research to try different variations on this important matter, perhaps with other environmental, war, and medical scenarios, to test the possible boundary conditions of psychophysical numbing. It is possible that variations in the scenario and the context will add fundamental insight to the nature and the limits of psychophysical numbing. We believe that testing the effect of evaluation (joint vs. separate) on psychophysical numbing, while outside of the scope of this research, can be very valuable to understand its scope and its boundary conditions.

Conclusion

More than twenty years later, we attempted very close replications and extensions Fetherstonhaugh et al. (1997)'s psychophysical numbing effects. We found evidence for a *reversal* of psychophysical numbing when attempting to replicate Study 1, which we deemed psychophysical *sensitization.* However, we successfully replicated Study 2, and we found that donation intentions are affected by psychophysical numbing in Study 2. Psychophysical numbing and sensitization may depend on procedural aspects of the experimental design that are yet to be identified.

References

Agnoli, F., Fraser, H., Thorn, F. S., & Fidler, F. (2020). *Australian and Italian Psychologists' View of Replication*. https://psyarxiv.com/ks48e/

Alicke, M. D. (1985). Global Self-Evaluation as Determined by the Desirability and Controllability of Trait Adjectives. *Journal of Personality and Social Psychology*, *49*(6), 1621–1630. https://doi.org/10.1037/0022-3514.49.6.1621

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In *Groups, leadership, and men. S* (pp. 222–236).

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods Matter: P-Hacking and Causal Inference in Economics. *American Economic Review*, *11796*. www.iza.org

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1–32. https://doi.org/10.1257/app.20150044

Butts, M. M., Lunt, D. C., Freling, T. L., Gabriel, A. S., Cox, E. L., Box, P. O., & States, U. (2019). Organizational Behavior and Human Decision Processes Helping one or helping many ? A theoretical integration and meta-analytic review of the compassion fade literature. *Organizational Behavior and Human Decision Processes*, *151*(May 2015), 16–33. https://doi.org/10.1016/j.obhdp.2018.12.006

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

https://doi.org/10.1038/s41562-018-0399-z

Caviola, L., Faulmüllert, N., Everett, J. A. C., Savulescu, J., & Kahane, G. (2014). The evaluability bias in charitable giving: Saving administration costs or saving lives? *Judgment and Decision Making*, *9*(4), 303–315. https://doi.org/10.13140/2.1.1028.9287

Caviola, L., Schubert, S., & Nemirow, J. (2020). The many obstacles to effective giving. *Judgment and Decision Making*, *15*(2), 159–172. https://doi.org/10.31234/osf.io/3z7hj

Caviola, L., Schubert, S., Teperman, E., Faber, N. S., & Moss, D. (2020). Donors vastly underestimate differences in charities ' effectiveness. *Judgement and Decision Making*, *15*(4), 509–516.

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. *The Behavioral and Brain Sciences*, *41*, e124. https://doi.org/10.1017/S0140525X18000596

Coppock, A. (2017). Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods*. https://doi.org/http://alexandercoppock.com/papers/Coppock_generalizability.pdf

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 1–15.

Evangelidis, I., & van Osselaer, S. M. J. (2017). Points of (Dis)Parity: Expectation Disconfirmation from Common Attributes in Consumer Choice. *Journal of Marketing Research*, *June 2014*, 1–55.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf und Härtel.

Fetherstonhaugh, D., Slovic, P., Johnson, S. M., & Friedrich, J. (1997). Insensitivity to the

Value of Human Life: A Study of Psychophysical Numbing. *Journal of Risk and Uncertainty*, *14*(3), 283–300. https://doi.org/10.1023/A:1007744326393

Fiennes, C. (2017). We need a science of philanthropy. *Nature*.

Frederick, S. (2012). Overestimating Others' Willingness to Pay. *Journal of Consumer Research*, *39*(1), 1–21. https://doi.org/10.1086/662060

Friedrich, J., Barnes, P., Chapin, K., Dawson, I., Garst, V., & Kerr, D. (1999). Psychophysical numbing: When lives are valued less as the lives at risk increase. *Journal of Consumer Psychology*, *8*(3), 277–299. https://doi.org/10.1207/s15327663jcp0803_05

Gneezy, U., Keenan, E. a, & Gneezy, A. (2014). Avoiding overhead aversion in charity. *Science (New York, N.Y.)*, *346*(6209), 632–635. https://doi.org/10.1126/science.1253932

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research*, *35*(3), 472–482. https://doi.org/10.1086/586910

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576–590. https://doi.org/10.1037/0033-2909.125.5.576

Jenni, K., & Loewenstein, G. (1997). Explaining the "Identifiable Victim Effect." *Journal of Risk and Uncertainty*, *14*(3), 235–257. https://doi.org/10.1023/A:1007740225484

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–

532. https://doi.org/10.1177/0956797611430953

Jung, M. H., Moon, A., & Nelson, L. D. (2019). Overestimating the valuations and preferences of others. *Journal of Experimental Psychology: General*.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Kant, I. (1785). *Fundamental principles of the metaphysics of ethics*. Longmans, Green, and co.

Karlan, D., & Wood, D. H. (2017). The effect of effectiveness: Donor response to aid effectiveness in a direct mail fundraising experiment. *Journal of Behavioral and Experimental Economics , 66*, 1–8. https://doi.org/10.1016/j.socec.2016.05.005

Kivetz, R., Urminsky, O., & Zheng, Y. (2006). The Goal-Gradient Hypothesis Resurrected: Purchase Acceleration, Illusionary Goal Progress, and Customer Retention. *Journal of Marketing Research, 43*(1), 39–58. https://doi.org/10.1509/jmkr.43.1.39

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology, 45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Kothe, E. J., & Ling, M. (2019). *Retention of participants recruited to a one-year longitudinal study via Prolific*. 1–6. https://doi.org/10.31234/osf.io/5yv2u

Kull, S. (2017). American public support for foreign aid in the age of Trump. *Brookings,* 1–7. www.brookings.edu/wp-content/uploads/2017/08/global-20170731-blum-stevenkull-brief-6.pdf

Landy, D., Guay, B., & Marghetis, T. (2017). Bias and ignorance in demographic perception. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-017-1360-2

Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. *Cognitive Science*, *37*(5), 775–799. https://doi.org/10.1111/cogs.12028

Latané, B., Williams, K., & Harkins, S. (2006). Many hands make light the work: The causes and consequences of social loafing. *Small Groups: Key Readings*, *37*(6), 297–308. https://doi.org/10.4324/9780203647585

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*(2), 254–261. https://doi.org/10.1037/pspi0000106

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2018). A Brief Guide to Evaluate Replications. *Meta-Psychology*, *541*, 1–17. https://doi.org/10.31219/osf.io/paxyn

Lee, S., & Feeley, T. H. (2016). The identifiable victim effect: A meta-analytic review. *Social Influence*, *11*(3), 199–215.

Lovakov, A., & Agadullina, E. (2021). Empirically Derived Guidelines for Interpreting Effect Size in Social Psychology. *European Journal of Social Psychology*. https://doi.org/doi: 10.1002/EJSP.2752

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The Generalizability of Survey Experiments. *Journal of Experimental Political Science*, *2*(2), 109–138. https://doi.org/10.1017/XPS.2015.19

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Sellke, T., Bayarri, J. J., & Berger, J. O. (2001). Calibration of p-values for testing precise hypotheses. *The American Statistican*, *55*(1), 62–71. https://doi.org/10.1198/000313001300339950

Shah, A. K., Mullainathan, S., & Shafir, E. (2019). An exercise in self-replication: Replicating Shah, Mullainathan, and Shafir (2012). *Journal of Economic Psychology*, *75*(May 2018), 102127. https://doi.org/10.1016/j.joep.2018.12.001

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Slovic, P. (2007). “If I look at the mass I will never act.” *Judgment and Decision Making*, *2*(2), 37–59. https://doi.org/10.1007/978-90-481-8647-1_3

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9). https://doi.org/10.1098/rsos.160384

Snowberg, E., & Yariv, L. (2018). *Testing the Waters: Behavior across Participant Pools*.

Stevens, S. S. (1975). *Psychophysics*. Wiley.

Sudhir, K., Roy, S., Cherian, M., Roy, S., & Cherian, M. (2016). Do Sympathy Biases Induce Charitable Giving ? The Effects of Advertising Content Do Sympathy Biases Induce Charitable Giving ? The Effects of Advertising Content. *Marketing Science*, *November 2020*.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463. https://doi.org/10.1037/a0018963

United Nations High Commissioner for Refugees. (2021). Figures at a Glance. *Unhcr.Org*. https://www.unhcr.org/figures-at-a-glance.html

Weber, E. H. (1834). De pulsu, resorptione, auditu et tactu. In *Annotationes Anatomicae et Physiologicae*. Koehler.

Williams, L. E., & Bargh, J. A. (2008). Keeping One ' s Distance. *Psychological Science2*, *19*(3), 302–308.

Zagefka, H., & James, T. (2015). The psychology of charitable donations to disaster victims and beyond. *Social Issues and Policy Review*, *9*(1), 155–192. https://doi.org/10.1111/sipr.12013

Ziano, I., Mok, P. Y. (Cora), & Feldman, G. (2020). Replication and Extension of Alicke (1985) Better-Than-Average Effect for Desirable and Controllable Traits Replication and Extension of Alicke (1985) Better-Than-Average Effect for Desirable and Controllable Traits. *Social Psychological and Personality Science*, *June*. https://doi.org/https://doi.org/10.1177/1948550620948973

Ziano, I., & Pandelaere, M. (2018). The majority premium: Competence inferences derived from majority consumption. *Journal of Business Research*, *92*. https://doi.org/10.1016/j.jbusres.2018.08.002

Ziano, I., Tsun, S. M., Lei, H. C., & Kamath, A. A. (2020). *Revisiting " Money Illusion ": Replication and Extension of Shafir et al . (1997)* (Issue February). https://doi.org/10.13140/RG.2.2.10679.47525

Ziano, I., & Villanova, D. (2021). *Using Self-Generated Priors in Division Leads to Biased Consumer Judgments*. PsyArXiv. https://psyarxiv.com/78xp4/

Ziano, I., & Wang, D. (2021). Slow Lies: Response Delays Promote Perceptions of Insincerity. *Journal of Personality and Social Psychology*. https://doi.org/https://doi.org/10.1037/pspa0000250

Ziano, I., Wang, Y. J., Sany, S. S., Feldman, G., Ngai, L. H., Lau, Y. K., Bhattal, I. K., Keung, P. S., Wong, Y. T., Tong, W. Z., Cheng, B. L., Yan, H., & Chan, C. (2020). Perceived morality of direct versus indirect harm : Replications of the preference for indirect harm effect In press at Meta Psychology . Accepted for publication on Jan 30 , 2020. *Meta-Psychology*. https://psyarxiv.com/bs7jf

Zwaan, R. A., Etz, A., Richard, L., & Donnellan, M. B. (2017). Making Replications Mainstream. *Behavioral and Brain Sciences*. https://doi.org/10.17605/OSF.IO/4TG9C

Fetherstonhaugh et al. (1997) Replications and extensions:

Supplementary

Table of contents

Open Science disclosures

Data and code + Pre-registrations

Data and code are shared using the Open Science Framework. Review link for data and code of the study: https://osf.io/786jg/?view_only=32a44611c63d4d2787ac139192d26c71

Procedure and data disclosures

Data collection

Data collection was completed before analyzing the data.

Conditions reporting

All collected conditions are reported.

Data exclusions

Details are reported in the materials section of this document

Variables reporting

All variables collected for this study are reported and included in the provided data.

Study 1 Supplementary

Analysis of the target study

Effect size calculation

The effect of our interest was the camp size main effect, i.e., $F(1, 52) = 8.24$, $p < .01$ with the full sample and $F(1, 21) = 3.92$, $p = .06$ with those who indicated equivalence between the two refugee programs. The former translates to a $\eta^2_p = 0.14$, 90% CI [0.02, 0.28], and the latter translates to a $\eta^2_p = 0.16$, 90% CI [0.00, 0.37] (its actual coverage should be larger than 90% because it was a non-significant effect).

The authors might have omitted reporting one case of exclusion. Otherwise, given a sample size $n = 54$ and a two-way repeated-measures ANOVA, the error df should be 53 instead of 52.

Power analyses

Because there are no analytic procedures available for the power analysis of a two-factor repeated-measures ANOVA, and because we are only interested in the main effect of a two-level within-subject factor (and the other main effect and their interaction were not significant), we conducted power analysis assuming what was done was a paired-sample t-test.

The *F*-statistics were converted to *t*-statistics, which were then used to determine Cohen's *d*s and the required *N* for the design to be powered at .95, assuming alpha = .05.

Our calculation shows that $F(1, 52) = 8.24$ is equivalent to $t(52) = 2.87$. The corresponding Cohen's *d* was 0.40, 95% CI [0.11, 0.68]. At least 84 participants are needed to detect this *d* at power = .95 in a paired-sample *t*-test. $F(1, 21) = 3.92$ is equivalent to $t(21) = 1.98$. The corresponding Cohen's *d* was 0.43, 95% CI [-0.02, 0.87]. At least 72 participants are needed to detect this *d* at power = .95. To sum up, our power analysis suggests that we need 84 participants.

Exclusion criteria

General criteria

1. Participants indicating a low proficiency of English (self-report < 5 on a 1-to-7 scale).
2. Participants reporting not being serious about the survey (self-report < 4 on a 1-to-5 scale).
3. Participants who correctly guessed the hypothesis of this study in the funneling section.
4. Participants who have already seen or done the survey before.
5. Participants who failed to complete the survey (duration = 0; leave questions blank).

Specific criteria

Study 1a

1. Participants who are not from the U.S.
2. Participants who failed the comprehension check question at the middle.

Comparisons and deviations

Original vs. replication

Table 1

Comparison and deviations from the original study, Study 1a

	Original	Replication	Reason for change
Study design	-	-	-
Procedure	Pairs of programs were presented in one of two randomized orders.	We randomized the order of all five pairs (Program A to D), which in theory should yield 120 possible orders.	To better control any possible order effects.
	Participants answered the verification questions after comparing the domestic and the Rwanda programs.	They answered the questions after evaluating both Rwanda and Haiti programs.	-
Condition	-	-	-

Table 2

Comparison and deviations from the original study, Study 1a

	Original	Replication	Reason for change
Study design	-	-	-
Procedure	Pairs of programs were presented in one of two randomized orders.	A and B were compared first. Then Its either A vs. C and A vs. D (the order of C and D was also randomized) or B vs. C and B vs. D.	To better control any possible order effects.
Condition	-	-	-

Deviations from pre-registration

The experiment was carried out as pre-registered. The data were analyzed, however, in a slightly different manner. We included the additional factor from our extensions in the ANOVAs we conducted (and hence 3-way repeated-measures ANOVAs were conducted instead of the original 2-way repeated-measures ANOVAs). This would provide better control for Type I error and have better ability to explore any interaction between the original factors and the new factor. The original (and pre-registered) analyses were nested within these ANOVAs.

Pre-exclusion vs. post-exclusion

Study 1a full sample results

Table 3

Study 1a full sample descriptive statistics

Refugee program location	Comparison program	Refugee camp size			
		Large		Small	
		Mean	*SD*	Mean	*SD*
Rwanda	Employment	0.09	4.33	-0.28	4.33
	Transportation	0.87	4.28	0.67	4.21
Haiti	Employment	0.04	4.42	-0.25	4.37
	Transportation	0.81	4.29	0.64	4.26

Note. Higher values indicate higher preferences for the refugee programs as compared with the employment or transportation programs.

Table 4

Three-way rmANOVA table:

Model term	dfs	MSE	*F*	*p*	η^2_p	90% CI
CP	1, 756	10.197	106.30	< .001	0.123	[0.089, 0.160]
L	1, 756	2.507	0.40	.527	0.001	[0.000, 0.007]
CS	1, 756	2.499	39.78	< .001	0.050	[0.028, 0.078]
CP × L	1, 756	1.778	0.23	.630	< 0.001	[0.000, 0.006]
CP × CS	1, 756	1.585	5.04	.025	0.007	[0.000, 0.020]
L × CS	1, 756	1.816	0.47	.493	0.001	[0.000, 0.007]
CP × L × CS	1, 756	1.622	0.13	.717	< 0.001	[0.000, 0.005]

Note. CP = comparison program, CS = camp size, L = camp location.

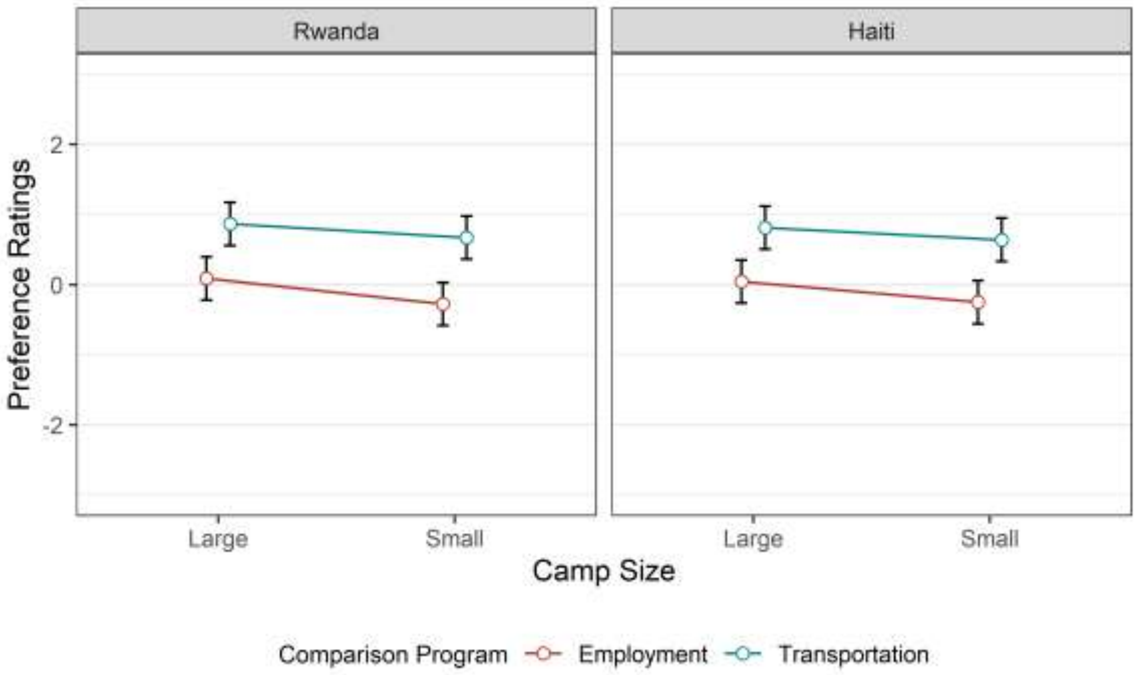


Figure 1

Interaction plot on estimated marginal means of the full sample (error bars represent 95% CIs of the estimates):

Unlike results for the sample after exclusion, the full sample results revealed a significant interaction between comparison program and camp size.

Follow-up analysis on the simple main effect of camp size on each level of comparison program revealed that participants preferred the larger camp program (*M* = 0.07) over the smaller camp program (*M* = -0.26) when they were compared to the employment program, *t*(1440) = 6.33, *p* < .001, *d* = 0.17, 95% CI [0.12, 0.22]. The same was true when the two programs were compared with the transportation program, but the effect size was smaller, *t*(1440) = 3.54, *p* < .001, *d* = 0.09, 95% CI [0.04, 0.15].

Table 5

Three-way rmANOVA table (confined analysis on those who indicated that the same number of people would be saved):

Model term	dfs	MSE	*F*	*p*	η^2_p	90% CI
CP	1, 373	9.279	61.43	< .001	0.141	[0.091, 0.196]
L	1, 373	2.385	0.20	.653	0.001	[0.000, 0.011]
CS	1, 373	2.393	20.60	< .001	0.052	[0.022, 0.094]
CP × L	1, 373	1.850	0.03	.872	< .001	[0.000, 0.005]
CP × CS	1, 373	1.572	1.43	.233	0.004	[0.000, 0.021]
L × CS	1, 373	1.741	1.29	.267	0.003	[0.000, 0.020]

CP × L × CS	1, 373	1.814	< .01	> .999	< .001	[0.000, 0.000]

Note. CP = comparison program, CS = camp size, L = camp location.

Participants preferred the larger camp program (*M* = 1.06) over the smaller camp program (*M* = 0.80), *t*(4.54), *p* < .001, *d* = 0.24, 95% CI [0.13, 0.34].

Table 6

Three-way rmANOVA table (confined analysis on those who preferred neither the larger nor the smaller camp program):

Model term	dfs	MSE	*F*	*p*	η^2_p	90% CI
CP	1, 119	8.290	6.71	.011	0.053	[0.007, 0.132]
L	1, 119	1.071	4.37	.039	0.035	[0.001, 0.105]
CS	1, 119	0.600	0.39	.533	0.003	[0.000, 0.041]
CP × L	1, 119	1.060	0.03	.876	< .001	[0.000, 0.015]
CP × CS	1, 119	0.641	0.59	.445	0.005	[0.000, 0.046]
L × CS	1, 119	0.661	2.91	.091	0.024	[0.000, 0.087]
CP × L × CS	1, 119	0.776	0.39	.535	0.003	[0.000, 0.041]

Note. CP = comparison program, CS = camp size, L = camp location.

We found a significant main effect of comparison program and one of refugee camp location. Participants preferred the Rwanda programs (*M* = -2.11) over the Haiti programs (*M* = -2.25), *t*(119) = 2.09, *p* < .039, *d* = 0.19, 95% CI [0.01, 0.37].

Overall, the comparison camp main effect (participants preferred the employment program over the transportation program) and the camp size main effect (participants preferred the larger camp program over the smaller camp program) remained largely consistent prior and after exclusion.

Study 1b full sample results

Unlike results for the sample after exclusion, the full sample results had a significant three-way interaction. The interaction between camp size and comparison program was significant when participants were not aware that 15 other countries were considering funding similar refugee programs, *F*(1, 1492.11) = 16.33, $p < .001$, $\eta^2_p = 0.011$, 90% CI [0.004, 0.021]. Nonetheless, participants still preferred the larger camp program over the smaller camp program, regardless of whether they were compared with the employment program (*M* = 0.62 vs. *M* = -0.44, *t*(2401) = 15.88, $p < .001$, *d* = 0.32, 95% CI [0.28, 0.37]) or with the transportation program (*M* = 1.93 vs. *M* = 1.17, *t*(2401) = 11.37, $p < .001$, *d* = 0.23, 95% CI [0.19, 0.27]). The interaction was not significant after participants were told that other countries were considering the programs, *F*(1, 1492.11) = 0.98, *p* = .323. Participants again preferred the larger camp program over the smaller camp program, *t*(1350) = 7.03, $p < .001$, *d* = 0.19, 95% CI [0.18, 0.25].

Table 7

Study 1b full sample descriptive statistics

Potential diffusion of responsibility	Comparison program	Refugee camp size Large Mean	*SD*	Small Mean	*SD*
No (i.e., 1st round evaluation)	Employment	0.62	3.97	-0.44	3.90
	Transportation	1.93	3.90	1.17	3.88
Yes (i.e., 2nd round evaluation)	Employment	0.11	4.00	-0.32	3.92
	Transportation	1.43	3.79	1.08	3.74

Note. Higher values indicate higher preferences for the refugee programs as compared with the employment or transportation programs.

Table 8

Three-way rmANOVA table:

Model term	dfs	MSE	*F*	*p*	η^2_p	90% CI
CP	1, 749	11.626	256.03	< .001	0.255	[0.213, 0.297]
CS	1, 749	3.051	206.43	< .001	0.216	[0.175, 0.258]
RE	1, 749	8.306	10.87	.001	0.014	[0.004, 0.032]
CP × CS	1, 749	1.098	11.90	001	0.016	[0.004, 0.034]
CP × RE	1, 749	2.235	1.82	.178	0.002	[0.000, 0.012]
CS × RE	1, 749	1.535	65.71	< .001	0.081	[0.052, 0.113]
CP × CS × RE	1, 749	0.969	4.97	.026	0.007	[0.000, 0.020]

Note. CP = comparison program, CS = camp size, R = round of evaluation.

Figure 2

Interaction plot on estimated marginal means of the full sample (error bars represent 95% CIs of the estimates), Study 1a. Generated with R.

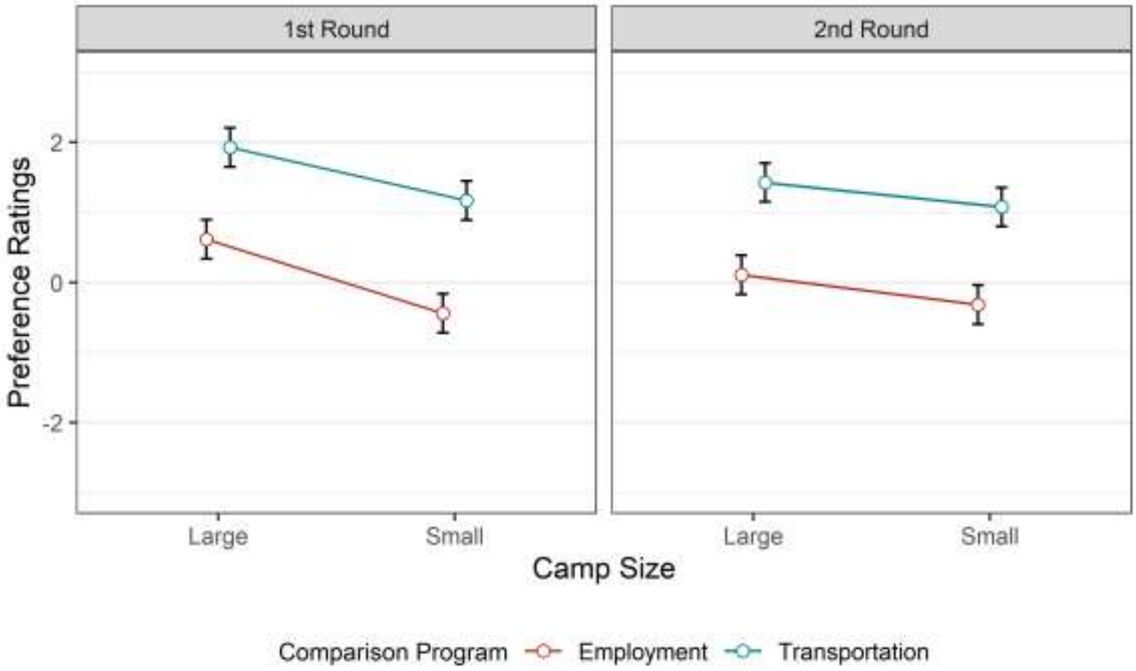


Table 9

Three-way rmANOVA table (confined analysis on those who indicated that it was the same to save lives in either camp):

Model term	dfs	MSE	*F*	*p*	η^2_p	90% CI
CP	1, 576	11.243	196.58	< .001	0.254	[0.206, 0.302]
CS	1, 576	2.779	185.83	< .001	0.244	[0.196, 0.292]
RE	1, 576	7.995	8.99	.003	0.015	[0.003, 0.036]
CP × CS	1, 576	1.004	12.22	.001	0.021	[0.006, 0.044]
CP × RE	1, 576	2.211	2.45	.118	0.004	[0.000, 0.018]
CS × RE	1, 576	1.376	50.81	< .001	0.081	[0.049, 0.119]
CP × CS × RE	1, 576	0.872	2.58	.109	0.005	[0.000, 0.018]

Note. CP = comparison program, CS = camp size, R = round of evaluation.

An analysis on the simple effect of camp size on each level of comparison program revealed that participants preferred the larger camp program over the smaller camp program, regardless of whether these programs were compared with the employment program (*M* = 0.45 vs. *M* = -

0.32, *t*(944) = 13.48, *p* < .001, *d* = 0.44, 95% CI [0.37, 0.51]) or with the transportation program (*M* = 1.73 vs. *M* = 1.17, *t*(944) = 9.88, *p* < .001, *d* = 0.32, 95% CI [0.26, 0.39]).

An analysis on the simple effect of evaluation round on each level of camp size revealed that participants' preferences for the larger camp program decreased in the second round (*M* = 1.34 vs. *M* = 0.85, *t*(768) = -5.50, *p* < .001, *d* = -0.20, 95% CI [-0.27, -0.13]), but those for the smaller camp program remained largely the same (*M* = 0.43 vs. *M* = 0.42, *t*(768) = -0.04, *p* = .969, *d* = -0.001, 95% CI [-0.072, 0.069]).

Overall, with the full sample, we still observed that participants generally preferred the larger camp program over the smaller camp program. We also observed the unexpected interaction between camp size and round of evaluation. Participants' preferences for the larger camp program were influenced by the knowledge that other countries were considering funding similar programs, which does not appear to be the case for the smaller camp program.

Study 1c – Analyses with the Preregistered Data Exclusions

Methods

Participants. We received complete responses from 451 participants recruited on Mechanical Turk (224 males, 222 females, 2 non-binary, 3 preferred not to disclose, M_{age} = 42.33, SD_{age} = 13.49), who were paid $1.00 for this task. Participants who participated in the other studies in this paper were barred from participation in this study. No participants failed the attention check at the end of the survey, by replying "Yes" to the question "Have you ever been on Jupiter". Four participants selected a value below "4" on a self-reported 5-point seriousness scale ranging from 1 (not at all serious) to 5 (very serious), measuring how serious they claimed to be when completing the survey, and were excluded from analyses. This left 447 valid participants (222 males, 220 females, 2 non-binary, 3 preferred not to disclose, M_{age} = 42.35, SD_{age} = 13.50). This sample is enough to detect a within-subjects effect size *d* = 0.40 (original effect size) at 99.9% power with alpha = .05, two-tailed (applied for all tests) and *d* = 0.13 (a small effect, following Lovakov and Agadullina 2020) at 80% power (alpha = .05, two-tailed).

Results

Replication. A two-way within-subjects ANOVA with domestic program (transportation vs. unemployment) and camp size (small vs. large) as factors found a statistically significant effect of domestic program, *F*(1, 446) = 63.47, *p* < .001, η^2_p = 0.12, such that transportation programs resulted in more favourable ratings for the aid to South Sudan; a statistically significant effect of camp size, *F*(1, 446) = 44.79, p < .001, η^2_p = 0.09, indicating a preference for aid towards the large camp size; and no statistically significant interaction between domestic program and camp size, , *F*(1, 446) = 0.01, *p* = .932, η^2 < .001 η^2_p < 0.001. When making comparisons with the domestic aid programs, participants were more likely to prefer to save 4,500 lives out of 250,000 (i.e., in the large camp, *M* = 0.27, *SE* = 0.19) than to save 4,500 lives out of 11,000 (i.e., in the small camp, *M* = -0.29, *SE* = 0.19). These results replicate the results of Study 1a and 1b, corroborating the existence of psychophysical sensitization within this paradigm.

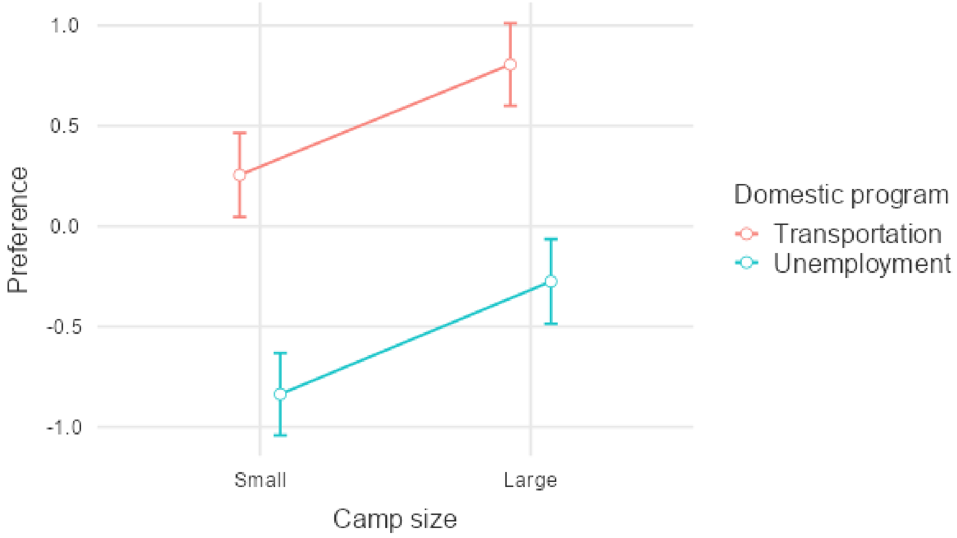


Figure 3. Preferences for aid, Study 1c. Higher values indicate a preference for the foreign aid program. Error bars indicate .± one standard error of the mean.

Extension, direct comparison. In order to directly test the notion of psychophysical numbing, we tested whether participants preferred the aid to save 4,500 lives to be sent to a camp with 11,000 total people or to a camp with 250,000 total people, by asking participants directly on one item in which -6 (negative number) indicated the strongest preference for aid to the larger camp and +6 indicated the strongest preference for aid to the smaller camp. A one-sample t-test against the scale midpoint (0) showed that participants preferred to allocate the aid towards the large camp, *M* = -3.10, *SD* = 3.28, *t*(446) = -19.94, *p* < .001, *d* = -0.94, 95% CI [-1.05, -0.83]. This extension also provides support for psychophysical sensitization and it does not provide support for psychophysical numbing.

Extension, political ideology and aid importance. We explored the correlation between our one-item measure of psychophysical sensitization, self-reported political ideology and several measures of importance of each problem tackled (unemployment, transportation, refugee crisis in South Sudan), and an imaginary $100 million budget that participants had to allocate across a program to alleviate unemployment in their country, a program to improve transportation in their country, and a program to alleviate the refugee crisis in South Sudan. As shown in Table 10, we did not find a statistically significant correlation between ideology and the one-item psychophysical sensitization measure. Participants considered unemployment and transportation programs in their countries more important than solving the refugee crisis in South Sudan. However, the extent to which they did was not correlated in a statistically significant fashion with the one-item psychophysical sensitization measure. In addition to not reaching statistical significance (the lowest p-value was .075), these correlations were all very small (the highest Pearson's r was = -.08, considered a very small effect, cf. Lovakov and Agadullina 2021). If these factors are indeed associated with psychophysical numbing or sensitization, these associations are quite small and would require a much larger sample size than the one we employed here - and a much larger budget – to be properly studied.

Table 10
Correlation of political ideology, importance, and mock allocation with psychophysical numbing one-item measure

		Correlation with psychophysical numbing
Political ideology	Pearson's r	0.034
	p-value	0.471
Importance employment program	Pearson's r	-0.062
	p-value	0.191
Importance transportation program	Pearson's r	0.082
	p-value	0.085
Importance refugee program	Pearson's r	-0.035
	p-value	0.456
Allocation to employment program	Pearson's r	-0.084
	p-value	0.075
Allocation to transportation program	Pearson's r	0.054

Table 10
Correlation of political ideology, importance, and mock allocation with psychophysical numbing one-item measure

		Correlation with psychophysical numbing
	p-value	0.253
Allocation to refugee aid	Pearson's r	0.036
	p-value	0.445

Table 11
Descriptive statistics, ideology, importance, and mock allocation, Study 1c

	Political ideology	**Importance employment**	**Importance transportation**	**Importance refugee aid**	**Budget employment**	**Budget transportation**	**Budget refugee aid**
M	3.620	4.893	3.906	3.532	45.306	29.613	25.081
SD	1.801	1.232	1.464	1.866	19.694	16.682	20.818

Note. Political ideology and program importance were measured on 0-6 scales. Participants allocated a fictitious budget of $100 million to employment and transportation programs in their own countries and refugee aid in South Sudan

Table 12
Correlation of familiarity and refugee estimation with psychophysical numbing, Study 1c

		Correlation with Psychophysical numbing	Correlation with psychophysical numbing a after removing values above 80th percentile
Familiarity Afghanistan	Pearson's r	-0.052	
	p-value	0.274	
Familiarity Syria	Pearson's r	-0.089	
	p-value	0.060	
Familiarity Venezuela	Pearson's r	0.048	
	p-value	0.309	
Familiarity South Sudan	Pearson's r	0.016	
	p-value	0.738	
Familiarity Rwanda	Pearson's r	0.016	
	p-value	0.732	
Estimation Afghanistan	Pearson's r	0.059	-0.074
	p-value	0.213	0.160
Estimation Syrian	Pearson's r	0.006	-0.019
	p-value	0.904	0.714
Estimation Venezuela	Pearson's r	0.041	-0.070
	p-value	0.385	0.184
Estimation South Sudan	Pearson's r	0.044	-0.053
	p-value	0.351	0.306
Estimation Rwanda	Pearson's r	-0.065	-0.105
	p-value	0.171	0.045

Table 13
Descriptive statistics, familiarity and refugee estimation

	Fam Afg.	Fam. Syria	Fam. Ven.	Fam. S.S.	Fam. Rwanda	Est. Afg.	Est. Syria	Est. Ven.	Est S.S.	Est. Rwa
M	1.776	2.649	1.662	1.651	1.454	6965000	26260000	15170000	10920000	475820
Mdn	2	3	1	1	1	100000	165000	100000	250000	100000
SD	1.685	1.893	1.728	1.617	1.554	119100000	376600000	284700000	212800000	1134000

Extension, familiarity and estimation. Further, we explored whether familiarity with refugee crises in Afghanistan, Syria, Venezuela, South Sudan, and Rwanda correlated with the one-item measure of psychophysical numbing. Overall, participants reported that they were not very familiar with any of the refugee crises (with perhaps the exception of Syria, which, however, still had a mean value below the scale midpoint 3). What is more important, familiarity self-reports were not significantly correlated with the extent of psychophysical sensitization, as shown in Table 12 (the lowest p-value was .060, but most of the p-values were higher than .20) and the correlations produced were quite weak (the highest correlation was Pearson's r = -.09, but most of the correlation were equal or lower than Pearson's r = .05). Similarly, the estimation of the refugee number in Afghanistan, Syria, Venezuela, South Sudan, and Rwanda were not significantly correlated with the extent of psychophysical sensitization, as shown in Table YYY (the lowest p-value was .171, but most of the p-values were higher than .20) and the correlations produced were quite weak (the highest correlation was Pearson's r = -.06, but most of the correlations were equal or lower to Pearson's r = .05). Since there was the possibility that some outliers may have severely skewed the distribution, we examined the correlation between each of the refugee estimate and the one-item measure of psychophysical numbing after excluding estimates above the 80th percentile (per each refugee estimate). We had not preregistered this decision, and it is arbitrary and possibly quite drastic, but we believed there was quite a lot of skewness and preferred to explore whether there was any correlation after excluding the largest outliers. All correlations (see Table 12) were small (all smaller than r = .11), and only one was statistically significant (Rwanda, p = .045). Since there is no clear theoretical rationale as to why this particular measure should explain our results, we consider it a false positive. Overall we could not find that knowing about or estimating a number of refugees was correlated with the one-item measure of psychophysical numbing.

Discussion

This study successfully replicates the results of Study 1a and Study 1b using a different context, but did not replicated the original study. Again, we find no evidence of psychophysical numbing, but we do find evidence of psychophysical sensitization. In an extension, we measured people's preferences for aid to the smaller or the larger camp directly, and we found strong evidence of psychophysical sensitization, and no evidence of psychophysical numbing. These results are hard to reconcile with the original study. Further, the additional measures we collected and the procedure we followed allow us to cast doubt on a number of alternative explanation. A number of exploratory analyses found no effect of political ideology, knowledge of refugee crises, estimation of refugee numbers, and relative importance of the aid programs considered on measures of psychophysical numbing/sensitization.

Sample comparisons

Table 14

Study 1a

	Fetherstonhaugh et al. (1997) Study 1	Study 1a (full sample)	Study 1a (sample after exclusion)
Sample size	54	757	386
Geographic origin	Undergraduate volunteers from two sections of an economics statistics course	US American	US American
Gender	N/A	372 males, 382 females, 3 other/rather not disclose	178 males, 205 females, 3 other/rather not disclose
Median age (years)	N/A	37	36
Average age (years)	N/A	39.51	38.92
Age range (years)	N/A	19 – 78	19 – 74
Medium (location)	In-person	Computer (online)	Computer (online)
Compensation	N/A	Nominal payment	Nominal payment
Year	Around 1997	2019	2019

Note. 18 were excluded for not understanding the study materials well, for not being serious enough about the study, or for having participated in similar studies before. 347 were excluded for failing the comprehension check at the middle of the survey. Five were excluded for not being from the U.S., and one was excluded for correctly guessing the hypothesis.

Table 15

Study 1b

	Fetherstonhaugh et al. (1997) Study 1	Study 1b (full sample)	Study 1b (sample after exclusion)
Sample size	54	750	723
Geographic origin	Undergraduate volunteers from two sections of an economics statistics course	British	British
Gender	N/A	311 males, 439 females	299 males, 424 females
Median age (years)	N/A	38	38
Average age (years)	N/A	40.47	40.48
Age range (years)	N/A	18 – 87	18 – 87
Medium (location)	In-person	Computer (online)	Computer (online)
Compensation	N/A	Nominal payment	Nominal payment
Year	Around 1997	2019	2019

Note. 14 were excluded for having unsatisfactory understanding of the English used in the study materials, for not being serious enough about the study, or for having participated in similar studies in the past. 13 were excluded for providing nonsensical answers in the funneling section or for guessing the hypotheses correctly.

Study 2 (Fetherstonhaugh, Slovic, Johnson, & Friedrich, 1997)

Replication Classification with LeBel, McCarthy, Earp, Elson, and Vanpaemel (2018)

Table 16

Classification of Replication Study 2A, based on LeBel et al. (2018)

Design facet	Replication	Details of deviation
Effect, Hypothesis	Same	
IV Construct	Same	
DV Construct	Same	
IV operationalization	Same	
DV operationalization	Same	
Population (e.g. age)	Similar	Same country, but more diverse population and higher mean age in the replication compared to the original
IV stimuli	Same	
DV stimuli	Same	
Procedural details	Similar	Very minor difference in font size. The original study had system reliability as a within-subject variable, but our replication had reliability as a between-subject variable.
Physical settings	Different	Online study vs hand-written study
Contextual variables	Similar	Different year. The original study was conducted in 1994, whereas the replication was conducted in 2019. There may be cultural changes.
Replication classification	Very close replication	

Note: IV = independent variable; DV: dependent variable

Table 17

Classification of Replication Study 2B, based on LeBel et al. (2018)

Design facet	Replication	Details of deviation
Effect, Hypothesis	Same	
IV Construct	Same	
DV Construct	Same	
IV operationalization	Same	
DV operationalization	Same	
Population (e.g. age)	Different	The sample in the replication was more diverse than university students in the original study, with a higher mean age in the replication. Original study participants were from US whereas the replication participants were from UK.
IV stimuli	Same	
DV stimuli	Same	
Procedural details	Similar	Very minor difference in font size. The original study had camp size, prior aid, and reliability as within-subject variables, but our replication had all three factors as between-subject variables
Physical settings	Different	Online study vs hand-written study
Contextual variables	Similar	Different year. The original study was conducted in 1994, whereas the replication was conducted in 2019. There may be socio-cultural changes. Moreover, the original study was conducted with American participants whereas this study was conducted with British participants. There may be minor cultural differences.
Replication classification	Close-Replication	

Note: IV = independent variable; DV: dependent variable

Comparison Between Original Study and Replications

Table 19

Difference and similarities between original study and replication study 2A

	Fetherstonhaugh, Slovic, Johnson, & Friedrich (1997)	American Amazon MTurk workers
Sample size	162	499
Geographic origin	University of Oregon students, United States	US American
Gender	Not provided	258 males, 238 females, 3 other
Median age (years)	Not provided	36
Average age (years)	Not provided	38.99
Age range (years)	Not provided, likely 17-22	20-83
Medium (location)	Not provided	Computer (online)
Compensation	Nominal payment ($4 per participant)	Nominal payment
Year	1994	2019

Table 20

Differences and similarities between original study and Study 2B

	Fetherstonhaugh, Slovic, Johnson, & Friedrich (1997)	British Prolific workers
Sample size	162	1606
Geographic origin	University of Oregon students, United States	United Kingdom British people
Gender	Not Provided	565 males, 1036 females, 5 other
Median age (years)	Not Provided	36
Average age (years)	Not Provided	38.31
Age range (years)	Not Provided	18-83
Medium (location)	Not provided	Computer (online)
Compensation	Nominal payment ($4 per participant)	Nominal payment
Year	1994	2019

Replication-Extension Experimental Design Tables

Table 21

Replication and Extension Experimental Design of Study 2A

Within-Subject IVs: IV1: size of refugee camp IV2: amount of pure-water aid before a water-purification plane was sent Between-Subject IV: IV3: reliability of the water-purification plane		IV1: small camp size (11,000) condition Participants will be presented with a scenario where the size of the refugee camp is small (i.e. 11,000) Manipulation example: Moga and Fizi having small camp size of 11,000	IV1: large camp size (250,000) condition Participants will be presented with a scenario where the size of the refugee camp is large (i.e. 250,000) Manipulation example: Uvira and Kalehe having large camp size of 250,000
IV2: Amount of prior aid (Low) The amount of pure-water aid a camp was receiving before a water-purification plane was sent is low Manipulation example: Moga and Uvira having low levels of prior aid at 5%	IV3: Low water system reliability (60%) Participants will be presented with a scenario where reliability of the equipment used to administer the aid (i.e. water system) is 60% IV3: High water system reliability (100%) Participants will be presented with a scenario where reliability of the equipment used to administer the aid (i.e. water system) is 100%	Replication Dependent Variables DV1: Benefit rating **Specific DV item:** What would be the benefit of sending this *Dash-8 plane to this camp? (on a scale of 0-8, with 0 being “extremely low benefit” and 8 being “extremely high benefit”) *Dash-8 plane is the transportation which sends the required water system to the refugee camps DV2: Send or not send **Specific DV item:** Given the benefit indicated on the scale above (referring to the benefit rating on the above), would it be worth sending the plane to this camp? (yes/no decision)	
IV2: Amount of prior aid (High) The amount of pure-water aid a camp was receiving before a water-purification plane was sent is high Manipulation example: Fizi and Kalehe having high levels of prior aid at 50% and 93% respectively	IV3: Low water system reliability (60%) Participants will be presented with a scenario where reliability of the equipment used to administer the aid (i.e. water system) is 60% IV3: High water system reliability (100%) Participants will be presented with a scenario where reliability of the equipment used to administer the aid (i.e. water system) is 100%	Extension Dependent Variable DV3: Amount of Donation **Specific DV item:** If we were to award you a pay bonus of $5 right now for this one question, how much of that would you be willing to donate to this refugee camp? from $0 to $5	

Table 22

Replication and Extension Experimental Design of Study 2B

	IV1: small camp size (11,000) condition Participants will be presented with a scenario where the size of the refugee camp is small (i.e. 11,000) Manipulation example: Moga: a small camp size of 11,000		IV1: large camp size (250,000) condition Participants will be presented with a scenario where the size of the refugee camp is large (i.e. 250,000) Manipulation example: Uvira: a large camp size of 250,000	
Between-Subject IVs: IV1: size of refugee camp IV2: amount of pure-water aid before a water-purification plane was sent IV3: reliability of the water-purification plane	IV3: Low water system reliability (60%) Participants will be presented with a scenario where reliability of the equipment used to administer the aid (i.e. water system) is 60%	IV3: High water system reliability (100%) Participants will be presented with a scenario where reliability of the equipment used to administer the aid (i.e. water system) is 100%	IV3: Low water system reliability (60%) Participants will be presented with a scenario where reliability of the equipment used to administer the aid (i.e. water system) is 60%	IV3: High water system reliability (100%) Participants will be presented with a scenario where reliability of the equipment used to administer the aid (i.e. water system) is 100%
IV2: Amount of prior aid (Low) The amount of pure-water aid a camp was receiving before a water-purification plane was sent is low Manipulation example: Moga and Uvira having low levels of prior aid at 5%	Replication Dependent Variables DV1: Benefit rating **Specific DV item:** What would be the benefit of sending this *Dash-8 plane to this camp? (on a scale of 0-8, with 0 being “extremely low benefit” and 8 being “extremely high benefit”) *Dash-8 plane is the transportation which sends the required water system to the refugee camps			
IV2: Amount of prior aid (High) The amount of pure-water aid a camp was receiving before a water-purification plane was sent is high Manipulation example: Fizi and Kalehe having high levels of prior aid at 50% and 93% respectively	DV2: Send or not send **Specific DV item:** Given the benefit indicated on the scale above (referring to the benefit rating on the above), would it be worth sending the plane to this camp? (yes/no decision) Extension Dependent Variable DV3: Donation **Extension DV item:** What percentage of your earnings would you be willing to donate?			

Materials, Procedures and Scales in the Study 2A

Procedure

An online Qualtrics survey will be used for this replication study for data collection. The Qualtrics survey consisted of the following items.

(1) Participants first completed a consent form

(2) Then they read an introduction about the survey

“The main survey consists of a total of 4 scenarios, with 4 comprehension questions and 3 decision questions each. There are 28 questions in total, not including final wrap up and demographics questions.

Please note: this study involved comprehension questions.

The scenarios are fairly similar but differ on several key parameters. It is important you pay attention to these factors (bolded to make things clear). Please read the descriptions and questions carefully, and note that your responses to each scenario should be independent of your responses to other scenarios.”

(3) They read a cover story about the Rwandan refugee crisis

“The U.N. High Commissioner for Refugees was coordinating a massive humanitarian aid campaign by requesting that able countries send assistance to the Rwandan refugees in Zaire. Many refugees had a water-borne disease and would die if purified water did not soon become available. One small country was considering sending one of two Dash-8 water-purification planes to Zaire. Although each water system was capable of producing only a small fraction of the water needed, each could keep about 1500 disease victims alive each day. Once a plane was operating in a camp, aid-workers will distribute the clean water to designated disease victims, which usually saves the victims' lives.

The cost to this small country of delivering and operating these purification systems is significant in light of its economy.”

(4) Participants underwent 4 scenarios, including four refugee camps (Moga, Fizi, Uvira and Kalehe), with identical structure, in 4 separate pages. Each scenario differed in camp, camp size, water system reliability, amount of prior and post aid. They were randomized into two blocks, a) 100% reliability block, or b) 60% reliability block, each with 4 scenarios:

Table 23

List of the Eight Scenarios

Scenario	Refugee camp	Camp size	Water system reliability	Amount of prior aid	Amount of post aid
1	Moga 1	11,000	100%	5%	50%
2	Moga 2	11,000	60%	5%	50%
3	Fizi 1	11,000	100%	50%	95%
4	Fizi 2	11,000	60%	50%	95%
5	Uvira 1	250,000	100%	5%	7%
6	Uvira 2	250,000	60%	5%	7%
7	Kalehe 1	250,000	100%	93%	95%
8	Kalehe 2	250,000	60%	93%	95%

5) For each of the scenarios, participants have to answer:

- ○ “How many refugees are now in the city?”
  - ■ Choose 1 from 4 options
    1) 11,000
    2) 250,000
    3) 1,000,000
    4) None of the other options
- ○ “How much of the clean water needed for disease victims in this camp is currently being met (prior to receiving aid)?”
  - ■ Choose 1 from 4 options
    1) 5%
    2) 50%
    3) 93%
    4) None of the other options
- ○ “How much of the clean water needed for disease victims in this camp would be met if aid is given (post aid)?”
  - ■ Choose 1 from 4 options
    1) 50%

2) 95%
3) 7%
4) None of the other options

- “What is the reliability of the water system considered to be sent?”
    - Choose 1 from 4 options
        1) 100%
        2) 60%
        3) 0%
        4) None of the other options
- “What would be the benefit of sending this Dash-8 plane to this camp?”
    - Nine-point Likert scale
    - From 0 (“extremely low benefit”) to 8 (“extremely high benefit”)
- “Given the benefit indicated on the scale above, would it be worth sending the plane to this camp?”
    - Choose either “Yes” or “No”
- “If we were to award you a pay bonus of $5 right now for this one question, how much of that would you be willing to donate to this refugee camp?”
    - From $0 to $5

b) Descriptions of 4 scenarios from the 60%-reliability block. For each of the scenarios, participants have to answer:

- “How many refugees are now in the city?”
    - Choose 1 from 4 options
        1) 11,000
        2) 250,000
        3) 1,000,000
        4) None of the other options
- “How much of the clean water needed for disease victims in this camp is currently being met (prior to receiving aid)?”
    - Choose 1 from 4 options
        1) 5%
        2) 50%
        3) 93%
        4) None of the other options
- “How much of the clean water needed for disease victims in this camp would be met if aid is given (post aid)?”
    - Choose 1 from 4 options
        1) 50%

        2) 95%
        3) 7%
        4) None of the other options

- "What is the reliability of the water system considered to be sent?"
    - Choose 1 from 4 options
        1) 100%
        2) 60%
        3) 0%
        4) None of the other options
- "What would be the benefit of sending this Dash-8 plane to this camp?"
    - Nine-point Likert scale
    - From 0 ("extremely low benefit") to 8 ("extremely high benefit")
- "Given the benefit indicated on the scale above, would it be worth sending the plane to this camp?"
    - Choose either "Yes" or "No"
- "If we were to award you a pay bonus of $5 right now for this one question, how much of that would you be willing to donate to this refugee camp?"
    - From $0 to $5

(5) They completed the comprehension check to ensure that they understand the number of lives saved is the same across both camps.

- "Please choose the statement you find the most appropriate."
    - Choose whether the water system saves the same number of lives, more lives in the larger camps, or more lives in the smaller camps

(6) After that, they completed the funneling section.

- "How serious were you in filling out this questionnaire?"
    - From 1 (Not at all) to 5 (Very much)
- "Have you ever seen the materials used in this study or similar before? If yes, please indicate where."
    - Choose either "Yes" or "No"
- "What do you think the purpose of the study was?"
- "Did you spot any errors? Anything missing or wrong? Something we should pay attention to in next runs?"
- "Please rate your satisfaction with the pay/compensation offered for this MTurk HIT."
    - From 0 (Extremely satisfied) to 6 (Very satisfied)

(7) They filled in the demographic questions

- Age

- ○ Gender
- ○ Country of origin
- ○ Family’s social class
- ○ English proficiency

(8) They read the debriefing statement.

Materials, Procedures and Scales in Study 2B

Procedure

An online Qualtrics survey will be used for this replication study for data collection. The Qualtrics survey consisted of the following items.

1. Participants first completed a consent form
2. Then they read an introduction about the survey

"This study is about a decision making regarding situations involving a refugee crisis.

Please note: this study involves comprehension questions.

You will read the general background, then presented with a very brief and clear scenario. You'll answer 4 comprehension questions and 3 follow-up evaluation questions."

3. They read a cover story about the Rwandan refugee crisis

"The U.N. High Commissioner for Refugees is coordinating a massive humanitarian aid campaign by requesting that able countries send assistance to the Rwandan refugees in Zaire. Many refugees have a water-borne disease and will die if purified water does not soon become available.

One small country is considering sending one of the two the Dash-8 water-purification planes to Zaire. Although each water system is capable of producing only a small fraction of water needed, each can keep about 1500 disease victims alive each day.Once a plane is operating in a camp, aid-workers will distribute the clean water to designated disease victims, which usually saves the victims' lives.

It should be noted that the cost to this small country of delivering and operating these purification systems is significant in light of its economy."

4. We presented one of the eight scenarios (i.e. Mogo 1, Fizi 1, Uvria 1, Kahele 1, Mogo 2, Fizi 2, Uvira 2, and Kahele 2.) to participants. The sequence of the scenarios shown to them are randomized, using the Qualtrics' "Randomizer" function as well as the "Evenly Present Elements" function, which can be selected under the tab "Survey Flow".

Table 24

List of the Eight Scenarios

Scenario	Refugee camp	Camp size	Water system reliability	Amount of prior aid	Amount of post aid
1	Moga 1	11,000	100%	5%	50%
2	Moga 2	11,000	60%	5%	50%
3	Fizi 1	11,000	100%	50%	95%
4	Fizi 2	11,000	60%	50%	95%
5	Uvira 1	250,000	100%	5%	7%
6	Uvira 2	250,000	60%	5%	7%
7	Kalehe 1	250,000	100%	93%	95%
8	Kalehe 2	250,000	60%	93%	95%

5) For each of the scenarios, participants have to answer:

- ○ “How many refugees are now in the city?”
  - ■ Choose 1 from 4 options
    1) 11,000
    2) 250,000
    3) 1,000,000
    4) None of the other options
- ○ “How much of the clean water needed for disease victims in this camp is currently being met (prior to receiving aid)?”
  - ■ Choose 1 from 4 options
    1) 5%
    2) 50%
    3) 93%
    4) None of the other options
- ○ “How much of the clean water needed for disease victims in this camp would be met if aid is given (post aid)?”
  - ■ Choose 1 from 4 options

        1) 50%
        2) 95%
        3) 7%
        4) None of the other options

- ○ "What is the reliability of the water system considered to be sent?"
    - ■ Choose 1 from 4 options
        1) 100%
        2) 60%
        3) 0%
        4) None of the other options
- ○ "What would be the benefit of sending this Dash-8 plane to this camp?"
    - ■ Nine-point Likert scale
    - ■ From 0 ("extremely low benefit") to 8 ("extremely high benefit")
- ○ "Given the benefit indicated on the scale above, would it be worth sending the plane to this camp?"
    - ■ Choose either "Yes" or "No"
- ○ "Suppose that you were given the option to donate some or all of the pay received for this tasks to support the above described refugee camp. What percentage of your earnings would you be willing to donate?"
    - ■ From 0 to 100

6) They completed the comprehension check to ensure that they understand the number of lives saved is the same across both camps.

- ○ "Please choose the statement you find the most appropriate."
    - ■ Choose whether the water system saves the same number of lives, more lives in the larger camps, or more lives in the smaller camps

7) After that, they completed the funneling section.

- ○ "How serious were you in filling out this questionnaire?"
    - ■ From 1 (Not at all) to 5 (Very much)
- ○ "Have you ever seen the materials used in this study or similar before? If yes, please indicate where."
    - ■ Choose either "Yes" or "No"
- ○ "What do you think the purpose of the study was?"
- ○ "Did you spot any errors? Anything missing or wrong? Something we should pay attention to in next runs?"
- ○ "Please rate your satisfaction with the pay/compensation offered for this MTurk HIT."
    - ■ From 0 (Extremely satisfied) to 6 (Very satisfied)

8) They filled in the demographic questions

- ○ Age
- ○ Gender
- ○ Country of origin
- ○ Family’s social class
- ○ English proficiency

9) They read the debriefing statement.

Results Summary and Interpretation based on Lebel et al. (2019)

Table 25

Summary of Mixed-Design ANOVA and Confidence Intervals of Camp Size and Prior Aid on DV1 - Beneficial Ratings & DV2 - Yes/No Decision to Send a Plane in Study 2A

	F	df	*p*	Sum of Square	η^2p and 90% CI	Interpretation
Within-Subject Contrast between Large Camp Size and Small Camp Size (*n* = 499)						
Rated Benefits	821.82	1, 497	< .001	3813.54	0.62 [0.58, 0.66]	Signal Consistent Similar Successful Replication
Sending Plane Decision	392.88	1, 497	<.001	78.60	0.44 [0.39, 0.49]	Signal Inconsistent Larger Successful Replication
Within-Subject Contrast between Low Prior Aid and High Prior Aid (*n* = 499)						
Rated Benefits	43.10	1, 497	<.001	90.00	0.08 [0.05, 0.12]	Signal Consistent Similar Successful Replication
Sending Plane Decision	0.05	1, 497	=.827	0.00	0.00 [0.00, 0.01]	No-signal Consistent Successful Replication
Between-Subject Contrast between 100% Reliability (*n* = 259) and 60% Reliability (*n* = 240)						
Rated Benefits	5.70	1, 497	=.017	43.89	0.01 [0.00, 0.03]	Signal Inconsistent Smaller Successful Replication
Sending Plane Decision	1.06	1, 497	=.303	0.24	0.00 [0.00, 0.01]	No-signal Inconsistent Failed Replication

Note. Mixed ANOVA, *N* = 499, see Study2A_AfterExclusion_DataAnalysis_YSK_V2.omv for code and statistics. CI for partial eta squared = 90% confidence intervals.

Table 26

Summary of Between-Design ANOVA and Confidence Intervals of Camp Size and Prior Aid on DV1 - Beneficial Ratings & DV2 - Yes/No Decision to Send a Plane in Study 2B

	F	df	p	Sum of Square	η^2p and CI	Interpretation
Between-Subject Contrast between Large Camp Size (*n* = 791) and Small Camp Size (*n* = 815)						
Rated Benefits	511.60	1,1598	*p*<.001.	1460.70	0.24 [0.21, 0.28]	Signal Inconsistent Smaller Successful Replication
Sending Plane Decision	9.10	1,1598	*p*=.003	25.98	0.01 [0.00, 0.02]	Signal Inconsistent Smaller Successful Replication
Between-Subject Contrast between Low Prior Aid (*n* = 806) and High Prior Aid (*n* = 800)						
Rated Benefits	9.10	1,1598	*p*=.003	0.27	0.01[0.00, 0.02]	Signal Inconsistent Smaller Successful Replication
Sending Plane Decision	25.98	1,1598	*p*=.115	0.27	0.00 [0.00, 0.01]	No signal Consistent Successful Replication
Between-Subject Contrast between 100% Reliability (*n* = 819) and 60% Reliability (*n* = 787)						
Rated Benefits	88.37	1,1598	*p*< .001	252.31	0.05 [0.03, 0.08]	Signal Consistent Similar Successful Replication
Sending Plane Decision	2.46	1,1598	*p*=.117	0.27	0.00 [0.00, 0.00]	No-signal Inconsistent Failed Replication

Note. [Between-Subject ANOVA, with Independent-Sample t-test], *N* =1,606. CI for partial eta squared = 90% confidence intervals.

Study 2A Pre-registration plan versus final report

Table 27

Preregistration Planning and Deviation Documentation (PPDD) of Study 2A

Components in your preregistration	Location of preregistered decision/plan	Were there deviations? What type?	If yes - describe details of deviation(s)	Rationale for deviation	How might the results be different if you had/had not deviated	Date/time of decision for deviation + stage
Study design	Folder for Pre-registered plan and Materials	Major	Pre-registration: 2x2x2 within-subject design Final: 2x2x2 mixed design	Low tolerance of MTurk participants	Participants may become inpatient or may not finish the survey, if we used within-subject design, which takes longer.	Data collection
Measured variables		Minor	For the extension question Pre-registration: "How much are you willing to donate to this camp if you are given $5?" Final: "If we were to award you a pay bonus of $5 right now for this one question, how much of that would you be willing to donate to this refugee camp?"	for participants to understand the question better	Our original question may result in misunderstandings. As participants understand the question better, validity of their responses increase.	Drafting Qualtrics survey before data collection

Exclusion criteria	Minor	Four additional comprehension questions were added in each scenario	ensure that participants were aware of the situation in each scenario	Not including the comprehension question may result in inclusion of participants who don’t fully understand the scenarios. Validity of responses increase.	Drafting Qualtrics survey before data collection
Data analysis	Major	Pre-registration: i) use of randomly generated dataset. Ii) only reported Cohen’s d. Final Manuscript: i) use of real dataset generated from MTurkers’ responses, ii) reported both Cohen’s d and partial eta squared, iii) for plane decision, in the main manuscript, we reported binomial logistic regression statistics instead of ANOVA and t-test results.	1) Usage of randomly generated dataset is well recognized for Stage 1 RR, to test the code and analyses method. We replaced the simulated statistics with real statistics 2) Partial eta squared was added for ANOVA main effect, which explains the amount of variance a fixed factor explains. 3) The original article used ANOVA for plane decisions, but it is generally better to analyze binary decisions with binomial logistic regression.	No substantial impact on the results with different statistics.	Final Analysis

Study 2B Pre-registration plan versus final report

Preregistration Planning and Deviation Documentation (PPDD) of Study 2B

Components in preregistration	Location of preregistered plan	Were there deviations?	If yes - describe details of deviation(s)	Rationale for deviation	How might the results be different if you had/had not deviated	Date/time of decision for deviation + stage
Study design	Folder for Pre-Registered Plan and Materials	Major	Changed from within-subject design to between-subject design	Low tolerance of the online participants	The data may be unreliable, and some participants may not be attentive given a very long survey.	During the data collection
			Sample size: Our target sample size was 405. However, we recruited 2020 participants at the end, in which 1606 were included.	We changed from within-subject design to between-subject design, so we need a larger sample to detect the effect.	If we use the initial target sample size, the effects may not be detected in a between-subject design.	
Measured variables		Minor	Changed from "How much would you be willing to donate to this camp (in USD)?" to "What percentage of your earnings would you be willing to donate?"	More reliable.	If we use amount of donation in USD, the findings may be confounded by earnings of participants	Before data collection
			For second follow-up question (i.e., "Given the benefit indicated on the scale above, would it be worth sending the plan to this camp?"), the order of the choices (i.e., "Yes" or "No") has been randomised		The changes to randomization of options likely do not bring any substantial change to the results.	
Exclusion criteria		Minor	Added four special exclusion criteria based on the comprehension questions	Due to the change of study design	The psychophysical numbing effect measured might be less significant, since the data of the people who have understood the questions incorrectly were retained	Before data collection

Data analysis		Major	Pre-registration: i) use of randomly generated dataset. ii) only reported Cohen's d, iii) Planned to report ANOVA for plane decision Final Manuscript: i) use of real dataset generated from MTurkers' responses, ii) reported both Cohen's d and partial eta squared, iii) We changed to binomial logistic regression analyses for plane decision.	1) Usage of randomly generated dataset is well recognized for Stage 1 RR, to test the code and analyses method. We replaced the simulated statistics with real statistics 2) Cohen's d and t-statistics were added for t-test pairwise comparison of conditions. 3) For binary decisions, binomial logistic regression is more suitable than ANOVA.	No substantial impact or difference in the results	Final Analysis

Note. *Categories for deviations: Minor - Change probably did not affect results or interpretations; Major - Change likely affected results or interpretations. Preregistration Planning and Deviation Documentation (PPDD) document (Van 't Veer et al., 2019) for latest updates

Original versus Replications

Table 28

Original Versus Replication Study 2A and Study 2B Comparison

	Original	Replication	Reason for change
Study Design	2 x 2 x 2 within-subject: Participants answer questions for all of the 8 scenarios in the 100% and 60%-reliability blocks	2A: 2 x 2 x 2 mixed-design: Participants answer questions only for 4 scenarios either in the 100% or 60%-reliability block 2B: 2 x 2 x 2 between-subject design. Participants underwent only 1 of the 8 scenarios. They were randomized to one of the scenarios.	Low tolerance of MTurkers
Sample Size	162 students from the University of Oregon	Study 2A: 821 participants from Amazon Mechanical Turk, 499 included in final analysis Study 2B: 2020 participants from British Prolific, 1606 included in final analysis	MTurk and Prolific samples are more diverse and representative of the population, compared to university student samples (check Ziano, Mok, & Feldman, 2020 for justifications of using MTurk and Prolific samples). Maximize statistical power. For both studies, we initially aimed for 405 participants, which is 162 X 2.5. However, as we changed the designs to between-subject design and within-subject design, we recruited more participants.
Material	2 questions in each scenario	1) An extension and scenario comprehension checks are added in this study	Extension was added to further investigate on the psychophysical numbing effect. Additional comprehension checks were added to ensure participants had fully understood the scenarios.

Exclusion Criteria of Replications

Generalized exclusion criteria

We exclude below participants for Both Study 2A and Study 2B:

1. Participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale)
2. Participants who self-report not being serious about filling in the survey (self-report < 4, on a 1-5 scale).
3. Participants who correctly guessed the hypothesis of this study in the funnelling section.

4. Participants who have already seen or done the survey before.
5. Participants who failed to complete the survey. (duration = 0, leave question blank)
6. (When target sample is MTurk:) Participants not from United States.
7. 18 years ago or below.

Specific criteria for Study 2A

Same with the original study, a comprehension check will be added after participants complete all tasks regarding the eight scenarios.

Participants will be asked to choose a statement they find the most appropriate:

(1) The water system saves about the same number of lives regardless of camp sizes;

(2) The water system saves more lives in the larger camps; and

(3) The water system saves more lives in the smaller camps.

Participants who choose the third statement will be excluded from data analysis as their responses may be biased due to their wrong perception that the intervention is able to save more lives, instead of psychophysical numbing.

Specific criteria for Study 2B

Other than the criteria mentioned above, four additional criteria were also added to assist the process of excluding relevant data. This criteria were accessed based on four questions which check whether the participants have correctly understood the scenario assigned to them.

1. Getting the wrong answer for “How many refugees are now in the city?” (e.g., For the Moga 1 scenario, if the participants answered the option other than “11,000”, then their data would be excluded)
2. Getting the wrong answer for “How much of the clean water needed for disease victims in this camp is currently being met (prior to receiving aid)?” (e.g., For the Moga 1 scenario, if the participants answered the option other than “5%”, then their data would be excluded)
3. Getting the wrong answer for “How much of the clean water needed for disease victims in this camp would be met if aid is given (post aid)?” (e.g., For the Moga 1 scenario, if the participants answered the option other than “50%”, then their data would be excluded)
4. Getting the wrong answer for “What is the reliability of the water system considered to be sent?” (e.g., For the Moga 1 scenario, if the participants answered the option other than “100%”)