



Moral framing effects within subjects

Paul Rehren & Walter Sinnott-Armstrong

To cite this article: Paul Rehren & Walter Sinnott-Armstrong (2021) Moral framing effects within subjects, *Philosophical Psychology*, 34:5, 611-636, DOI: [10.1080/09515089.2021.1914328](https://doi.org/10.1080/09515089.2021.1914328)

To link to this article: <https://doi.org/10.1080/09515089.2021.1914328>



Published online: 19 Apr 2021.



Submit your article to this journal [↗](#)



Article views: 80



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

ARTICLE



Moral framing effects within subjects

Paul Rehren ^a and Walter Sinnott-Armstrong^{a,b}

^aEthics Institute, Department of Philosophy and Religious Studies, Utrecht University, Utrecht, The Netherlands; ^bDepartment of Philosophy, Duke University, Durham, USA

ABSTRACT

Several philosophers and psychologists have argued that evidence of moral framing effects shows that many of our moral judgments are unreliable. However, all previous empirical work on moral framing effects has used between-subject experimental designs. We argue that between-subject designs alone do not allow us to accurately estimate the extent of moral framing effects or to properly evaluate the case from framing effects against the reliability of our moral judgments. To do better, we report results of our new within-subject study on four types of moral framing effects, and we discuss the implications of our findings for the reliability of moral judgments. Overall, our results strengthen the evidence from moral framing effects against the reliability of some of our moral judgments.

ARTICLE HISTORY

Received 21 December 2019

Accepted 28 January 2021



KEYWORDS

Framing effects; moral judgment; unreliability; between-subject design; within-subject design

1. Introduction

Over the past decade, the phenomenon of framing effects on moral judgments has received lots of attention from philosophers. In this paper, we will adopt the definition given by Demaree-Cotton (2016), according to which moral framing effects occur when “morally irrelevant differences in the way a scenario is presented affect people’s moral intuitions regarding that scenario” (p. 1).¹ We will, however, refer to moral judgments instead of intuitions, because the relevant experiments almost never provide evidence that participants’ responses express real moral intuitions (Bengson, 2013).

The primary examples of framing effects in the literature vary the wording and the order of moral scenarios (for a review, see Demaree-Cotton, 2016).² Word framing effects involve descriptions of the same scenario using different but morally equivalent language. To the extent that people’s moral judgments are affected by morally irrelevant differences in wording, those people are susceptible to a word framing effect. To date, at least two types of word framing effect have been reported:

CONTACT Paul Rehren  p.rehren@uu.nl  Ethics Institute, Department of Philosophy and Religious Studies, Utrecht University, Utrecht, The Netherlands

Save vs. Kill: Petrinovich and O'Neill (1996, Exp. 1) reported that some participants indicated that they would do one act when the act was described in terms of how many people it would save (Save condition) but would not do that act when the act was described in terms of how many people it would kill (Kill condition), even though they had been told that the act would have both effects.

Actor vs. Observer: The scenario refers to the agent either in the second person ("You do the act"), which we call the Actor condition, or in the third person ("She or he does the act"), which we call the Observer condition. Using a standard trolley dilemma, Nadelhoffer and Feltz (2008) found that a larger proportion of participants judge an action as permissible in the Observer condition. Similar results were reported on a different sacrificial dilemma for judgments of moral obligation (Tobia et al., 2013, Exp. 1). Curiously, in a sample of professional philosophers, Tobia et al. find opposite effects on both scenarios: participants were more likely to judge the action as permissible/obligatory in the Actor condition than in the Observer condition (Exp. 2).

Order framing effects involve a pair of scenarios in both of which optional actions have the same consequences (such as 1 versus 5 deaths) but produce those consequences in different ways. What is manipulated between the different framing conditions is the order in which the pair is presented. Two kinds of scenario pairs have been researched, giving rise to two more types of moral framing effects:

Action then Omission vs. Omission then Action ($A \rightarrow O$ vs. $O \rightarrow A$): As the name suggests, in one scenario, an outcome is brought about by an action, while in the other scenario, that same outcome is brought about by an omission. Across two studies, Haidt and Baron (1996, Exp. 1, 2) found that when Omission is presented before Action ($O \rightarrow A$), compared with the reverse order ($A \rightarrow O$), a larger proportion of subjects rate the actor as less moral in the action scenario compared to the omission scenario. Schwitzgebel and Cushman (2012) compared ratings of moral badness on the omission scenario across the different framing conditions. Their study included two Action-Omission scenario pairs. For one pair, they found that subjects gave higher moral badness ratings on the omission scenario in the $A \rightarrow O$ condition than in the $O \rightarrow A$ condition. For the second pair, they found the reverse.

Means then Side-effect vs. Side-effect then Means ($M \rightarrow SE$ vs. $SE \rightarrow M$): In one of the scenarios, an outcome is brought about as a means to an end; in the other, it is brought about as a side-effect. Previous literature (Lanteri et al., 2008; Lombrozo, 2009; Schwitzgebel & Cushman, 2012; Wiegmann & Waldmann, 2014, Exp. 1, 4, 5) unanimously reports that a hypothetical action in the side-effect scenario is judged more harshly (e.g., as less morally permissible, less morally good, etc.) when the means scenario is presented first ($M \rightarrow SE$) than when it is presented second ($SE \rightarrow M$).³

Notice that while the differences between action and omission and between means and side-effect are often seen as morally relevant, the order in which two scenarios are presented is not relevant to whether the acts in the two scenarios are morally right or wrong.

Several philosophers and psychologists have used these kinds of framing effects to cast doubt on our moral judgments (e.g., Andow, 2016; Nadelhoffer & Feltz, 2008; Schwitzgebel & Cushman, 2012; Sinnott-Armstrong, 2008). The argument for this claim is rather straightforward: When people make (two or more) incompatible moral judgments about the same scenario merely because that scenario is presented in different but morally equivalent ways, at most one of those moral judgments is correct. Yet processes that fail to consistently lead to correct judgments are unreliable, by definition. Therefore, framing effects show that the processes behind moral judgments are unreliable for scenarios in which they occur. If a large enough proportion of moral judgments are affected by framing effects, then the reliability of moral judgments in general would be in doubt.

Critics often respond by denying that framing effects are widespread enough to support any conclusion about moral judgments in general (e.g., Shafer-Landau, 2008). The crucial question, then, is not *whether*, but *how often* and *when* moral framing effects occur. These questions can be answered only by careful empirical research.

Unfortunately, previous work on moral framing effects has followed the trend in non-moral framing effects research (see Piñon & Gambara, 2005; Steiger & Anton, 2018) in employing mostly *between-subject* experimental designs. In this kind of design, experimenters split participants into groups, assign each framing condition to a different group, and compare group responses. Crucially, each participant receives only one frame. The alternative is a *within-subject* experimental design, in which each participant receives all framing conditions over the course of the experiment. The experimenter then compares the responses that the same participants gave in the different framing conditions.

Both designs have advantages and disadvantages (Charness et al., 2012). However, for the purpose of estimating the frequency of moral framing effects and assessing the reliability of our moral judgments, it is inadvisable to rely solely upon between-subject designs. In the next section, we point out a number of reasons for this advice.

2. Why use a within-subject design?

There is disagreement about precisely how framing effects are supposed to show the unreliability of moral judgments. Some authors have proposed a counterfactual account, according to which evidence of framing effects shows that some people *would have given* a different response from the one

they actually gave, *had they received* a different frame than the one they in fact received (e.g., Demaree-Cotton, 2016, pp. 3–5; Sinnott-Armstrong, 2008, pp. 52–54). Because the frames are morally equivalent and the resulting moral judgments are incompatible, framing effects show that the processes underlying these moral judgments are affected by irrelevant factors. These processes and the moral judgments they produce are therefore unreliable.

Other authors seem to have in mind a somewhat different version of the argument that avoids counterfactuals. According to them, evidence of framing effects suggests that some people *will in fact* endorse different moral judgments at different times when they encounter different frames (e.g., Nadelhoffer & Feltz, 2008, pp. 136–37). This account implies the first, but it is more demanding in a crucial way. On the counterfactual account, moral judgments are unreliable even if framing affects only the initial formation of a moral judgment, but no subsequent moral judgments. This might happen because after forming the initial moral judgment, the person strives to remain consistent (Campbell & Kumar, 2012) or because they update their judgment in light of new evidence (Horne & Livengood, 2017). After all, that person would still have formed a different moral judgment if they had initially been presented with a different frame. Thus, on the counterfactual account, their judgment would be unreliable.

Things look different on the second account, however, because the person would not make different moral judgments at different points in time depending on the frame if they made the same moral judgment that they initially formed in the first trial on all subsequent trials. On the second account, then, framing does not show unreliability unless it affects people's moral judgments beyond the initial trial in repeated encounters with the same scenario in different frames.

In our view, both accounts warrant some skepticism toward our moral judgments, but repeated effects of frame allow for a stronger case. This is because framing effects that occur only initially but are not repeated later have less severe implications for what number of people are unreliable. To see why, suppose on a certain moral scenario, 20% of people are vulnerable to framing effects. If this effect is repeated, then every time someone encounters that moral scenario, there is a 20% chance that her judgment cannot be trusted because it is affected by framing. This means that in the case of repeated framing, 20% of people being vulnerable to framing effects on a scenario implies that 20% of people's judgments on that scenario will be unreliable. The same is not true for initial framing. In that case, we need only to be worried if a person's moral judgment was affected by framing the first time she encountered the scenario. Yet even if framing is prevalent in moral life, not every moral problem that people encounter is always going to be presented in some distorting frame (Tolhurst, 2008, pp. 80–81). This

means that there is a non-zero chance that her initial moral judgment of the scenario was not problematically affected by framing, meaning that all of her subsequent judgments of the scenario are going to be reliable (or at least not unreliable due to any effect of framing). Thus, in the case of initial framing, the percentage of people whose judgment of the scenario is unreliable will generally be lower than 20%.

Unfortunately, between-subject designs fail to differentiate between initial and repeated framing effects. Participants in between-subject designs respond in a single framing condition, so data from such experiments cannot show how individual subjects would respond if they were to receive a different frame at some other time. In contrast, subjects in within-subject designs receive the entire set of framing conditions over the course of the study. Therefore, their responses to the different frames at different times can be directly compared. In this way, within-subject designs avoid a crucial limitation for between-subject designs.

Ultimately, no matter which of the two views turns out to be right, moral framing effects provide *some* evidence for *some* unreliability in our moral judgments. But *how much* unreliability? To discredit our moral judgments, it will not be enough to point to just any very small bit of unreliability, no matter how small (Demaree-Cotton, 2016, p. 3; Shafer-Landau, 2008, p. 86). What the argument needs, then, is an estimate of the proportion of people and of moral judgments that are affected by framing effects. The choice of experimental design – between-subject versus within-subject – bears heavily on this issue as well.

Demaree-Cotton (2016) valiantly tried to provide such an estimate. Based on a meta-analysis of the literature, she calculated that on average 20% of the moral judgments that had been investigated were vulnerable to framing effects. However, this estimate is subject to at least two complications.

First, any between-subject result is compatible with a range of estimates for the proportion of participants or judgments that are unreliable (Hershey & Paul, 1980). Suppose that a study shows that 60% of participants rate an action as impermissible in frame A, while 80% of participants rate the same action as impermissible in frame B. What could be going on here is that 20% of participants are unreliable: minimally, 20% of participants judged the act impermissible in B, but would have made a contrary judgment in A. However, the same results are consistent with much higher levels of underlying unreliability. For instance, 20% of participants might have judged the action as impermissible in A but would have judged it permissible in B, while another 40% rated the action as permissible in A but would have judged it impermissible in B. In that case, not just 20%, but 60% of participants are subject to framing effects and have unreliable moral judgments. Any other proportion between 20% and 60% is similarly possible.

The point, then, is that people's moral judgments may be affected by framing in more complex ways than what can be captured by between-subject designs. Thus, since the studies that Demaree-Cotton analyzes all use between-subject designs, her estimate of 20% is only the minimum rate of framing effects among the population, while the actual rate might be much higher.

The use of a within-subject design could provide a better estimate of the actual rate of framing effects. Because multiple frames are presented to each participant, that experimental design can directly investigate how the moral judgments of individual participants change between different frames. Therefore, a within-subject design could be able to reveal the more complex ways in which participants' moral judgments are affected by framing.

A second complication is that Demaree-Cotton arrives at her estimate of 20% by averaging over results for a number of different types of framing effects. However, this averaging neglects the possibility that different types of framing effects affect different groups of people, as McDonald et al. (2019) point out:

Imagine that we find five framing effects that each affects 20 people out of 100. If the same 20 people are vulnerable to all five framing effects, then the remaining 80 are not vulnerable to any of those effects. However, if each framing effect occurs in a separate 20 with no overlap, then every one of the 100 is vulnerable to one framing effect. (p. 36)

In this way, relatively low levels of unreliability on any given type of framing may still lead to high percentages of people being unreliable, depending on the amount of overlap between groups who are affected by different types of framing effect.

How can this overlap be determined? Between-subject studies do not help here, for they do not reveal which individual participants are affected by framing. A within-subject design, on the other hand, can determine the effects of framing on individual participants by comparing their responses across the different framing conditions. If a within-subject study included multiple types of framing manipulations, then the percentage of people vulnerable to at least one of them could be estimated.

We have argued that between-subject designs are limited in the extent to which they can help us answer three questions that we need to answer in order to assess the strength of the arguments from moral framing effects to the unreliability of moral judgments:

Question 1: Are the effects of framing on moral judgments repeated or only initial?

Question 2: What percentage of people are affected in their moral judgments by any given type of frame?

Question 3: What percentage of people are affected in their moral judgments by at least one type of frame?

We explained how a within-subject approach might be able to help answer these questions. We now report and discuss the results of a new within-subject study of a number of different types of moral framing effects that we conducted.

3. Method

3.1 Design

Our study included the four types of framing effects we discussed in the first section: Save vs. Kill, Actor vs. Observer, $A \rightarrow O$ vs. $O \rightarrow A$, and $M \rightarrow SE$ vs. $SE \rightarrow M$. The experiment consisted of two sessions, each containing multiple scenarios divided by filler tasks. [Figure 1](#) shows a schematic of a single session. In each session, participants were asked to read a total of six short moral scenarios that were divided into four units, one for each type of framing effect under investigation. The first two units contained a single scenario each, with the scenario in the first unit being presented in one of the two possible Save vs. Kill frames, while the scenario in the second unit was presented in one of the two possible Actor vs. Observer frames. The third and fourth unit each contained a scenario pair, with the first pair being presented in one of the two possible $A \rightarrow O$ vs. $O \rightarrow A$ frames and the second pair being presented in one of the two possible $M \rightarrow SE$ vs. $SE \rightarrow M$ frames. This order of the four units was left unchanged throughout the study. To counteract unintended order effects within a single session, all adjacent units were separated by two filler tasks. Each session thus consisted of six moral scenarios plus six filler tasks.

A major challenge to investigating framing effects using a within-subject design is that participants receive both framing conditions, so they might recognize the similarity between the two conditions and strive to make their choices consistent, thereby decreasing the sensitivity of the design (Aczel et al., 2018). The two most common strategies researchers have employed to combat this challenge are the insertion of unrelated filler materials between the two framing conditions (e.g., Bruine De Bruin et al., 2007) and using a time delay (e.g., Levin et al., 2002). Here, we chose the latter approach: Having completed the first session, participants returned for the second session after a delay of between 165 h and 208 h. In addition, we changed the names of the agents in

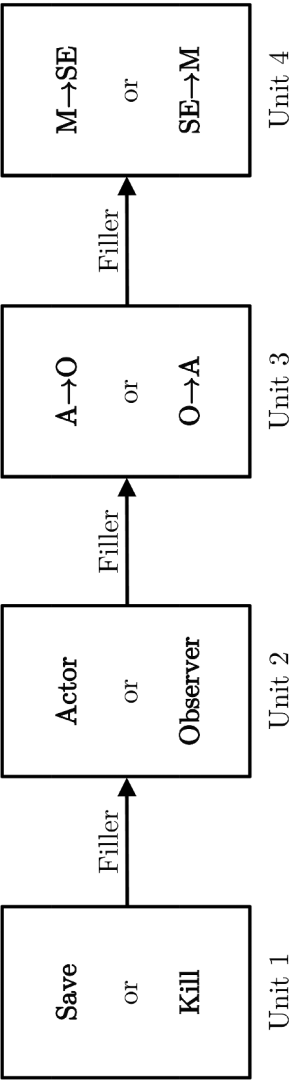


Figure 1. Schematic of one session of the experiment.

our scenarios between sessions, being careful to match these names in both gender and length (compare LeBoeuf & Shafir, 2003, Exp. 2).

3.2 Materials

Prior research on Save vs. Kill, Actor vs. Observer, and $M \rightarrow SE$ vs. $SE \rightarrow M$ almost exclusively employed variants of the well-known trolley dilemma (Foot, 1967; Thomson, 1976). We used similar scenarios in order to investigate whether those reported effects would recur in a within-subject design. Specifically, we picked three kinds of scenario, each of which has previously been used to research at least one type of moral framing effect. The first was the standard *Trolley* scenario (e.g., Lombrozo, 2009; Nadelhoffer & Feltz, 2008; Petrinovich & O'Neill, 1996): an out-of-control railroad car is threatening to overrun and kill five people; the only way to prevent these deaths is by pulling a lever, which will divert the railroad car onto a side track. Unfortunately, on this side track, there is a sixth person whom the railroad car would kill instead.

The two other kinds of scenario were chosen to be equivalent to *Trolley* in structure, effects, and intention, but different in context and content. In *Shark* (Petrinovich & O'Neill, 1996, Exp. 2; Wiegmann & Waldmann, 2014, Exp. 4), the railroad car is replaced by a shark swimming toward a group of five swimmers. The only way to save the five swimmers is by making loud noises that will redirect the shark toward the noise maker, but the shark will then kill a sixth swimmer in the water between them. Likewise, in *Gas* (Wiegmann & Waldmann, 2014, Exp. 4), poisonous gas is being emitted into a hospital room with five patients. The only way to save them is to push a switch that will redirect the gas into a different room, where it will kill a sixth patient. We created versions of *Trolley*, *Shark*, and *Gas* for Save vs. Kill, Actor vs. Observer, and $M \rightarrow SE$ vs. $SE \rightarrow M$.

The situation for $A \rightarrow O$ vs. $O \rightarrow A$ is somewhat different. We know of no previous research employing trolley-type scenarios to explore this type of moral framing effect. Instead, Haidt and Baron (1996) used one scenario about lying and another about workplace neglect; Schwitzgebel and Cushman (2012) employ two scenarios about a trade-off between the safety of an agent and that of a third party. Because the study by Schwitzgebel and Cushman was more highly powered, we used their materials. Out of their two scenario pairs, we adapted the pair that yielded the larger effect size. In *Oxygen*, a person goes into a diving pod at an aquarium, when an earthquake hits. In the action version of the scenario, the earthquake pinches off the person's own oxygen supply line. In the omission version, the earthquake obstructs the oxygen supply line of another person. The agent must now decide whether to switch the supply lines (in the action version) or whether to do nothing (in the omission version).

To minimize the risk of confounding idiosyncrasies of individual scenarios, all scenarios were matched as closely as possible in both length and structure. All study materials, including the scenarios, are available online at the following link: <https://osf.io/b6nk2/>.

As filler material, we inserted a number of tasks commonly used in the heuristics and biases literature (Aczel et al., 2015), which have previously been used for masking purposes in within-subject research on non-moral framing effects (Aczel et al., 2018). From pilot research, the average completion time for each of the filler tasks was estimated to be between one and two minutes. All adjacent units of scenarios were separated by two such tasks (see Figure 1).

3.3 Variables

For every scenario, participants were asked to rate the hypothetical action on a seven-point scale labeled at both endpoints, from “Completely morally permissible” (= 1) to “Not at all morally permissible” (= 7). We adopted this “permissibility”-scale because it has been used by the largest number of publications on moral framing effects to date.

Most existing empirical evidence for moral framing effects is of one of two forms. Some studies (e.g., Nadelhoffer & Feltz, 2008) focus on *judgment reversals*. A judgment reversal occurs when a participant makes opposite judgments in the different frames (e.g., permissible in one frame, not permissible in the other frame). Other studies (e.g., Wiegmann & Waldmann, 2014) instead focus on *judgment shifts*. This notion includes but is broader than judgment reversals. A judgment shift is any change in judgment between two frames. For example, participants in both frames might judge that an action should be carried out, but the strength of their judgment changes between frames.

On the definition used in this paper, both judgment shifts and judgment reversals provide evidence of moral framing effects: they both indicate that morally irrelevant differences in the way a scenario is presented affect people’s moral judgments. However, it has been suggested that they may have different implications for the question to what extent moral framing effects show unreliability (Andow, 2016). While we are not aware of anyone who has spelled this out in detail, the intuition seems to be that if moral framing effects lead us to make opposite judgments in different frames, this would clearly be a problem. In contrast, if framing effects generally affect only the strength of our moral judgments without affecting their polarity, people would still be making the same moral judgment (in some sense) in both frames. Therefore, judgment shifts would be much less worrying for the reliability of moral judgment.

What this would mean is that we should focus on studies of judgment reversals for the purpose of inferring unreliability from framing effects.

Demaree-Cotton (2016) does this in her meta-analysis, explicitly excluding a couple of studies which only report results about judgment shifts (p. 9).

On the other side, Andow (2016) has argued that framing effects which only affect the strength, not the polarity of moral judgments still cast serious doubts on their reliability.⁴ If this is right, then both evidence of judgment reversals and evidence of judgment shifts would be relevant for the argument.

We are inclined to agree with Andow. Nevertheless, we here do not wish to adjudicate between the two views. Therefore, we strove to include operationalizations for both judgment shifts and judgment reversals in our experiment. We say that a participant exhibits a judgment shift for a given type of framing manipulation if they give different ratings on our seven-point scale in the two framing conditions. To operationalize judgment reversals, we dichotomize our raw ratings (cf. Matthew et al., 2012; Wiegmann et al., 2012). We label all responses above the scale midpoint (= 4) “Impermissible”, and all other responses “Not Impermissible”. A participant is then showing a rating reversal if she responded Impermissible in one frame and Not Impermissible in the other frame.

3.4 Participants

All participants were recruited through the online subject pool Prolific (<https://www.prolific.co>). For the reliability of data gathered through Prolific for experimental studies, see Peer et al. (2017). We restricted participation to current residents of the US, the UK, and Ireland whose first language was English and who had at least a 95% acceptance rating on Prolific.

Pilot research indicated that some participants spend implausibly short amounts of time on reading and responding to the scenarios. Since such insufficient effort responding is likely to artificially inflate the amount of inconsistent responses, we used response time as a criterion for excluding some participants (as recommended by Huang et al., 2012; Mason & Suri, 2012; Wood et al., 2017). Following Mason and Suri (2012), we empirically determined a response time threshold for exclusion by conducting a small subsidiary study. 25 participants (13 female, 1 not specified; M (SD) = 35.56(11.44)y; 12.0% students) were recruited through Prolific. For their participation, participants were compensated 0.75. USD The procedure through the survey instructions was identical to the main experiment (see Procedure). Participants were then asked to read and rate a version of *Shark*, *Gas*, *Oxygen* and *Trolley*, in that order. We next asked a series of four multiple-choice questions, which required that participants had properly read and understood the scenarios. Six

participants got at least one of these questions wrong and were excluded. From the remaining sample of 19 (10 female; $M(SD) = 34.00(10.91)$ y; 15.8% students), we calculated an overall mean scenario response time by averaging both over scenario and participants. We then subtracted two standard deviations to account for the top 95% of readers. In this way, we arrived at an estimate of 18.10s for the minimum average time required for reading and understanding one session's worth of moral scenarios.

A simulation-based power analysis (Green et al., 2016) based on our pilot data indicated that to detect moderate effect sizes of at least 0.25, a sample size of $N = 250$ would be needed to achieve acceptable power of at least 0.80 (Cohen, 1992). To account for both dropout and exclusion of participants due to timing, we recruited a total of 366 participants for the first session. Seven participants did not complete all materials and were excluded, leaving us with 359 participants (212 female, 1 not specified; $M(SD) = 35.02(12.93)$ y; 21.4% students). Of those participants, 320 (185 female; $M(SD) = 35.84(12.66)$ y; 18.1% students), or 81%, returned to complete the second session. 59 participants were excluded by our response time threshold, leaving us with a final sample of $N = 261$ (156 female; $M(SD) = 36.60(13.07)$ y; 14.2% students).

3.5 Procedure

For both sessions, participants were contacted by Prolific with an invitation to participate in the study. Upon accepting, participants were redirected to the study, which was hosted on LimeSurvey (<https://www.limesurvey.org>). After giving informed consent and receiving instructions, participants were asked to complete the study materials. Participants were compensated 1.75 USD per session. In addition, participants who completed both sessions received a bonus payment of 1.50 USD. The study was approved by the Duke University Campus Institutional Review Board.

In the first session, for each type of framing manipulation, participants were randomly assigned to one of the two possible framing conditions; in the second session, they were then assigned to the other condition. For example, for Save vs. Kill, participants were either assigned to the framing condition Save in the first session and to the framing condition Kill in the second session, or to Kill in the first session and Save in the second session. Assignments for the different types of framing manipulation were made independently of each other. Each of Save vs. Kill, Actor vs. Observer, and $M \rightarrow SE$ vs. $SE \rightarrow M$ was presented in one of the three trolley-type scenario contexts (*Trolley*, *Shark*, *Gas*). This assignment was made pseudo-randomly, such that for each participant, each scenario context was used exactly once, and was kept constant across both sessions of the experiment.

4. Results and discussion

Analysis was carried out in R (R Core Team, 2019). Figures were created using the package ggplot2 (Wickham, 2016). All the data collected for this experiment, as well as the analysis code, are available online at the following link: <https://osf.io/b6nk2/>.

4.1 Question 1: Are the effects of framing on moral judgments repeated or only initial?

4.1.1 Results

Descriptive statistics are presented in Table 1. As described above, we report results for both judgment shifts and judgment reversals.

For analysis, we employed general linear mixed-effect models (Bates et al., 2015). *t*-tests were computed using Satterthwaite's approximation for degrees of freedom (Kuznetsova et al., 2017). We added a random intercept for participant. For Save vs. Kill, Actor vs. Observer, and M→SE vs. SE→M, we in addition added a random intercept for scenario context. judgment shifts and judgment reversals were analyzed separately. For the first, we chose an identity link function for the model. For the second, as the outcome variable was discrete, we chose a binomial link function. For both analyses, we entered frame and session as predictors into the model. Nowhere did the interaction of frame and session improve model fit significantly, so we did not include their interaction term into the final model. The addition of sex, age, student status, and their interaction with frame did not significantly improve model fit, with one exception: for M→SE vs. SE→M, adding sex as a predictor to the judgment shifts model was found to significantly improve model fit, $\chi^2(1) = 4.19$, $p = 0.04$; male participants tended to give higher impermissibility ratings, $b = -0.42$, $SE = 0.20$,

Table 1. For judgment shifts, we report mean impermissibility ratings and their standard deviations. for judgment reversals, we report the percentages of impermissible judgments.

	judgment shifts	judgment reversals
Frame	<i>M</i> (<i>SD</i>)	% Impermissible
Save vs. Kill		
Save	3.00(1.60)	18.4%
Kill	3.23(1.78)	24.1%
Actor vs. Observer		
Actor	3.21(1.72)	21.5%
Observer	3.28(1.75)	21.8%
A→O vs. O→A		
A→O	3.33(1.99)	27.2%
O→A	3.12(1.83)	22.6%
M→SE vs. SE→M		
M→SE	3.69(1.88)	32.2%
SE→M	3.26(1.75)	22.6%

$t = -2.04$, $p = 0.04$. However, adding a term for the interaction between frame and sex did not further improve model fit, $\chi^2(1) = 0.10$, $p = 0.76$.

For Save vs. Kill, participants' impermissibility ratings in the Kill condition were higher overall than in the Save condition. This pattern was statistically significant, $b = 0.19$, $SE = 0.09$, $t = 2.07$, $p = 0.04$. In addition, we found that session significantly predicted impermissibility ratings, in that overall, participants gave higher ratings in the second session when compared to the first session, $b = 0.32$, $SE = 0.09$, $t = 3.41$, $p < 0.001$. Looking at judgment reversals, we found that a significantly larger proportion of participants judge the hypothetical action as Impermissible in the Kill condition than in the Save condition, $OR = 2.76$, $z = 2.23$, $p = 0.03$. The effect of session found for judgment shifts was also again observed. More participants gave judgments of Impermissible in the second session than in the first session, $OR = 5.14$, $z = 3.26$, $p = 0.001$.

Turning to Actor vs. Observer, we found no significant effect of frame, neither in the sense of judgment shifts, $b = 0.06$, $SE = 0.08$, $t = 0.68$, $p = 0.50$, nor of judgment reversals, $OR = 1.75$, $z = 1.00$, $p = 0.32$. In contrast, session again had a significant effect. Overall, participants tended to give higher ratings of impermissibility in the second session than in the first session, $b = 0.43$, $SE = 0.09$, $t = 4.90$, $p < 0.001$. Similarly, a larger proportion of participants judged the hypothetical action as Impermissible in the second session, $OR = 233.32$, $z = 5.79$, $p < 0.001$.

For A→O vs. O→A, frame did not significantly predict impermissibility ratings, $b = -0.18$, $SE = 0.14$, $t = -1.27$, $p = 0.20$, or Impermissible judgments, $OR = 0.79$, $z = -1.00$, $p = 0.32$. No significant effect of session was found, $b = -0.19$, $SE = 0.14$, $t = -1.34$, $p = 0.18$, and $OR = 0.65$, $z = -1.82$, $p = 0.07$.

Finally, for M→SE vs. SE→M, we found that participants overall rated the hypothetical action as more impermissible when the side-effect scenario was presented after the means scenario than when it preceded it. This pattern was statistically significant, $b = -0.44$, $SE = 0.10$, $t = -4.41$, $p < 0.001$. The effect of session did not reach statistical significance, $b = 0.09$, $SE = 0.10$, $t = 0.86$, $p = 0.39$.

Turning to judgment reversals, significantly more participants in the M→SE condition judged the hypothetical action in the side-effects scenario to be Impermissible than in the SE→M condition, $OR = 0.43$, $z = -3.14$, $p = 0.002$. In contrast, session did not significantly predict judgments of Impermissible, $OR = 1.59$, $z = 1.79$, $p = 0.07$.

4.1.2 Discussion

For two out of the four types of framing manipulation included in our experiment, we find evidence of effects of frame, both in the sense of

participants' exhibiting judgment shifts and in the sense of participants' exhibiting judgment reversals.

For Save vs. Kill, in line with previous work, we find that the same participants will give significantly higher impermissibility ratings in the Kill condition than in the Save condition. Likewise, a significantly larger proportion of participants judged a hypothetical action to be Impermissible when the scenario was presented in the Kill condition. For $M \rightarrow SE$ vs. $SE \rightarrow M$, for a sequence of two trolley-style scenarios, the same participants rated the action in the side-effect scenario as significantly more impermissible if that scenario was presented after, rather than before, a means scenario. Similarly, significantly more participants gave judgments of Impermissible in the $M \rightarrow SE$ condition than in the $SE \rightarrow M$ condition. Again, these results agree with what had previously been reported in between-subject studies.

There was no interaction of frame with session, meaning that which framing condition participants received first made no significant difference for the effect of frame on participants' responses. Thus, our results suggest that for Save vs. Kill and for $M \rightarrow SE$ vs. $SE \rightarrow M$, the effects of frame are at least in part repeated rather than only initial.

In contrast, we find no evidence for repeated effects of frame for either Actor vs. Observer or $A \rightarrow O$ vs. $O \rightarrow A$. Neither participants' raw ratings, nor their dichotomized ratings, were found to systematically vary with the frame.

Our analysis also revealed that session significantly predicted impermissibility ratings and Impermissible judgments for two out of the four framing manipulations. Participants overall tended to give higher ratings of impermissibility and higher proportions of Impermissible judgments in the second session than in the first. We know of only one prior report of this curious phenomenon (Borg et al., 2010) and we have no convincing explanation for it. We report it here in case some people might like to know about it.

4.2 Question 2: What percentage of people are affected in their moral judgments by any given type of frame?

4.2.1 Results

To get a sense of the ways in which individual participants' responses changed between frames, we calculated a difference score for each type of framing manipulation and participant. This difference score is defined as the impermissibility rating given in one frame (Kill; Observer; $A \rightarrow O$; $M \rightarrow SE$) minus the impermissibility rating given in the other. Positive difference scores indicate a difference in responses between frames that is in line with the overall trends reported in Table 1.⁵ Figure 2 shows the observed difference score distributions.

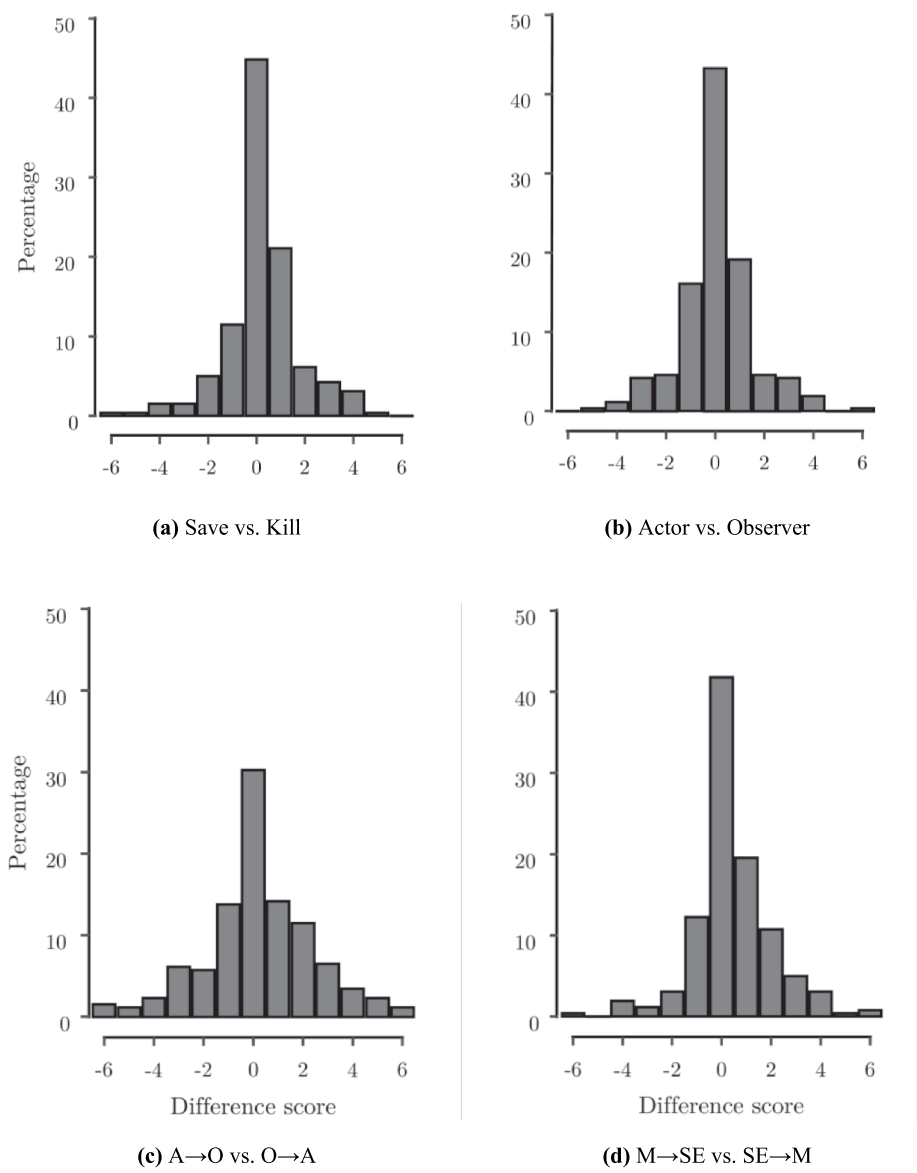


Figure 2. Barplots of observed difference score distributions.

Table 2 contains more detailed descriptive statistics. We again report results on both judgment shifts and judgment reversals. For all types of framing manipulation, a substantial number of participants changed their ratings between frames in a direction opposite from the overall trend. Therefore, we separate results into those in line with the overall trends (marked in bold) and those opposite to the overall trends (not marked in bold) for each type of framing manipulation.

Table 2. Summary statistics of differences in participants' responses between frames. for judgment shifts, we report the percentage of participants exhibiting a judgment shift and corresponding mean impermissibility ratings. for judgment reversals, we report the percentage of participants exhibiting a judgment reversal and the corresponding percentage of impermissible judgments. square brackets show 95% confidence intervals. for each type of framing manipulation, the results in line with the overall trend (Table 1) are marked in bold.

	judgment shift		judgment reversal	
	% of participants	<i>M</i> (<i>SD</i>)	% of participants	% Impermissible
Save vs. Kill				
Save	34.9%	2.70(1.35)	13.0%	4.6%
Kill		4.43(1.59)		17.6%
Save	20.3%	4.43(1.59)	7.3%	8.0%
Kill		2.23(1.14)		0.8%
Total	55.2% [49.1%, 61.1%]		20.3% [15.9%, 25.6%]	
Actor vs. Observer				
Actor	30.3%	2.66(1.31)	10.4%	3.4%
Observer		4.34(1.65)		13.8%
Actor	26.4%	4.36(1.57)	10.0%	11.9%
Observer		2.68(1.28)		1.9%
Total	56.7% [50.6%, 62.6%]		20.3% [15.9%, 25.6%]	
A→O vs. O→A				
A→O	30.7%	2.24(1.46)	11.5%	3.1%
O→A		4.45(1.57)		14.6%
A→O	39.1%	4.65(1.66)	16.1%	19.2%
O→A		2.37(1.36)		3.1%
Total	69.7% [63.9%, 75.0%]		27.6% [22.5%, 33.3%]	
M→SE vs. SE→M				
M→SE	18.8%	2.59(1.31)	7.7%	1.5%
SE→M		4.29(1.72)		9.2%
M→SE	39.5%	4.29(1.72)	17.3%	21.5%
SE→M		2.83(1.37)		4.2%
Total	58.2% [52.2%, 64.1%]		24.9% [20.0%, 30.5%]	

4.2.2 Discussion

In the sense of exhibiting judgment shifts, we find that on each of the four types of framing manipulation investigated here, more than half of participants are affected by framing (on Table 2, see the totals under the column labeled “% of participants”). This number goes as high as 69.7% of participants on A→O vs. O→A. The majority of differences in ratings between frames occurred in line with the overall trend; however, a sizable proportion of differences in the opposite direction were also observed.

If we instead look at judgment reversals, between 20.3% and 27.6% of participants show evidence of framing effects (on Table 2, see the totals under the column labeled “% of participants”). While the majority of these effects again fell in line with overall trends, between 7.3% and 11.5% of participants exhibited reversals in the opposite direction. This more complex picture of the effects of moral framing on individuals could not have been found by previous research, because between-subject designs compare only group-level tendencies.

What do these results mean for Demaree-Cotton's (2016) estimate of the proportion of moral judgments that are vulnerable to framing effects? In her meta-analysis, she reports an estimate of 20% for the rate of framing in the sense of judgment reversal. All of the estimates we report here with our operationalization of judgment reversal are larger than that (between 20.3% and 27.6%, though two of the confidence intervals include 20%). If we instead look at the rate of framing effects in the sense of judgment shifts, we arrive at estimates which are very much larger than hers (over 50%). We conclude that at a minimum, Demaree-Cotton's estimate of 20% is likely somewhat optimistic.

4.3 Question 3: What percentage of people are affected in their moral judgments by at least one type of frame?

4.3.1 Results

We again report results for both judgment shifts and judgment reversals. For both, we calculated the percentages of participants who showed evidence of zero, one, two, three and four types of framing effects. We also determined the proportion of participants who showed evidence for at least one type of framing effect. Table 3 contains the results.

In addition, we calculated an estimate of the overlap between groups of participants affected by different types of framing effects. For two types of framing effects, call the sets of participants who show evidence of being affected A and B. We then define their overlap as the quotient of the sizes of their intersection $A \cap B$ and their union $A \sqcup B$: $\text{overlap} = |A \cap B|/|A \sqcup B|$. In other words, overlap is defined as the proportion of participants who show evidence of both types of framing effect relative to all of the participants who show evidence of at least one of the two types.

Since our study included four types of framing manipulation, there are six different pairs of types of framing effect. Looking at judgment shifts, we find a mean overlap of $M = 49.8\%$, $SD = 7.0\%$. Looking at judgment reversals, the mean amount of overlap is $M = 22.1\%$, $SD = 11.8\%$.

Table 3. Percentages of participants showing evidence of some number n of different framing effects. results for both judgment shifts and judgment reversals are reported. square brackets show 95% confidence intervals.

	judgment shifts	judgment reversals
0	6.1% [3.8%, 9.7%]	43.3% [37.4%, 49.4%]
1	19.5% [15.2%, 24.8%]	30.3% [25.0%, 36.1%]
2	22.6% [17.9%, 28.1%]	17.2% [13.1%, 22.3%]
3	31.8% [26.4%, 37.7%]	8.4% [5.6%, 12.4%]
4	19.9% [15.5%, 25.2%]	0.8% [0.2%, 2.8%]
$n \geq 1$	93.9% [90.3%, 96.2%]	56.7% [50.6%, 62.6%]

4.3.2 Discussion

Our results for Question 3 suggest that groups of people who are vulnerable to different types of framing effect only partially overlap. Regardless of whether we consider judgment reversals or judgment shifts as evidence for an effect of frame on people's moral judgments, we find that on average, less than 50% of participants who show evidence of being susceptible to one type of framing effect also show evidence of being susceptible to another type. Such an estimate might prove useful if in the future, additional types of moral framing effects are discovered. In that case, it could be used to calculate a rough update of our answer to Question 3.

We find that more than half of participants show evidence of at least one type of moral framing effect, both in the sense of exhibiting judgment shifts and in the sense of exhibiting judgment reversals. If we consider all participants who exhibited judgment shifts as showing evidence of being framed, then 93.9% of participants show evidence of at least one type of framing effect. If we instead require that participants exhibited judgment reversals, still 56.7% of participants show evidence of at least one type of framing effect. Both of these values are much larger than Demaree-Cotton's estimate of 20%. Thus, we again conclude that her estimate is too optimistic.

5. General discussion

We argued that between-subject designs cannot help us answer a number of questions that are crucial if we want to evaluate the strength of the argument from framing effects against the reliability of our moral judgments. We then reported and discussed the results of a new within-subject study of four different types of moral framing effects.

What do our results mean for the argument? We think that the argument looks stronger in light of our new results than it did before. In particular, we find some evidence of repeated effects of frame on people's moral judgments in two out of the four types of framing manipulation included in our experiment, Save vs. Kill and M→SE vs. SE→M. This evidence of repeated framing strengthens the argument by suggesting more unreliability in our moral judgments than if the framing effects were only in the initial presentation. Furthermore, our results on Questions 2 and 3 both indicate that the proportion of moral judgments which are vulnerable to moral framing effects is likely larger than previously estimated. For Question 2, we find evidence for effects of frame at the level of individual participants that may have been missed by previous research using between-subject designs. For Question 3, we find that the proportion of people who show effects of frame for at least one

type of framing manipulation is likely larger than previous estimates. All of these results hold whether we consider judgment shifts or judgment reversals as evidence for moral framing effects. All of this strengthens the argument from framing effects against the reliability of our moral judgments.

5.1 Limitations

Our research has a number of limitations. First, all of our participants were from a small number of English-speaking Western countries. It is thus unclear whether our results are representative of human beings in general (Henrich et al., 2010).

Second, while we found a statistically significant effect of frame for Save vs. Kill, our study did not have adequate power to reliably detect effect sizes as small as the one we observed. Future research should therefore strive to replicate our results in larger samples.

Third, in contrast to most previous work on moral framing effects, we included multiple types of framing manipulations in our study. We took some steps to prevent the materials used to probe different types of effect from influencing each other, such as the inclusion of filler tasks between moral scenarios. Still, such influences might have increased the amount of noise in our dataset. Future research should try to replicate our results using fewer materials.

Lastly, our discussion of Questions 2 and 3 observed that a majority of participants changed their rating between frames for all four types of framing manipulation. We also reported that for each type, a substantial proportion of participants gave judgments of Impermissible in one frame while giving judgments of Not Impermissible in the other. Both changes seemed to occur rather unsystematically, with a sizable minority of participants going against the observed overall trends.

This poses a problem for the interpretation of our data. On one hand, these patterns may have been caused by our framing manipulations – different groups of participants being affected by framing in opposite ways has not been unheard of (Tobia et al., 2013). On the other hand, they may simply indicate that our instruments had low test-retest reliability. Our study itself does not provide any way of distinguishing between the two explanations. Since we do not know of any study looking into the test-retest reliability of the kinds of moral scenarios used here, future research will be needed to settle this question.⁶ We encourage such research.

5.2 More research should be done

Our results point toward several potential avenues for future research. In addition to those discussed in the previous section, we want to mention two more.

First, we found that *some* people are susceptible to moral framing effects. But *which* people are more susceptible to moral framing effects? Future research might investigate variations in people's susceptibility to moral framing effects by deploying a number of individual differences measures, such as personality, cognitive style, cognitive ability, and emotion differentiation.⁷ This research could further our understanding of moral framing effects, and might even suggest some practical interventions to counter some effects of frames on people's moral judgments.

Second, moral framing effects have been explored for only a very thin slice of morality. Existing research, ours included, has overwhelmingly relied on trolley-style sacrificial dilemmas. This limitation makes it very hard to support any strong claim about what proportion of our moral judgments in general might be affected by framing. Future research should expand the empirical investigation of moral framing effects both to harm-based scenarios other than sacrificial dilemmas, but also to other domains of morality such as fairness, loyalty, or sanctity (Haidt and Joseph, 2004), as well as honesty. Only then will we know how far moral framing effects extend, and the degree to which they introduce unreliability into our moral judgments.

Notes

1. See also Andow (2018) and Sinnott-Armstrong (2008).
2. Studies have also found effects of language (Costa et al., 2014; Geipel et al., 2015) and font (Spears et al., 2018). If the circumstances in which a scenario is presented are seen as part of "the way a scenario is presented", then our and Demaree-Cotton's definition counts many other phenomena as framing effects, including the influence on moral judgment of disgust (Schnall et al., 2008; Wheatley & Haidt, 2005), eye gaze (Pärnamets et al., 2015), cues to cleanliness (Helzer & Pizarro, 2011) or social power (Lammers & Stapel, 2009), sleep deprivation (Cho et al., 2017; Olsen et al., 2010), blood alcohol level (Duke & Laurent, 2015), social information (Kelly et al., 2017) and even room temperature (Nakamura et al., 2014). While these circumstantial effects are not usually called framing effects, they could form the basis for very similar arguments against the reliability of our moral judgments. However, while fascinating in their own right, this article will not focus on such effects of the circumstances in which a scenario is presented.
3. A slightly different version of this effect employs sets of more than two means and side-effect scenarios, the order of which is then manipulated (Matthew et al., 2012; Petrinovich & O'Neill, 1996, Exp. 2; Wiegmann et al., 2012). The findings reported by these authors have been entirely consistent with the literature on $M \rightarrow SE$ vs. $SE \rightarrow M$.
4. Andow himself does not refer to this problem as unreliability.

5. That is, for Save vs. Kill, higher ratings of impermissibility in Kill than in Save; for Actor vs. Observer, higher ratings in Observer than in Actor; for $A \rightarrow O$ vs. $O \rightarrow A$, higher ratings in $A \rightarrow O$ than in $O \rightarrow A$; for $M \rightarrow SE$ vs. $SE \rightarrow M$, higher ratings in $M \rightarrow SE$ than in $SE \rightarrow M$.
6. Hannikainen et al. (2018) do investigate test-retest reliability for a series of sacrificial moral dilemmas, but over a much longer period than the time delay used in this study ($M = 8.4$ years).
7. Similar projects have been undertaken for non-moral framing effects: e.g., Bruine De Bruin et al. (2007); Levin et al. (2002); Mahoney et al. (2011); Stanovich and West (1998).

Acknowledgments

The authors wish to thank Thomas Nadelhoffer and Valerij Zisman for valuable comments on an earlier version of this paper. The research reported in this paper was presented at two meetings of MAD Lab at Duke University and at a meeting of the Moral Psychology Research Group at the University of Utah. We thank all participants for their helpful feedback and suggestions. The first author gratefully acknowledges the German-American Fulbright Commission for its financial support.

Disclosure statement

The authors declare no conflict of interest.

Notes on contributor

Paul Rehren is a PhD student in Philosophy at Utrecht University. His research is at the intersection of moral philosophy and experimental psychology.

Walter Sinnott-Armstrong is Chauncey Stillman Professor of Practical Ethics in the Philosophy Department and the Kenan Institute for Ethics at Duke University. He has secondary appointments in the Psychology and Neuroscience Department and the Law School. He publishes widely in ethics, moral psychology and neuroscience, philosophy of law, epistemology, and informal logic.

ORCID

Paul Rehren  <http://orcid.org/0000-0003-4267-3097>

References

- Aczel, B., Bago, B., Szollosi, A., Foldes, A., & Lukacs, B. (2015). Measuring individual differences in decision biases: Methodological considerations. *Frontiers in Psychology*, 6 (November), 1770, <https://doi.org/10.3389/fpsyg.2015.01770>
- Aczel, B., Szollosi, A., & Bago, B. (2018). The effect of transparency on framing effects in within-subject designs. *Journal of Behavioral Decision Making*, 31(1), 25–39. <https://doi.org/10.1002/bdm.2036>

- Andow, J. (2016). Reliable but not home free? What framing effects mean for moral intuitions. *Philosophical Psychology*, 29(6), 904–911. <https://doi.org/10.1080/09515089.2016.1168794>
- Andow, J. (2018). Are intuitions about moral relevance susceptible to framing effects? *Review of Philosophy and Psychology*, 9(1), 115–141. <https://doi.org/10.1007/s13164-017-0352-5>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1. <https://doi.org/10.18637/jss.v067.i01>
- Bengson, J. (2013). Experimental attacks on intuitions and answers. *Philosophy and Phenomenological Research*, 86(3), 495–532. <https://doi.org/10.1111/j.1933-1592.2012.00578.x>
- Borg, J. S., Sinnott-Armstrong, W., & Grafton, S. (2010). Moral decision-making...the second time around [Poster]. Society for Cognitive Neuroscience meeting, Montreal, CA
- Bruine De Bruin, W., Andrew, M., Parker, J., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, 92(5), 938–956. <https://doi.org/10.1037/0022-3514.92.5.938>
- Campbell, R., & Kumar, V. (2012). Moral reasoning on the ground. *Ethics*, 122(2), 273–312. <https://doi.org/10.1086/663980>
- Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1), 1–8. <https://doi.org/10.1016/j.jebo.2011.08.009>
- Cho, K., Barnes, C. M., & Guanara, C. L. (2017). Sleepy punishers are harsh punishers: Daylight saving time and legal sentences. *Psychological Science*, 28(2), 242–247. <https://doi.org/10.1177/0956797616678437>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Core Team, R. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., Keysar, B., & Sigman, M. Your morals depend on language. (2014). *PLoS ONE*, 9(4), e94842. Edited by Mariano Sigman. <https://doi.org/10.1371/journal.pone.0094842>
- Demaree-Cotton, J. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology*, 29(1), 1–22. <https://doi.org/10.1080/09515089.2014.989967>
- Details omitted for double-blind reviewing.
- Duke, A. A., & Laurent, B. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. *Cognition*, 134(January), 121–127. <https://doi.org/10.1016/j.cognition.2014.09.006>
- Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, 5, 5–15.
- Geipel, J., Hadjichristidis, C., Surian, L., & Brañas-Garza, P. (2015). The foreign language effect on moral judgment: The role of emotions and norms. *Plos One*, 10(7), e0131529. <https://doi.org/10.1371/journal.pone.0131529>
- Green, P., MacLeod, C. J., & Nakagawa, S. (2016). SIMR an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493–498. <https://doi.org/10.1111/2041-210X.12504>
- Haidt, J., & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology*, 26(2), 201–218. [https://doi.org/10.1002/\(SICI\)1099-0992\(199603\)26:2<201::AID-EJSP745>3.0.CO;2-J](https://doi.org/10.1002/(SICI)1099-0992(199603)26:2<201::AID-EJSP745>3.0.CO;2-J)

- Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4), 55–66. <https://doi.org/10.1162/0011526042365555>
- Hannikainen, I. R., Machery, E., & Cushman, F. A. (2018). Is utilitarian sacrifice becoming more morally permissible? *Cognition*, 170(January), 95–101. <https://doi.org/10.1016/j.cognition.2017.09.013>
- Helzer, E. G., & Pizarro, D. A. (2011). Dirty liberals!: Reminders of physical cleanliness influence moral and political attitudes. *Psychological Science*, 22(4), 517–522. <https://doi.org/10.1177/0956797611402514>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hershey, J. C., & Paul, J. H. S. (1980). Prospect theory's reflection hypothesis: A critical examination. *Organizational Behavior and Human Performance*, 25(3), 395–418. [https://doi.org/10.1016/0030-5073\(80\)90037-9](https://doi.org/10.1016/0030-5073(80)90037-9)
- Horne, Z., & Livengood, J. (2017). Ordering effects, updating effects, and the specter of global skepticism. *Synthese*, 194(4), 1189–1218. <https://doi.org/10.1007/s11229-015-0985-9>
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114. <https://doi.org/10.1007/s10869-011-9231-8>
- Kelly, M., Ngo, L., Chituc, V., Huettel, S., & Sinnott-Armstrong, W. (2017). Moral conformity in online interactions: Rational justifications increase influence of peer opinions on moral judgments. *Social Influence*, 12(2–3), 57–68. <https://doi.org/10.1080/15534510.2017.1323007>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lammers, J., & Stapel, D. A. (2009). How power influences moral thinking. *Journal of Personality and Social Psychology*, 97(2), 279–289. <https://doi.org/10.1037/a0015437>
- Lanteri, A., Chelini, C., & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics*, 83(4), 789–804. <https://doi.org/10.1007/s10551-008-9665-8>
- LeBoeuf, R. A., & Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. *Journal of Behavioral Decision Making*, 16(2), 77–92. <https://doi.org/10.1002/bdm.433>
- Levin, I. P., Gaeth, G. J., Schreiber, J., & Lauriola, M. (2002). A new look at framing effects: Distribution of effect sizes, individual differences, and independence of types of effects. *Organizational Behavior and Human Decision Processes*, 88(1), 411–429. <https://doi.org/10.1006/obhd.2001.2983>
- Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, 33(2), 273–286. <https://doi.org/10.1111/j.1551-6709.2009.01013.x>
- Mahoney, K. T., Walter Buboltz, I. P., Dennis Doverspike, L., Daniel, J., & Svyantek. (2011). Individual differences in a within-subjects risky-choice framing study. *Personality and Individual Differences*, 51(3), 248–257. <https://doi.org/10.1016/j.paid.2010.03.035>
- Mason, W., & Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, 44(1), 1–23. <https://doi.org/10.3758/s13428-011-0124-6>
- Matthew, L. S., Wiegmann, A., Alexander, J., & Vong, G. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology*, 25(5), 661–671. <https://doi.org/10.1080/09515089.2011.627536>

- McDonald, K., Yin, S., Weese, T., & Sinnott-Armstrong, W. (2019). Do framing effects debunk moral beliefs? *The Behavioral and Brain Sciences*, 42(e162), 35–36. <https://doi.org/10.1017/S0140525X18002662>
- Nadelhoffer, T., & Feltz, A. (2008). The actor–observer bias and moral intuitions: Adding fuel to sinnott-armstrong's fire. *Neuroethics*, 1(2), 133–144. <https://doi.org/10.1007/s12152-008-9015-7>
- Nakamura, H., Ito, Y., Honma, Y., Mori, T., & Kawaguchi, J. (2014). Cold-hearted or cool-headed: Physical coldness promotes utilitarian moral judgment. *Frontiers in Psychology*, 5(October), 1086. <https://doi.org/10.3389/fpsyg.2014.01086>
- Olsen, O. K., Pallesen, S., & Jarle, E. (2010). The impact of partial sleep deprivation on moral reasoning in military officers. *Sleep*, 33(8), 1086–1090. <https://doi.org/10.1093/sleep/33.8.1086>
- Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, 112(13), 4170–4175. <https://doi.org/10.1073/pnas.1415250112>
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017, May). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163. <https://doi.org/10.1016/j.jesp.2017.01.006>
- Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, 17(3), 145–171. [https://doi.org/10.1016/0162-3095\(96\)00041-6](https://doi.org/10.1016/0162-3095(96)00041-6)
- Piñon, A., & Gambara, H. (2005). A meta-analytic review of framing effect: Risky, attribute and goal framing. *Psicothema*, 17(2), 325–331. <https://www.redalyc.org/articulo.oa?id=72717222>
- Schnall, S., Haidt, J., Gerald, L., Clore, J., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality & Social Psychology Bulletin*, 34(8), 1096–1109. <https://doi.org/10.1177/0146167208317771>
- Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, 27(2), 135–153. <https://doi.org/10.1111/j.1468-0017.2012.01438.x>
- Shafer-Landau, R. (2008). Defending ethical intuitionism. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 83–95). MIT Press.
- Sinnott-Armstrong, W. (2008). Framing moral intuitions. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 47–76). MIT Press.
- Spears, D., Fernández-Linsenbarth, I., Okan, Y., Ruz, M., & Felisa, G. (2018). Disfluent fonts lead to more utilitarian decisions in moral dilemmas. *Psicológica Journal*, 39(1), 41–63. <https://doi.org/10.2478/psicolj-2018-0003>
- Stanovich, K. E., & West, R. F. (1998). Individual differences in framing and conjunction effects. *Thinking & Reasoning*, 4(4), 289–317. <https://doi.org/10.1080/135467898394094>
- Steiger, A., & Anton, K. (2018). A meta-analytic re-appraisal of the framing effect. *Zeitschrift Für Psychologie*, 226(1), 45–55. <https://doi.org/10.1027/2151-2604/a000321>
- Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *Monist*, 59(2), 204–217. <https://doi.org/10.5840/monist197659224>
- Tobia, K., Buckwalter, W., & Stich, S. (2013). Moral intuitions: Are philosophers experts? *Philosophical Psychology*, 26(5), 629–638. <https://doi.org/10.1080/09515089.2012.696327>

- Tolhurst, W. (2008). Moral intuitions framed. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 77–82). MIT Press.
- Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, 16(10), 780–784. <https://doi.org/10.1111/j.1467-9280.2005.01614.x>
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer.
- Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, 25(6), 813–836. <https://doi.org/10.1080/09515089.2011.631995>
- Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, 131(1), 28–43. <https://doi.org/10.1016/j.cognition.2013.12.004>
- Wood, D. P., Harms, D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, 8(4), 454–464. <https://doi.org/10.1177/1948550617703168>