On the Stability of Moral Judgment Over Time

Paul Rehren

Faculty Adviser: Walter Sinnott-Armstrong

Department of Philosophy

12/2020

This project was submitted in partial fulfillment of the requirements for the degree of Master of Arts in the Graduate Liberal Studies Program in the Graduate School of Duke University.

Abstract

Stability over time is often seen as a signature feature of moral judgment. Yet to date, little focused empirical examination of this assumption exists. In this study, we[1] compare the stability over time of moral judgments about acts in sacrificial dilemmas, moral judgments about the items on the Moral Foundations Questionnaire, and moral judgments about the items on the Morality-as-Cooperation Questionnaire. We find that on three metrics of stability over time, the different types of moral judgment all performed similarly. We also found that changes in moral judgment, when they occurred, could not be easily explained by people changing their mind in light of reasons. We discuss potential implications of our findings for moral psychology and moral philosophy.

1 This thesis is part of a larger research project that I am involved in with Walter Sinnott-Armstrong (who served as faculty adviser of this thesis). For this reason, I will use the first-person plural throughout the thesis, even though I am its single author.

Table of Contents

List of Tables

List of Figures

Introduction

Moral judgment and stability over time

Suppose you and your friend are sitting on a bench in a park somewhere. A volunteer approaches you and asks whether either of you would consider donating to his charity. Your friend agrees, and hands over $50. Later, you ask your friend why he donated; he says: “I believe I have a moral duty to donate to charity.”

A week later, the two of you are taking a walk together when you are again approached by someone asking you to donate to charity. This time, however, your friend declines. Asked for an explanation, he says: “I don’t believe I have a moral duty to donate to charity.”

If you are anything like us, this response would likely puzzle you. You may even protest: “Wait! Didn’t you say the opposite last week?” We typically expect that other people will remain consistent in their moral judgments (Campbell 2007; Campbell and Kumar 2012).[2] Pinning down a precise definition of “moral judgment” is a complicated issue (Cullity 2016). Like much other empirical work, we here approximate moral judgments (or beliefs) as judgments (or beliefs) that “refer to the rightness or wrongness of specific acts or policies” (Waldmann, Nagel, and Wiegmann 2012, 274). Research suggests that moral beliefs are seen as more committal and less flexible than non-moral beliefs (Kreps and Monin 2014).

2 Unless we say otherwise, the empirical results cited in this thesis are based on research with modern people from Western countries (in particular the US, Canada, and Western Europe). As is common in the social sciences, we nevertheless discuss the results in a way that suggests that they apply to everyone, at all times. Clearly, it is not obvious that this is always unproblematic. For example, there is reason to think that research on so-called WEIRD (western, educated, industrialized, rich, democratic) populations will often *not* generalize to other kinds of populations (Henrich, Heine, and Norenzayan 2010). Likewise, it likely will often be a bad idea to make straight-forward inferences about the psychology of people who lived hundreds or thousands of years ago from research on modern people. Everything we say needs to be read with these caveats in mind.

Moreover, we expect moral stability over time from our leaders: Leaders who change their opinion after having taken a moral stance on an issue are perceived as more hypocritical, less effective and less worthy of support than leaders who take a non-moral stance (Kreps, Laurin, and Merritt 2017).

In fact, many have suggested that stability across time is a signature feature of the way that we think, talk and make judgments about morality. Famously, Turiel (1983) argued that one of the key differences between moral rules and conventional rules is that moral rules are considered to be universal: they are valid for everyone, at all times and in all places. Slavery is not only wrong in the U.S. today, but was wrong in ancient Greece, too.

A number of modern moral philosophers seem to agree. Authors like Mackie (1977), Smith (1994), and Darwall (1998) have suggested that (most) people are moral objectivists. That is, (most) people think that moral judgments are right and wrong in the same way that judgments about scientific fact are right and wrong, and that they apply to everyone at all times and in all places. Here is Mackie (1977, 33):

> The ordinary user of moral language means to say something about whatever it is that he characterizes morally, for example a possible action, as it is in itself [...] and not about, or even simply expressive of, his, or anyone else's, attitude or relation to it [...] one that is absolute, not contingent upon any desire or preference or policy or choice.

More recently, psychologists have gotten in on the action. Their research broadly seems to confirm the philosophers' suggestion of wide-spread moral objectivism. For example, in one study, "participants attributed almost as much objectivity to ethical statements as they did to statements of physical fact and significantly more objectivity to ethical statements than to statements about preferences or tastes" (Beebe and Sackris 2016, 912; also Goodwin and Darley

2008). Along similar lines, attitudes that people believe to be based on moral conviction are regarded as more objectively true and more universally applicable than non-moral attitudes of comparable strength (Luttrell et al. 2016; Skitka, Washburn, and Carsel 2015).

Sacrificial dilemmas and stability over time

In the last section, we have suggested that stability over time is often seen as a signature feature of moral judgment. But is this really true of all moral judgments? A lot of what we know about the psychology of moral judgment comes from studying sacrificial dilemmas. Sacrificial dilemmas are scenarios in which an agent must sacrifice something valuable to prevent the loss of something else that is valuable. In a classic example, the *Switch* dilemma, an out-of-control trolley is speeding towards five people. The only way to save them is to divert the trolley onto a separate track by hitting a switch. However, there is a sixth person on this other track who would then be run over and killed by the trolley instead (Foot 1967). Psychologists (for a review, Christensen and Gomila 2012), behavioral economists (for a review, Mudrack and Mason 2013) and philosophers (McIntyre 2019; Woollard and Howard-Snyder 2016) all have devoted significant study to the question of how people make judgments regarding the acts described in such dilemmas.

Sacrificial dilemmas are designed to be difficult. Because agents in sacrificial dilemmas must decide between sacrificing something valuable to prevent the loss of something else that is valuable, they are often faced with a choice between two acts that are both required, but which they cannot both do at the same time (McConnell 2018). For example, plausibly, saving people who are in danger is a moral obligation;

therefore, the agent in the Switch dilemma should hit the switch. At the same time, not killing people is also a moral obligation; therefore, the agent should not hit the switch. Yet the agent can only either hit the switch or not hit the switch; they cannot do both.

In light of this, it may be argued (not unreasonably) that moral judgments about sacrificial dilemmas will be atypical in terms of their stability over time. Because of their difficulty, we may expect that people will often waver between the different options described in such dilemmas (Giner-Sorolla 2020; Greene et al. 2004; Kahane et al. 2012). If so, then their judgments may remain considerably less stable over time than moral judgments about issue, scenarios or problems that do not involve choosing between the different options of a dilemma.

A finding like this would be significant for two reasons. First, it would establish moral judgments about sacrificial dilemmas as a boundary condition for the presumed general stability over time of moral judgment. Second, it would suggest that moral judgments about sacrificial dilemmas are atypical in an important way. This could have important methodological implications. If judgments about sacrificial dilemmas are not representative members of the set of all moral judgments, then "examining people's responses to sacrificial dilemmas may provide only a partial view of how people tend to confront moral situations in their everyday lives" (Bauman et al. 2014, 545). Given the large amount of research that does use sacrificial dilemmas to understand "how people tend to confront moral situations in their everyday lives", this could mean that our current picture of moral judgment is incomplete or distorted (Bauman et al. 2014).

The first aim of this study is to investigate this question. To this end, we

conducted a three-wave longitudinal study of moral judgments about sacrificial dilemmas. In each wave, participants were asked to read and make moral judgments about the same series of sacrificial dilemmas. To estimate the stability of these judgments over time, we compared participant's ratings across the different waves. We then compared these estimates to the stability over time of other moral judgments—more specifically, moral judgments made on the Moral Foundations Questionnaire (MFQ; Graham et al. 2011) and the Morality-as-Cooperation Questionnaire (MAC-Q; Curry, Jones Chesters, and Van Lissa 2019). The MFQ and the MAC-Q are two individual difference measures that are meant to assess the extent to which people rely on and endorse various facets of morality. Our hypothesis was that moral judgments about sacrificial dilemmas would be less stable over time than judgments about the items on the MFQ and the MAC-Q.

Stability over time in light of reasons

We have suggested that stability over time is often seen as a signature feature of moral judgment. However, there are circumstances in which it is both expected and reasonable for people to change their moral judgments. For example, when someone is presented with a compelling counter-argument, they should—and sometimes will—change their mind (Bloom 2010; Haidt 2012; Paxton, Ungar, and Greene 2012). Likewise, when someone realizes that one of their moral judgments is incompatible with other judgments, they should—and sometimes will—change their mind (Campbell 2017; Horne, Powell, and Hummel 2015). What this suggests is there is an important nuance to the general assumption of moral stability: Moral judgments are expected to be stable over time, *unless* they are changed in light of

reasons (compelling counter-arguments, inconsistencies with other moral judgments, etc.).

The second aim of this study is to investigate whether when people do not remain stable in their moral judgments over time, this can be explained by them changing their minds in light of reasons. Call this the *reasons hypothesis*. To test this hypothesis, we propose two approaches.

The first approach is to investigate the persistence of changes observed between waves. Recall that our study consisted of three waves; participants made judgments about the same series of sacrificial dilemmas in each wave. Similarly, Hatemi, Crabtree, and Smith (2019) used the MFQ as part of a three-wave panel study, and made their dataset publicly available. Hence, for moral judgments about sacrificial dilemmas and about items on the MFQ, we can investigate whether changes that occur between the first and second wave tend to persist when the scenario or item is judged for a third time. Changes of moral judgment in light of reasons should typically last (e.g. Horne, Powell, and Hummel 2015; Pizarro and Bloom 2003; Sauer 2012). Thus, if participants who give different ratings in the first and second wave tend to stick with their ratings in the third wave, this would provide evidence for the reasons hypothesis.

A second approach is to look at individual differences in who changes their moral judgment(s) between waves. If many changes between waves occurred in light of reasons, participants who are generally more open to reconsidering and revising their judgments should be more likely to exhibit these changes. To measure this tendency, we here use the Actively Open-minded Thinking scale (AOT; Baron 1993; Stanovich and West 1997). As a construct, actively open-minded thinking

encompasses “the cultivation of reflectiveness rather than impulsivity, the seeking and processing of information that disconfirms one's belief [...], and the willingness to change one's beliefs in the face of contradictory evidence” (Stanovich and West 1997, 346). Higher scores on the AOT have been associated with the ability to evaluate arguments objectively (Stanovich and West 1997), the tendency to seek out information before making a decision (Haran, Ritov, and Mellers 2013), providing better evidence for one’s views (Sá et al. 2005), and the rejection of traditional moral values (Pennycook et al. 2019).

Method

Design

The study consisted of three waves. Waves were separated by 6-8 days each. In each wave, participants were asked to read and make moral judgments regarding the same series of short scenarios.

Materials

Scenarios. We chose six sacrificial dilemmas from previous research on moral judgment and decision making (Greene et al. 2001; Christensen et al. 2014). Scenarios were matched closely in length and structure.

For each scenario, participants were asked whether the agent should carry out an action that would sacrifice (cause the death of) one person so that five others can live. Participants indicated their answer on a seven-point scale labeled at both endpoints, from “Definitely should do it” (= 1) to “Definitely should not do it” (= 7). The midpoint of the scale (4) was labeled “Neither”. Here are the scenarios:[3]

> Shark
>
> [Agent] is visiting the beach. He is walking along a pier when he notices a shark swimming towards the beach. [Agent] can see that this shark will attack and kill everyone in the first group of swimmers it encounters. Directly in the path of the shark is a group of five swimmers. The shark is too fast for them to get out of the water in time. There is one swimmer in the water between [agent] and the shark.
>
> [Agent] realizes that he can make loud noises that he knows will attract the shark, diverting it from its current path. If [agent] does not make loud noises, the shark will attack the five swimmers in its path, but not the one swimmer in the water between him and the shark. If [agent] makes loud noises, the shark will kill the one swimmer in the water between him and the shark, but not the five swimmers in its original path.

3 Note that the scenarios describe situations that are quite outlandish. Because of that, at the start of the study, all participants were instructed to take the scenarios at face value and to “suspend disbelief”.

Should [agent] make loud noises?

Gas

[Agent] is visiting a chemical plant when she notices that there is gas leaking into a room close by. [Agent] can see that this gas is poisonous and will kill anyone it comes into contact with. There are five workers in the room. They are unaware of the gas and will not be able to get out of the room in time. Nearby, there is a second, smaller room with one worker in it.

[Agent] realizes that there is a switch next to her, which she can push to direct the gas away from the room with the five workers and into the smaller room with one worker. If [agent] does not push the switch, the gas will kill the five workers, but not the one worker. If [agent] pushes the switch, the gas will kill the one worker, but not the five workers.

Should [agent] push the switch?

Footbridge

[Agent] is standing on a footbridge across the railroad tracks when he notices an empty railroad car out of control. [Agent] can see that this railroad car is moving so fast that it will kill anyone it hits. On the track ahead are five people. The banks are so steep that they will not be able to get off the track in time. On the bridge with [agent] is a large man wearing a backpack.

[Agent] realizes that he can push the large man off the footbridge onto the railroad tracks in front of the railroad car that is out of control. He knows that the weight of the large man and his backpack would be sufficient to stop the railroad car. If [agent] does not push the large man off the footbridge, the railroad car will kill the five people on the track, but not the large man on the footbridge. If [agent] pushes the large man off the footbridge, the railroad car will kill the large man, but not the five people on the track.

Should [agent] push the large man off the footbridge?

Transplant

[Agent] is a surgeon in a hospital. She has five patients suffering from different organ failures. All five patients are in critical condition and will die if they do not get an immediate organ transplant. [Agent] is currently performing a routine surgery on a sixth patient who is otherwise healthy. From reading the patient's file, [agent] knows that this patient's organs are compatible with the five other patients.

[Agent] realizes that she can cut the sixth patient's main artery during surgery. This will kill that patient, but will make the required transplants available for the five patients suffering from organ failure. If [agent] does not cut the one patient's main artery, this patient will live, but the five patients will die. If [agent] cuts the

one patient's main artery, this will kill that patient, but the five patients will live.

Should [agent] cut the patient's main artery?

Expedition

[Agent] is part of a group of six ecologists who are studying wildlife in a remote stretch of jungle. The group has been taken hostage by a group of paramilitary terrorists. [Agent] understands the terrorist's language and knows that their leader will kill him and the other five hostages the following morning. The hostages are guarded day and night and will not be able to escape.

One of the guards has taken a liking to [agent]. He informs him that he is willing to help [agent] and the other hostages escape. However, as an act of good faith, he wants [agent] to cut the throat of one of his fellow hostages whom the guard does not like. If [agent] does not cut the throat of the fellow hostage, [agent] and the remaining five hostages will be killed the next morning. If [agent] cuts the throat of the fellow hostage, this will kill that hostage, but [agent] and the remaining four hostages will not be killed the next morning.

Should [agent] cut the throat of the fellow hostage?

Baby

[Agent] lives in a small village that has been taken over by enemy soldiers. [Agent] understands the enemy soldiers' language and knows that they are under orders to kill every villager they find. [Agent] is hiding out in the cellar of a large house with her baby and four neighbors. Upstairs, they can hear the voices of soldiers who have come to search the house for valuables. [Agent]'s baby begins to cry loudly. [Agent] knows that this will attract the attention of the soldiers. The cellar only has one exit, so the group will not be able to escape.

[Agent] realizes that she can press her hand over her baby's mouth to stop the crying. However, she knows that this will cut off the baby's breath, ultimately suffocating and killing it. If [agent] does not press her hand over the baby's mouth, [agent], the baby, and the four neighbors will all be killed by the enemy soldiers. If [agent] presses her hand over the baby's mouth, this will kill the baby, but [agent] and her four neighbors will not be killed by the enemy soldiers.

Should [agent] press her hand over the baby's mouth?

Actively Open-minded Thinking. The original scale (Stanovich and West 1997) consists of 41 items. We here use a 10-item short form (AOT-10; Baron 2019). Items include "People should take into consideration evidence that goes against conclusions they favor" and "When we are faced with a new question, the first

answer that occurs to us is usually best." For each item, participants indicate their response on a five-point scale from "Strongly disagree" (= 1) to "Strongly agree" (= 5), with the midpoint labeled "Neither agree nor disagree" (= 3). The mean score was 3.14(SD = 0.52). Cronbach's alpha (a measure of internal consistency for multi-item instruments that are supposed to measure a single construct) was 0.68, comparable to the value of 0.75 reported by Baron (2019).

Exclusion of participants

Low effort responding would likely inflate the amount of rating changes between waves (cf. Buchanan and Scofield 2018). Therefore, an important consideration was how to exclude low effort participants. Here, we chose to use instructional attention checks (Oppenheimer, Meyvis, and Davidenko 2009). We inserted an attention check scenario into each wave. These scenarios read very similar to our sacrificial dilemmas, but end by directing participants to select a certain answer. All participants who failed to pick this answer in either of the three waves were excluded from all analyses.

Participants

Participants were recruited through the online subject pool Prolific (https://www.prolific.co). For the reliability of data gathered through Prolific for use in psychological research, see Peer et al. (2017). We restricted participation to current residents of the US and the UK whose first language was English and who had at least a 95% acceptance rating on Prolific (that is, at least 95% of their previous submissions on Prolific had been accepted).

Sample size was determined before any data analysis. For each scenario, our main quantity of interest was the proportion of participants who gave ratings on opposite sides of our scale midpoint in subsequent waves. In pilot research, the largest proportion we found for any scenario was 0.14. We therefore aimed for a sample size large enough for us to detect proportions of 0.15 within a 95% Wilson score confidence interval (Agresti and Coull 1998) of size ≤ 0.10. Power analysis (Vallejo et al. 2013) indicated that we needed a minimum sample size of 236. To allow for attrition and exclusions, we recruited an initial sample of n = 461 (307 female, 1 not specified; M(SD) = 32.8(12.4) y; 24.7% students; 93.3% UK nationals).[4]

Of the initial sample, we excluded 67 participants who failed the attention check. 356 participants returned for the second wave, of whom 68 failed the attention check. 262 participants completed the third wave within a set time window (6-8 days after completing the second wave). Of those 262 participants, we excluded 21 participants who failed the attention check. Thus, the final sample included n = 241 participants (165 female; M(SD) = 33.1(12.0) y; 27.0% students; 95.9% UK nationals).

Procedure

Participants were contacted by Prolific with an invitation to participate in the study, which was hosted on LimeSurvey (https://www.limesurvey.org). In all three waves, participants gave informed consent, received instructions, and then read and rated the sacrificial dilemmas and the attention check scenario. We randomized the

4 Information about nationality and student status was collected by Prolific.

presentation order of scenarios for each participant in the first wave. This order was then repeated in the second and third wave. To make it less likely in waves two and three that participants would remember already having read the scenarios a week earlier, thereby potentially introducing experimenter demand (Aczel, Szollosi, and Bago 2018), we changed the names of the agents between sessions, matching them in gender, ethnicity and length.

In addition, in the third wave, participants completed the AOT-10.[5] For each participant, we randomized the order in which items on the AOT were presented. Following this, participants in the third wave answered a series of self-report questions about the survey[6] and provided demographic information.

Participants were compensated $1.00 per wave. An additional bonus of $1.25 was offered and payed to participants who completed all three waves. The study was approved by the Duke University Campus Institutional Review Board.

Comparison datasets

In order to compare the stability of moral judgments about sacrificial dilemmas to other moral judgments, we turn to existing studies that have reported longitudinal data on moral judgments. We identified three studies, all of which included one of two individual difference measures: The Moral Foundations Questionnaire (MFQ; Graham et al. 2011), and the Morality-as-Cooperation Questionnaire (MAC-Q Curry, Jones Chesters, and Van Lissa 2019).

5 In addition, participants were asked to complete two other individual difference measures, the REI-10 (Epstein et al. 1996) and the CRT-7 (Frederick 2005; Toplak, West, and Stanovich 2014). However, since they do not bear on the issue discussed here, we will not mention them further.

6 Since these self-report questions do not bear on the issues discussed here, we will not mention them further.

The MFQ consists of 30 items. It measures the extent to which people endorse the five psychological foundations of morality posited by Moral Foundations Theory (Haidt and Graham 2007): Harm/Care, Fairness/Reciprocity, Ingroup/Loyalty, Authority/Respect, and Purity/Sanctity. Half of the items on the MFQ are short moral statements that respondents are asked to make a judgment about. Items include "One of the worst things a person could do is hurt a defenseless animal", "I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing" and "People should not do things that are disgusting, even if no one is harmed." Respondents rate each statement on a scale from "Strongly disagree" to "Strongly agree."

The MAC-Q has a similar structure to the MFQ, but is based on Morality-as-Cooperation theory. Morality-as-Cooperation theory posits that morality is "a collection of biological and cultural solutions to the problems of cooperation and conflict recurrent in human social life" (Curry 2016, 29). It identifies seven distinct types of cooperation, each of which gives rise to a distinct moral domain. The seven domains are Family, Group, Reciprocity, Heroism, Deference, Fairness, and Property. The MAC-Q is supposed to measure the extent to which people endorse each of these seven domains of morality. It includes 42 items, half of which are short moral statements that respondents are asked to make a judgment about. Items include "People should be willing to do anything to help a member of their family", "You have an obligation to help those who have helped you" and "It's acceptable to steal food if you are starving." Respondents rate each statement on a scale from "Strongly disagree" to "Strongly agree."

To date, three studies have reported longitudinal data on the MFQ (Curry,

Jones Chesters, and Van Lissa 2019; Graham et al. 2011; Hatemi, Crabtree, and Smith 2019). Graham et al. and Curry et al. were interested in the test-retest reliability of the MFQ. Graham et al. administered the MFQ to 123 participants (86 female; M = 20.1 y; 100% students) twice, with an average delay of 37.4 days between waves. Curry et al. had 137 participants (69 female; M(SD) = 53.1(15.8) y); their delay between the first and second wave was one month. Hatemi et al. used the MFQ as part of a three-wave panel study testing for a causal relationship between different moral foundations and political orientation. Their final sample includes 127 participants. The first and second wave were separated by approximately 2.5 years; the second and third wave were separated by approximately one year.

To date, one paper has reported longitudinal data on the MAC-Q (Curry, Jones Chesters, and Van Lissa 2019). Again, Curry et al. were interested in test-retest reliability. They used the same sample and procedure as in their investigation of the test-retest reliability of the MFQ (described above).

Analysis

All analyses reported in this paper were carried out in R (R Core Team 2020). Hatemi et al. made their dataset public (available at: https://doi.org/10.7910/DVN/O6VHZZ). Thus, we could run our own analyses on their data.

Test-retest correlations

For the moral judgment items on the MFQ, Graham et al. and Curry et al. report Pearson's product-moment correlation coefficients of first wave and second wave ratings averaged across foundation. Table 1 shows their results. For the judgment items on the MAC-Q, Curry et al. provide Pearson's product-moment correlation coefficients of first wave and second wave ratings averaged across moral domain. Table 2 shows their results.

For the Hatemi et al. dataset, we calculated Pearson's product-moment correlations between first wave and second wave ratings, and between second wave and third wave ratings of the MFQ moral judgment items. We report the results averaged across moral foundations (Table 1). To get the average correlation coefficients, we first calculated the correlations for each judgment item individually. Individual item correlation coefficients were then transformed to Fisher *z* coefficients, averaged, and back-transformed (Silver and Dunlap 1987).

We followed the same procedure for our own data. In addition to the overall correlations (averaged over dilemmas), we also report Pearson's product-moment correlation coefficients for the individual dilemma ratings. Table 3 shows our results.

MFQ				
	Graham et al. (2011)	**Hatemi, Crabtree, and Smith (2019)**		**Curry, Jones Chesters, and Van Lissa (2019)**
n	123	127		137
Delay	37.4 days	2.5 y (1st wave/ 2nd wave)	1 y (2nd wave/ 3rd wave)	30 days
		r		
Harm	0.71	0.69 [0.62, 0.75]	0.81 [0.76, 0.85]	0.51
Fairness	0.68	0.71 [0.64, 0.77]	0.73 [0.66, 0.78]	0.46
Ingroup	0.69	0.72 [0.66, 0.78]	0.76 [0.7, 0.81]	0.62
Authority	0.71	0.68 [0.6, 0.74]	0.81 [0.76, 0.85]	0.59
Purity	0.82	0.73 [0.66, 0.78]	0.83 [0.79, 0.87]	0.75

Table 1. Pearson's product-moment correlations for the moral judgment items on the MFQ averaged across foundation. Square brackets show 95% confidence intervals.

MAC-Q	
n	137
Delay	30 days
	r
Family	0.87
Group	0.74
Reciprocity	0.71
Heroism	0.78
Deference	0.71
Fairness	0.66
Property	0.71

Table 2. Pearson's product-moment correlations for the moral judgment items on the MAC-Q averaged across moral domain (Curry, Jones Chesters, and Van Lissa 2019).

Sacrificial dilemmas		
n	241	
Delay	6-8 days (1^{st} wave/ 2^{nd} wave)	6-8 days (2^{nd} wave/ 3^{rd} wave)
	r	
Shark	0.62 [0.53, 0.69]	0.68 [0.61, 0.74]
Gas	0.58 [0.49, 0.66]	0.65 [0.57, 0.72]
Footbridge	0.74 [0.68, 0.80]	0.85 [0.81, 0.88]
Transplant	0.66 [0.59, 0.73]	0.69 [0.62, 0.75]
Expedition	0.64 [0.55, 0.71]	0.74 [0.68, 0.79]
Baby	0.74 [0.68, 0.80]	0.77 [0.72, 0.82]
Overall	0.67 [0.59, 0.73]	0.74 [0.68, 0.79]

Table 3. Pearson's product-moment correlations between sacrificial dilemmas ratings. Square brackets show 95% confidence intervals.

Rating shifts and rating reversals

Next, we were interested in two quantities (cf. Andow 2016). We define a *rating shift* as any change in rating between waves. We define a *rating reversal* as any rating shift that crosses the scale-midpoint. A rating reversal indicates a change of moral judgment where the two judgments are polar opposites (push the switch vs. do not push the switch; agree vs. disagree that there is an obligation to help those who have helped you). In contrast, moral judgments that shift differ in strength, but may retain their polarity (push the switch vs. definitely push the switch; strongly agree vs. slightly agree that there is an obligation to help those who have helped you).

We calculated the proportions of rating shifts and rating reversals between the first and second wave, and the second and third wave for our data and for the MFQ data by Hatemi et al. For our data, we report results by scenario and overall (Table

4). For the data by Hatemi et al., we report results averaged across foundation (Table 5).

	MFQ			
n	127		127	
Delay	2.5 y (1st wave/ 2nd wave)	1 y (2nd wave/ 3rd wave)	2.5 y (1st wave/ 2nd wave)	1 y (2nd wave/ 3rd wave)
	Shifts		**Reversals**	
Harm	0.44	0.34	0.13	0.10
Fairness	0.46	0.40	0.16	0.14
Ingroup	0.48	0.43	0.13	0.14
Authority	0.49	0.38	0.17	0.11
Purity	0.51	0.45	0.20	0.15

Table 4. Proportions of MFQ moral judgment item rating shifts and rating reversal. Based on data by Hatemi, Crabtree, and Smith (2019).

	Sacrificial dilemmas			
n	241		241	
Delay	6-8 days (1st wave/ 2nd wave)	6-8 days (2nd wave/ 3rd wave)	6-8 days (1st wave/ 2nd wave)	6-8 days (2nd wave/ 3rd wave)
	Shifts		**Reversals**	
Shark	0.55 [0.48, 0.61]	0.44 [0.38, 0.51]	0.17 [0.13, 0.22]	0.17 [0.13, 0.22]
Gas	0.56 [0.49, 0.62]	0.52 [0.46, 0.58]	0.20 [0.15, 0.25]	0.16 [0.12, 0.21]
Footbridge	0.44 [0.38, 0.51]	0.39 [0.33, 0.45]	0.11 [0.07, 0.15]	0.07 [0.05, 0.11]
Transplant	0.28 [0.23, 0.34]	0.26 [0.21, 0.32]	0.06 [0.04, 0.1]	0.06 [0.03, 0.1]
Expedition	0.55 [0.48, 0.61]	0.5 [0.44, 0.56]	0.17 [0.13, 0.23]	0.15 [0.11, 0.2]
Baby	0.52 [0.46, 0.59]	0.46 [0.4, 0.52]	0.11 [0.08, 0.16]	0.1 [0.07, 0.14]
Overall	0.48	0.43	0.14	0.12

Table 5. Proportions of sacrificial dilemma rating shifts and rating reversals. Square brackets show 95% Wilson score confidence intervals.

The reasons hypothesis

According to the reasons hypothesis, when people change their moral judgments,

they usually do so in light of reasons. To test this hypothesis, we use two approaches. First, changes to moral judgments in light of reasons would generally be expected to last. Therefore, for the MFQ data by Hatemi et al. and for our own data, we look at the persistence of rating changes that occurred between the first and second wave. To this end, we fit a pair of linear mixed-effect models (Bates et al. 2015), predicting rating shifts that occurred between the first and second wave by rating shifts that occurred between the second and third wave. We added random intercepts for scenario and participant (Baayen, Davidson, and Bates 2008). For the sacrificial dilemmas, there was a strong effect, b = -0.39 [-0.43, -0.34], such that rating shifts between the first and second wave negatively predicted rating shifts between the second and third wave. We found a similar effect for the MFQ, b = -0.32 [-0.35, -0.28]. Here too, rating shifts that occurred between the first and second wave negatively predicted rating shifts between the second and third wave.

For rating reversals, we fit a pair of binomial mixed-effects model with logit link (Bates et al. 2015), predicting rating reversals that occurred between the first and second wave by rating reversals that occurred between the second and third wave. We included random intercepts for participant and scenario. We again found strong effects for both the sacrificial dilemmas and the moral judgment items on the MFQ. The odds of a participant exhibiting a rating reversal between the second and third wave were higher if they had previously reversed their rating between the first and second wave, OR = 7.61 [4.92, 11.9] and OR = 5.18 [3.72, 7.23].

Our second approach to investigating the reasons hypothesis was to test for a relationship between actively open-minded thinking on the one hand, and rating changes on the other hand. Again, we used (generalized) linear mixed-effects

models. We fit a linear mixed-effect model to our sacrificial dilemma data, predicting rating shift magnitude by mean-standardized scores on the AOT-10. We also fit a binomial mixed-effects model with logit link, predicting rating reversal by mean-standardized scores on the AOT-10. Both models included a dummy-coded variable that indicated whether a given rating shift occurred between the first and second wave, or between the second and third wave. Moreover, both models included a random intercept for scenario. Scores on the AOT-10 predicted neither rating shift magnitude, b = 0.04 [-0.01, 0.08], nor rating reversal, OR = 0.99 [0.88, 1.11].

Discussion

Sacrificial dilemmas and stability over time

In this study, we investigated the stability of moral judgment over time. For repeated moral judgments about sacrificial dilemmas, items on the MFQ and items on the MAC-Q, we looked at three metrics of stability: test-retest correlation, rating shifts and rating reversals. How do the different moral judgments compare?

We start with test-retest correlation. Figure 1 compares test-retest correlation coefficients obtained for the moral judgment items on the MFQ averaged across foundation (Table 1), the moral judgment items on the MAC-Q averaged across moral domain (Table 2), and our overall sacrificial dilemma ratings (Table 3). It shows that our results lie squarely in the middle of this range. Moreover, most of the other test-retest correlations fall within the 95% confidence intervals of our estimates. This suggests that on this metric of stability over time, moral judgments of sacrificial dilemmas are not atypical.

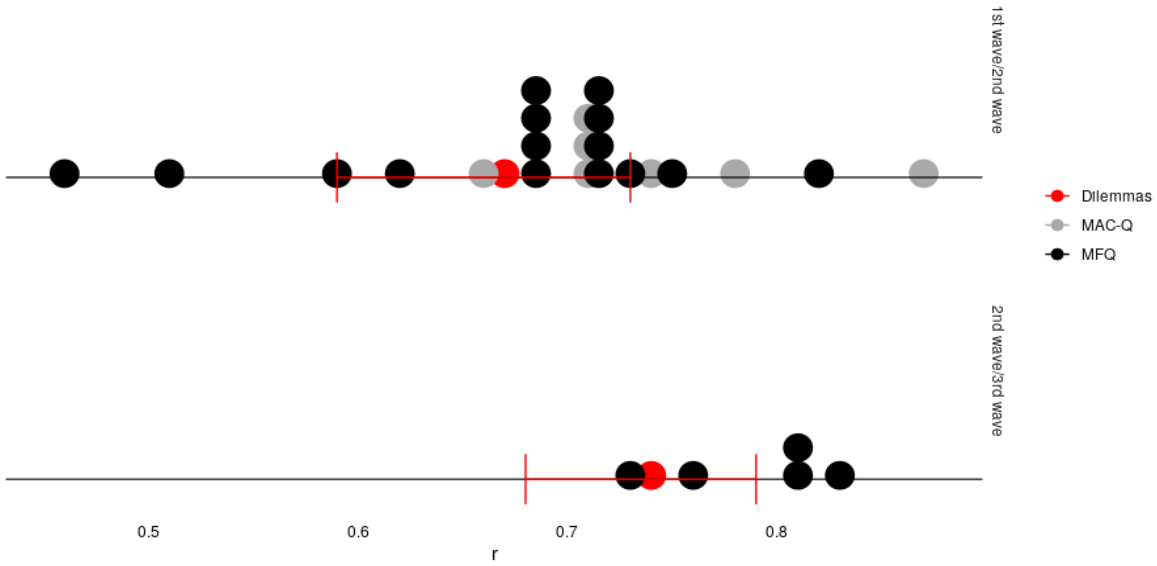

Figure 1. Dotplot showing the range of test-retest correlation coefficients found for moral judgments on the MFQ, MAC-Q and about sacrificial dilemmas.

Figure 2 compares the proportions of MFQ rating shifts and rating reversals (Table 4) with the overall proportions we obtained on our sacrificial dilemmas (Table 5). Again, all values are quite close together, further suggesting that moral judgments of sacrificial dilemmas are not atypical when it comes to their stability over time.

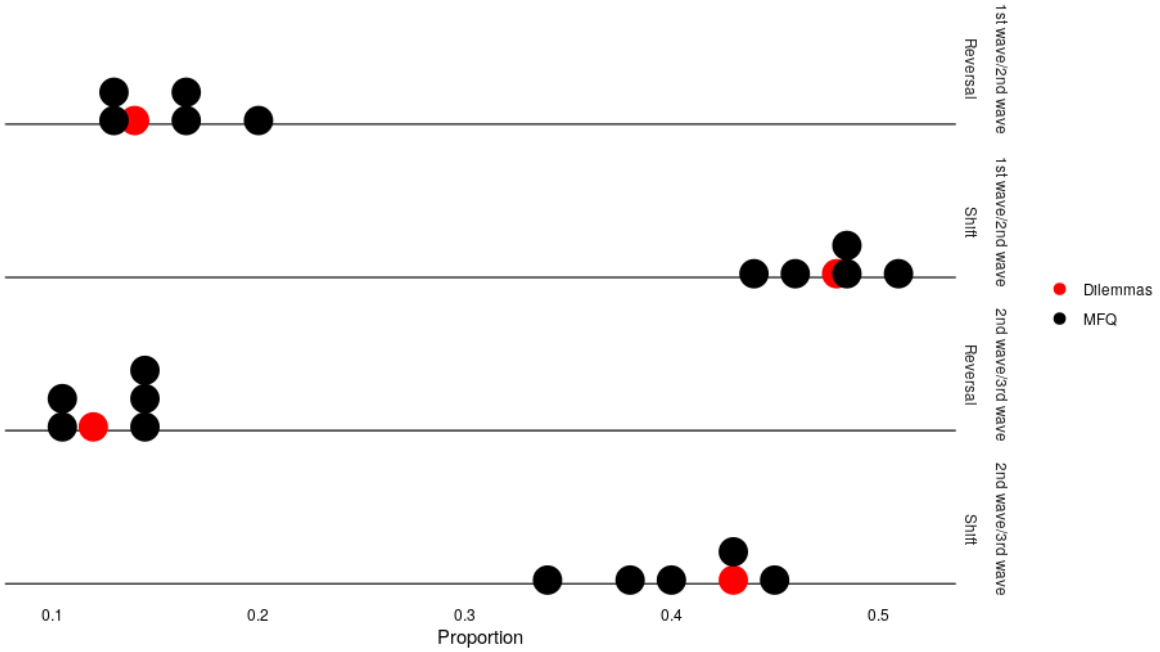

Figure 2. Dotplot showing the range of proportions of rating shifts and rating reversals found for moral judgments on the MFQ and about sacrificial dilemmas.

Stability over time in light of reasons

Next, we turn to the reasons hypothesis. According to the reasons hypothesis, when people change their moral judgments, they usually do so in light of reasons. Our study did not find much evidence for this hypothesis. First, for both judgments of sacrificial dilemmas and judgments of items on the MFQ, changes in rating that occurred between the first and second wave did not generally persist. Instead, the more participants shifted between the first and second wave, they more they tended to shift between the second and third wave. Likewise, participants who exhibited a rating reversal between the first and second wave were much more like to again reverse their rating between the second and third wave. Both speak against the reasons hypothesis, because we would generally expect that changes of moral

judgment in light of reasons will last (Horne, Powell, and Hummel 2015; Pizarro and Bloom 2003; Sauer 2012).

Second, for judgments of sacrificial dilemmas, scores on the AOT-10 did not predict the occurrence of rating shifts and rating reversals at all. The AOT-10 is supposed to measure “the cultivation of reflectiveness rather than impulsivity, the seeking and processing of information that disconfirms one's belief [...], and the willingness to change one's beliefs in the face of contradictory evidence” (Stanovich and West 1997, 346). If many of these rating changes had occurred in light of reasons, we would have expected that participants who are in general more open to reconsidering and revising their judgments would be more likely to exhibit these changes. Yet this is not what we found.

Implications for moral psychology

This research has potential implication for both moral psychology (a branch of psychology that studies human thought and behavior in moral contexts) and moral philosophy (a branch of philosophy that studies what is right and wrong, how people should live their lives, and what the nature of morality is).

First, moral psychology. Stability over time is often seen as a signature feature of moral judgment. And yet, the single most researched type of moral scenario, the sacrificial dilemma, is deliberately designed to be difficult and to "get people to waver between two choices" (Giner-Sorolla 2020). This raises the possibility that moral judgments about acts in sacrificial dilemmas are less stable over time than most other moral judgments, and that they are hence atypical in an important way.

However, the results of this study do not support this possibility. On three metrics of stability over time (test-retest correlation, rating shifts, rating reversals), moral judgments about sacrificial dilemmas performed similar to judgments regarding statements related to various facets of morality, including family, reciprocity, fairness, property, harm, authority, and purity. Our results suggest that to the extent that other moral judgments are stable over time, so too are moral judgments about sacrificial dilemmas. At least in this respect, then, there is no problem in using them to understand how people tend to confront moral situations in their everyday lives—though of course, other problems may still exist (Bauman et al. 2014).

Implications for moral philosophy

Our results could have serious implications for moral philosophy. This is because they suggest that a significant percentage of moral judgments are unreliable. Recall that averaged across moral foundation, 13-20% of participants in Hatemi et al. reversed their judgments between the first and second wave. Between the second and third wave, it was 10-15%. Moreover, on average, 14% of participants reversed their judgment on a given sacrificial dilemmas between our first and second wave. Between our second and third wave, it was 12%.

Because rating reversals indicate a change of moral judgment where the two judgments are polar opposites, in all of these cases, 10-20% of people are making incompatible moral judgments about the same scenario or item. We have already mentioned that there are circumstances in which this would be unproblematic (even expected); namely, if most of the reversals occurred in light of reasons. However, our results do not suggest that this is the case.

Suppose, then, that is not the case. In that case, our results suggest that in a number of domains of morality (sacrificial dilemmas, harm, fairness, loyalty, authority, sanctity), 10-20% of people will make incompatible moral judgments at different points in time without good reason. Because the judgments are opposites, at most one of them can be correct. Yet processes that fail to consistently lead to correct judgments are unreliable, by definition. Therefore, if our results cannot be explained in terms of changes in light of reason, then they suggest that 10-20% of moral judgments (about sacrificial dilemmas, harm, fairness, loyalty, authority, and sanctity) are unreliable.

What would this imply for moral philosophy? First, it may spell trouble for

moral intuitionism. Second, many moral philosophers may have to reconsider the way they typically do their work.

Moral intuitionism

Moral intuitionism is a theory about moral statements. While there are various different flavors of moral intuitionism, all moral intuitionists claim that moral statements "attribute irreducible, objective, evaluative properties to things, and that we sometimes know these claims to be true intuitively" (Huemer 2009, 192). Most modern moral intuitionists spell out this second commitment (that we sometimes know that moral claims are true intuitively) by claiming that at least some moral claims are justified non-inferentially—that is, justified even in the absence of independent confirmation (Audi 2008; Huemer 2005; Shafer-Landau 2003; cf. Tropman 2011). However, if our moral judgments are often not reliable, then a given moral claim would seem not to be justified without at least some independent confirmation (Nadelhoffer and Feltz 2008; Sinnott-Armstrong 2008).

For an analogy, consider a box of 100 thermometers. You know that many (say, 20) of the thermometers do not show the temperature accurately. You take one of the thermometers out of the box at random. Are you justified in trusting what this particular thermometer says? The answer seems to be "No". Unless you have some independent reason to think that this particular thermometer is accurate, the fact that many of the thermometers in the box are not accurate should make you skeptical of this one, too. This point applies no matter which thermometer you pick out of the box, so all of them are subject to the same doubts.

If this is correct, then, by analogy, if enough moral judgments are unreliable,

then no moral judgment should be trusted without independent confirmation. However, that conclusion conflicts directly with the central claim made by moral intuitionists that moral judgments are justified non-inferentially, since independent confirmation requires inference. Thus, temporal instability poses a serious challenge to moral intuitionism, provided it affects a wide enough variety of moral judgments.

How should we do moral philosophy?

It is widely accepted that one of the main ways (perhaps the main way) to do moral philosophy is to use responses to particular moral problems, issues, or scenarios (cf. Deutsch 2010; Williamson 2007). Such responses have been called the "data of ethics" (Ross 2002, 43). The most widely used method that relies on judgments of particular cases is reflective equilibrium (Kamm 1993; Rawls 1971). It consists in "working back and forth among our considered judgments [...] about particular instances or cases, the principles or rules that we believe govern them, and the theoretical considerations that we believe bear on accepting these considered judgments, principles, or rules, revising any of these elements wherever necessary in order to achieve an acceptable coherence among them" (Daniels 2020).

The problem for these methods that is posed by a lack of stability over time of moral judgments is then straight-forward: If our moral judgments are not generally stable over time, then they do not make for a solid foundation on top of which to build moral theory (cf. Alexander, Mallon, and Weinberg 2014; Paulo 2020). As Jesus is reported to have said (Matthew 7:26), only a foolish man would build his house on sand. Why? Because sand shifts, just like moral judgments that vary from week to week. In order to have a secure foundation, moral philosophers who rely on their

moral judgments about particular cases may need to reconsider the way they build their houses.

How much unreliability is too much?

Some philosophers have doubted that 10-20% of moral judgments being unreliable would really be all that bad. For example, in the context of her meta-analysis of moral framing effects, Demaree-Cotton (2016) claims that 20% would not constitute "a large probability of error" (19), and even adds, "I might be happy to accept the possibility that my moral judgments are off-track 20% of the time" (17). This sentiment is echoed by May (2019, 9) and by Sauer (2018, 76).

We disagree, and think that 20% (or even 10%) would be a problem. One reason for this is that morality is so important. Mistakes in moral judgments can lead to hurt feelings, antagonism, bad laws, and even war. Therefore, it is crucial that we get them right. Moreover, we would not accept scientific judgments if they were unreliable 20% (or even 10%) of the time. Yet the stakes in science are frequently much lower than in morality. This makes it hard to see why we should be happy to accept it if 10-20% of moral judgments could not be trusted.

Limitations and directions for future research

Our research is subject to several limitations. First, all of the data that we used were gathered in online convenience samples of European and U.S. American participants. This seriously limits the extent to which our findings can be generalized to other populations (Henrich, Heine, and Norenzayan 2010).

Second, the datasets we used in this study differed considerably in terms of their delay periods between waves of data collection. At the low end is our own survey, with 6-8 days between waves. At the high end is the dataset by Hatemi, Crabtree, and Smith (2019), where waves were separated by 2.5 years and 1 year, respectively. Had Hatemi et al. used delay periods more similar to ours, perhaps their data would have looked different. In particular, it is a distinct possibility that their participants would have exhibited far fewer rating shifts and rating changes on smaller time-scales. Since Graham et al. and Curry et al. reported similar test-retest correlations to what we obtained for the data by Hatemi et al. while using time delays much closer to ours, there is some reason to doubt this. Nevertheless, future research will need to investigate this possibility.

Third, and most notably, we have very little conclusive to say about mechanism. While we did investigate the reasons hypothesis, our tests could only have confirmed this hypothesis; they cannot reject it. In other words: If we had found that moral judgment changes tend to persist, or that there is a positive relationship between actively open-minded thinking and the occurrence of judgment changes, this would have given us reason to think that the reasons hypothesis may be true. However, the absence of such findings does not rule out the reasons hypothesis. For example, it is possible that most of the rating changes that occurred between the first

and second wave did occur in light of reasons, but that on further reflection, participants changed their mind again between the second and third wave.

One way to do better would be to directly manipulate reflective engagement. For example, participants could be asked to read a series of arguments against one of their moral judgments between waves (Stanley et al., 2018). If the reasons hypothesis were true, we would expect these moral judgments to change more than judgments which participants did not read arguments against.

An alternative explanation for our findings is that they were (largely) due to participants who were not taking their task seriously. Behavior like this is not at all uncommon in online survey research (Huang et al. 2012; Buchanan and Scofield 2018). In our survey, we took steps to mitigate this by using a fairly strict exclusion criterion. However, as far as we can tell, similar quality controls were not put in place by the authors of the other datasets, meaning that this explanation remains a live possibility.

A third possibility is that our findings are the result of external influences on moral judgment. A whole host of such influences on moral judgment have been reported in previous literature, including framing effects (Rehren and Sinnott-Armstrong under review), cues to cleanliness (Helzer and Pizarro 2011), sleep deprivation (Killgore et al. 2007), blood alcohol level (Duke and Bègue 2015), social conformity effects (Chituc and Sinnott-Armstrong 2020), room temperature (Nakamura et al. 2014), noise (Seidel and Prinz 2013) and music (Ansani, D'Errico, and Poggi 2017), and even the level of air pollution (Li et al. 2019). Thus, the thought is that participants may have made different moral judgments in different waves because in one of the waves, they were tired. Or drunk. Or in a cold room. Or

their computer was on a dirty desk.

We encourage future research to overcome these limitations.

Analysis code

All analyses reported in this paper were carried out in R (R Core Team 2020). Below, we provide the full R code necessary to reproduce all analyses, tables and figures reported in this paper.

```
# Load required packages
library(lme4); library(lmerTest)
library(Hmisc)
library(ggplot2)

## Our data
# Read in data
data.s1 <- read.csv("RESULTS S1.csv", header = TRUE, stringsAsFactors = FALSE)
data.s2 <- read.csv("RESULTS S2.csv", header = TRUE, stringsAsFactors = FALSE)
data.s3 <- read.csv("RESULTS S3.csv", header = TRUE, stringsAsFactors = FALSE)

# Exclude participants
data.s1 <- data.s1[data.s1$STATUS == "APPROVED", ]; n.s1.approved <- nrow(data.s1)
data.s1 <- data.s1[data.s1$ATTENTION_CHECK == 7, ]; n.s1.checked <- nrow(data.s1)
data.s2 <- data.s2[data.s2$STATUS == "APPROVED", ]; n.s2.approved <- nrow(data.s2)
data.s2 <- data.s2[data.s2$ATTENTION_CHECK == 1, ]; n.s2.checked <- nrow(data.s2)
data.s3 <- data.s3[data.s3$STATUS == "APPROVED", ]; n.s3.approved <- nrow(data.s3)
data.s3 <- data.s3[data.s3$ATTENTION_CHECK == 1, ]; n.s3.checked <- nrow(data.s3)

# Create SESSION column
data.s1$SESSION <- 1; data.s2$SESSION <- 2; data.s3$SESSION <- 3

## Clean-up, preparation
# Convert answer codes to numerical
data.s3[data.s3 == "A1"] <- 1
data.s3[data.s3 == "A2"] <- 2
data.s3[data.s3 == "A3"] <- 3
data.s3[data.s3 == "A4"] <- 4
data.s3[data.s3 == "A5"] <- 5
data.s3[, c("REI1", "REI2", "REI3", "REI4", "REI5", "REI6",
            "REI7", "REI8", "REI9", "REI10", "AOT1", "AOT2",
            "AOT3", "AOT4", "AOT5", "AOT6", "AOT7", "AOT8",
            "AOT9", "AOT10")] <-
  sapply(data.s3[, c("REI1", "REI2", "REI3", "REI4", "REI5",
                     "REI6", "REI7", "REI8", "REI9", "REI10",
                     "AOT1", "AOT2", "AOT3", "AOT4", "AOT5",
                     "AOT6", "AOT7", "AOT8", "AOT9", "AOT10")],
         function(x) as.numeric(x))

# Reverse code REI and AOT items
data.s3[, c("REI1", "REI2")] <-
  sapply(data.s3[, c("REI1", "REI2")], function(x) -(x - 6))
data.s3[, c("AOT3", "AOT5", "AOT7", "AOT8")] <-
  sapply(data.s3[, c("AOT3", "AOT5", "AOT7", "AOT8")], function(x) -(x - 6))

# Convert CRT answers to binary
data.s3$CRT1 <- ifelse(data.s3$CRT1 == 5, 1, 0)
data.s3$CRT2 <- ifelse(data.s3$CRT2 == 5, 1, 0)
data.s3$CRT3 <- ifelse(data.s3$CRT3 == 47, 1, 0)
data.s3$CRT4 <- ifelse(data.s3$CRT4 == 4, 1, 0)
data.s3$CRT5 <- ifelse(data.s3$CRT5 == 29, 1, 0)
```

```r
data.s3$CRT6 <- ifelse(data.s3$CRT6 == 20, 1, 0)
data.s3$CRT7 <- ifelse(data.s3$CRT7 == "c", 1, 0)

# Create overall NFC, FI, AOT and CRT scores (mean-scaled)
data.s3$NFC <- scale(apply(data.s3[,c("REI1", "REI2", "REI3", "REI4", "REI5")],
                           1, function(x) mean(x)))[,1]
data.s3$FI <- scale(apply(data.s3[,c("REI6", "REI7", "REI8", "REI9", "REI10")],
                          1, function(x) mean(x)))[,1]
data.s3$AOT <- scale(apply(data.s3[,c("AOT1", "AOT2", "AOT3", "AOT4", "AOT5",
                                      "AOT6", "AOT7", "AOT8", "AOT9", "AOT10")],
                           1, function(x) mean(x)))[,1]
data.s3$CRT <- scale(apply(data.s3[,c("CRT1", "CRT2", "CRT3", "CRT4",
                                      "CRT5", "CRT6", "CRT7")],
                           1, function(x) sum(x)))[,1]

# Convert categorical variables to factors
data.s1$SEX <- as.factor(data.s1$SEX)
data.s2$SEX <- as.factor(data.s2$SEX)
data.s3$SEX <- as.factor(data.s3$SEX)
data.s1$NATIONALITY <- as.factor(data.s1$NATIONALITY)
data.s2$NATIONALITY <- as.factor(data.s2$NATIONALITY)
data.s3$NATIONALITY <- as.factor(data.s3$NATIONALITY)
data.s1$EMPLOYMENT <- as.factor(data.s1$EMPLOYMENT)
data.s2$EMPLOYMENT <- as.factor(data.s2$EMPLOYMENT)
data.s3$EMPLOYMENT <- as.factor(data.s3$EMPLOYMENT)
data.s1$STUDENT_STATUS <- as.factor(data.s1$STUDENT_STATUS)
data.s2$STUDENT_STATUS <- as.factor(data.s2$STUDENT_STATUS)
data.s3$STUDENT_STATUS <- as.factor(data.s3$STUDENT_STATUS)
data.s3$RELIGION <- as.factor(data.s3$RELIGION)
data.s3$EDUCATION <- as.factor(data.s3$EDUCATION)
data.s3$ETHICS_COURSE <- as.factor(data.s3$ETHICS_COURSE)

# Mean-scale demographic variables
data.s3$RELIGIOSITY <- scale(data.s3$RELIGIOSITY)[,1]
data.s3$SES <- scale(data.s3$SES)[,1]
data.s3$POLITICAL_ATTITUDE <- scale(data.s3$POLITICAL_ATTITUDE)[,1]

# Concatenate datasets from different waves
data.s1 <- data.s1[, c("ID", "STATUS", "CODE", "AGE", "SEX",
                       "NATIONALITY", "EMPLOYMENT", "STUDENT_STATUS",
                       "SHARK", "GAS", "FOOTBRIDGE", "TRANSPLANT",
                       "EXPEDITION", "BABY",
                       "ATTENTION_CHECK", "SESSION")]
data.s2 <- data.s2[, c("ID", "STATUS", "CODE", "AGE", "SEX",
                       "NATIONALITY", "EMPLOYMENT", "STUDENT_STATUS",
                       "SHARK", "GAS", "FOOTBRIDGE", "TRANSPLANT",
                       "EXPEDITION", "BABY",
                       "ATTENTION_CHECK", "SESSION")]
data.s1.s2 <- rbind(data.s1[is.element(data.s1$ID, data.s2$ID), ], data.s2)
data.s1.s2 <- data.s1.s2[order(data.s1.s2$ID, data.s1.s2$SESSION),]

data.s3 <- data.s3[, c("ID", "STATUS", "CODE", "AGE", "SEX",
              "NATIONALITY", "EMPLOYMENT", "STUDENT_STATUS",
             "RELIGION", "RELIGIOSITY", "POLITICAL_ATTITUDE",
             "SES", "EDUCATION", "ETHICS_COURSE",
             "SHARK", "GAS", "FOOTBRIDGE", "TRANSPLANT",
             "EXPEDITION", "BABY",
             "ATTENTION_CHECK", "SESSION",
             "REI1", "REI2", "REI3", "REI4", "REI5",
             "REI6",  "REI7", "REI8", "REI9", "REI10",
             "CRT1", "CRT2", "CRT3", "CRT4", "CRT5", "CRT6", "CRT7",
             "AOT1", "AOT2", "AOT3", "AOT4", "AOT5",
             "AOT6", "AOT7", "AOT8", "AOT9", "AOT10",
```

```
             "NFC", "FI", "CRT", "AOT")]

data.s2 <- data.s2[is.element(data.s2$ID, data.s3$ID),]
data.s1 <- data.s1[is.element(data.s1$ID, data.s3$ID),]

data.s1 <- cbind(data.s1[, c("ID", "STATUS", "CODE", "AGE", "SEX",
                             "NATIONALITY", "EMPLOYMENT", "STUDENT_STATUS")],
                 data.s3[, c("RELIGION", "RELIGIOSITY", "POLITICAL_ATTITUDE",
                             "SES", "EDUCATION", "ETHICS_COURSE")],
                 data.s1[, c( "SHARK", "GAS", "FOOTBRIDGE", "TRANSPLANT",
                              "EXPEDITION", "BABY",
                              "ATTENTION_CHECK", "SESSION")],
                 data.s3[, c("REI1", "REI2", "REI3", "REI4", "REI5",
                             "REI6", "REI7", "REI8", "REI9", "REI10",
                             "CRT1", "CRT2", "CRT3", "CRT4",
                             "CRT5", "CRT6", "CRT7",
                             "AOT1", "AOT2", "AOT3", "AOT4", "AOT5",
                             "AOT6", "AOT7", "AOT8", "AOT9", "AOT10",
                             "NFC", "FI", "CRT", "AOT")])
data.s2 <- cbind(data.s2[, c("ID", "STATUS", "CODE", "AGE", "SEX",
                             "NATIONALITY", "EMPLOYMENT", "STUDENT_STATUS")],
                 data.s3[, c("RELIGION", "RELIGIOSITY", "POLITICAL_ATTITUDE",
                             "SES", "EDUCATION", "ETHICS_COURSE")],
                 data.s2[, c( "SHARK", "GAS", "FOOTBRIDGE", "TRANSPLANT",
                              "EXPEDITION", "BABY",
                              "ATTENTION_CHECK", "SESSION")],
                 data.s3[, c("REI1", "REI2", "REI3", "REI4", "REI5",
                             "REI6", "REI7", "REI8", "REI9", "REI10",
                             "CRT1", "CRT2", "CRT3", "CRT4",
                             "CRT5", "CRT6", "CRT7",
                             "AOT1", "AOT2", "AOT3", "AOT4", "AOT5",
                             "AOT6", "AOT7", "AOT8", "AOT9", "AOT10",
                             "NFC", "FI", "CRT", "AOT")])

data <- rbind(data.s1, data.s2, data.s3)
data <- data[order(data$ID, data$SESSION),]

# Define N
N <- nrow(data)/3

## Create dataframe for rating shifts and rating reversals
vars <- c("ID", "AGE", "SEX", "NATIONALITY", "EMPLOYMENT",
          "STUDENT_STATUS", "RELIGION",
          "RELIGIOSITY", "POLITICAL_ATTITUDE",
          "SES", "EDUCATION", "ETHICS_COURSE",
          "NFC", "FI", "CRT", "AOT")
n <- c("WAVE", "SCENARIO", vars, "FIRST_SESSION", "SECOND_SESSION")

df.full <- rbind(setNames(cbind("S1S2", "SHARK",
                    data[data$SESSION == 1, c(vars, "SHARK")],
                    data$SHARK[data$SESSION == 2]), n),
      setNames(cbind("S1S2", "GAS",
                     data[data$SESSION == 1, c(vars, "GAS")],
                     data$GAS[data$SESSION == 2]), n),
      setNames(cbind("S1S2", "FOOTBRIDGE",
                     data[data$SESSION == 1, c(vars, "FOOTBRIDGE")],
                     data$FOOTBRIDGE[data$SESSION == 2]), n),
      setNames(cbind("S1S2", "TRANSPLANT",
                     data[data$SESSION == 1, c(vars, "TRANSPLANT")],
                     data$TRANSPLANT[data$SESSION == 2]), n),
      setNames(cbind("S1S2", "EXPEDITION",
                     data[data$SESSION == 1, c(vars, "EXPEDITION")],
                     data$EXPEDITION[data$SESSION == 2]), n),
```

```
      setNames(cbind("S1S2", "BABY",
                     data[data$SESSION == 1, c(vars, "BABY")],
                     data$BABY[data$SESSION == 2]), n),
      setNames(cbind("S2S3", "SHARK",
                     data[data$SESSION == 2, c(vars, "SHARK")],
                     data$SHARK[data$SESSION == 3]), n),
      setNames(cbind("S2S3", "GAS",
                     data[data$SESSION == 2, c(vars, "GAS")],
                     data$GAS[data$SESSION == 3]), n),
      setNames(cbind("S2S3", "FOOTBRIDGE",
                     data[data$SESSION == 2, c(vars, "FOOTBRIDGE")],
                     data$FOOTBRIDGE[data$SESSION == 3]), n),
      setNames(cbind("S2S3", "TRANSPLANT",
                     data[data$SESSION == 2, c(vars, "TRANSPLANT")],
                     data$TRANSPLANT[data$SESSION == 3]), n),
      setNames(cbind("S2S3", "EXPEDITION",
                     data[data$SESSION == 2, c(vars, "EXPEDITION")],
                     data$EXPEDITION[data$SESSION == 3]), n),
      setNames(cbind("S2S3", "BABY",
                     data[data$SESSION == 2, c(vars, "BABY")],
                     data$BABY[data$SESSION == 3]), n))
df.full$WAVE <- as.factor(df.full$WAVE)

# Create SHIFT (rating shifts) and REVERSAL (rating reversals)
df.full$SHIFT <- df.full$SECOND_SESSION - df.full$FIRST_SESSION
df.full$REVERSAL <- ifelse(df.full$FIRST_SESSION < 4
                           & df.full$SECOND_SESSION > 4, 1,
       ifelse(df.full$FIRST_SESSION > 4
              & df.full$SECOND_SESSION < 4, -1, 0))


## Data by Hatemi et al. (2019)
# Read data from file
data.hatemi <- read.csv("hatemiEtAl2019.csv",
                 header = TRUE, stringsAsFactors = FALSE)

# Create SHIFT (rating shifts) and REVERSAL (rating reversals)
data.hatemi$SHIFT <- data.hatemi$RATING2 - data.hatemi$RATING
data.hatemi$REVERSAL <- ifelse(data.hatemi$RATING < 3 &
                                 data.hatemi$RATING2 >= 3, 1,
                  ifelse(data.hatemi$RATING >= 3 &
                           data.hatemi$RATING2 < 3, -1, 0))

# Define N
N.hatemi <- nrow(data.hatemi)/2/3/5


### Test-retest correlations

## Our data
df.12 <- df.full[df.full$WAVE == "S1S2", ]
df.23 <- df.full[df.full$WAVE == "S2S3", ]

# Calculate correlation coefficients by scenario
r.shark.12 <- cor.test(df.12$FIRST_SESSION[df.12$SCENARIO == "SHARK"],
                df.12$SECOND_SESSION[df.12$SCENARIO == "SHARK"])
r.shark.23 <- cor.test(df.23$SECOND_SESSION[df.23$SCENARIO == "SHARK"],
                df.23$FIRST_SESSION[df.23$SCENARIO == "SHARK"])
r.gas.12 <- cor.test(df.12$FIRST_SESSION[df.12$SCENARIO == "GAS"],
              df.12$SECOND_SESSION[df.12$SCENARIO == "GAS"])
r.gas.23 <- cor.test(df.23$FIRST_SESSION[df.23$SCENARIO == "GAS"],
              df.23$SECOND_SESSION[df.23$SCENARIO == "GAS"])
r.foot.12 <- cor.test(df.12$FIRST_SESSION[df.12$SCENARIO == "FOOTBRIDGE"],
```

```
                df.12$SECOND_SESSION[df.12$SCENARIO == "FOOTBRIDGE"])
r.foot.23 <- cor.test(df.23$FIRST_SESSION[df.23$SCENARIO == "FOOTBRIDGE"],
                df.23$SECOND_SESSION[df.23$SCENARIO == "FOOTBRIDGE"])
r.transpl.12 <- cor.test(df.12$FIRST_SESSION[df.12$SCENARIO == "TRANSPLANT"],
                  df.12$SECOND_SESSION[df.12$SCENARIO == "TRANSPLANT"])
r.transpl.23 <- cor.test(df.23$FIRST_SESSION[df.23$SCENARIO == "TRANSPLANT"],
                  df.23$SECOND_SESSION[df.23$SCENARIO == "TRANSPLANT"])
r.exped.12 <- cor.test(df.12$FIRST_SESSION[df.12$SCENARIO == "EXPEDITION"],
                 df.12$SECOND_SESSION[df.12$SCENARIO == "EXPEDITION"])
r.exped.23 <- cor.test(df.23$FIRST_SESSION[df.23$SCENARIO == "EXPEDITION"],
                 df.23$SECOND_SESSION[df.23$SCENARIO == "EXPEDITION"])
r.baby.12 <- cor.test(df.12$FIRST_SESSION[df.12$SCENARIO == "BABY"],
                df.12$SECOND_SESSION[df.12$SCENARIO == "BABY"])
r.baby.23 <- cor.test(df.23$FIRST_SESSION[df.23$SCENARIO == "BABY"],
                df.23$SECOND_SESSION[df.23$SCENARIO == "BABY"])
rs.12 <- c(r.shark.12$estimate, r.gas.12$estimate, r.foot.12$estimate,
    r.transpl.12$estimate, r.exped.12$estimate, r.baby.12$estimate)
rs.23 <- c(r.shark.23$estimate, r.gas.23$estimate, r.foot.23$estimate,
    r.transpl.23$estimate, r.exped.23$estimate, r.baby.23$estimate)

# Fisher z transform correlation coefficients, average, re-transform
zs.12 <- atanh(rs.12); zs.23 <- atanh(rs.23)
z.12 <- mean(zs.12); z.23 <- mean(zs.23)
z.lower.12 <- z.12 - qnorm(0.975)*sqrt(1/(N-3))
z.lower.23 <- z.23 - qnorm(0.975)*sqrt(1/(N-3))
z.upper.12 <- z.12 + qnorm(0.975)*sqrt(1/(N-3))
z.upper.23 <- z.23 + qnorm(0.975)*sqrt(1/(N-3))
r.12 <- tanh(z.12); r.23 <- tanh(z.23)
r.lower.12 <- tanh(z.lower.12); r.upper.12 <- tanh(z.upper.12)
r.lower.23 <- tanh(z.lower.23); r.upper.23 <- tanh(z.upper.23)

# Generate Table 3
table3 <-
  data.frame("Scenario" = rep(c("Shark", "Gas", "Footbridge", "Transplant",
                     "Expedition", "Baby", "Total"), 2),
  "Delay" = rep(c("W1/W2", "W2/W3"), each = 7),
  "Estimate" = c(rs.12, r.12, rs.23, r.23),
  "95%-CI (lower)" = c(r.shark.12$conf.int[1],
  r.gas.12$conf.int[1], r.foot.12$conf.int[1],
  r.transpl.12$conf.int[1], r.exped.12$conf.int[1],
  r.baby.12$conf.int[1], r.lower.12, r.shark.23$conf.int[1],
  r.gas.23$conf.int[1], r.foot.23$conf.int[1],
  r.transpl.23$conf.int[1], r.exped.23$conf.int[1],
  r.baby.23$conf.int[1], r.lower.23),
 "95%-CI (upper)" = c(r.shark.12$conf.int[2],
  r.gas.12$conf.int[2], r.foot.12$conf.int[2],
  r.transpl.12$conf.int[2], r.exped.12$conf.int[2],
  r.baby.12$conf.int[2], r.upper.12, r.shark.23$conf.int[2],
  r.gas.23$conf.int[2], r.foot.23$conf.int[2],
  r.transpl.23$conf.int[2], r.exped.23$conf.int[2],
  r.baby.23$conf.int[2], r.upper.23))


## Hatemi et al's data
df.12 <- data.hatemi[data.hatemi$WAVE == "12", ]
df.23 <- data.hatemi[data.hatemi$WAVE == "23", ]

rown <- c("COMPASSION", "ANIMAL", "KILL", "FAIRLY", "JUSTICE",
           "RICH", "HISTORY","FAMILY","TEAM","RESPECT",
           "SEXROLES","SOLDIER","HARMLESSDG","UNNATURAL", "CHASTITY")

# Calculate correlation coefficients by item
```

```r
cor.df.12 <- data.frame("Item" = rown, "Estimate" = 0,
                        "95%-CI (lower)" = 0, "95%-CI (upper)" = 0)
cor.df.23 <- data.frame("Item" = rown, "Estimate" = 0,
                          "95%-CI (lower)" = 0, "95%-CI (upper)" = 0)
for(j in 1:length(rown)){
  c.12 <- cor.test(df.12$RATING[df.12$ITEM == rown[j]],
                   df.12$RATING2[df.12$ITEM == rown[j]])
  cor.df.12$Estimate[j] <- c.12$estimate
  cor.df.12$X95..CI..lower.[j] <- c.12$conf.int[1]
  cor.df.12$X95..CI..upper.[j] <- c.12$conf.int[2]
  c.23 <- cor.test(df.23$RATING[df.23$ITEM == rown[j]],
                   df.23$RATING2[df.23$ITEM == rown[j]])
  cor.df.23$Estimate[j] <- c.23$estimate
  cor.df.23$X95..CI..lower.[j] <- c.23$conf.int[1]
  cor.df.23$X95..CI..upper.[j] <- c.23$conf.int[2]
}

# Generate Table 1
table1 <- data.frame("Foundation" = c("Harm", "Fairness", "Ingroup",
                       "Authority", "Purity"), "Estimate" = 0,
                       "95%-CI (lower)" = 0, "95%-CI (upper)" = 0,
                       "Estimate" = 0, "95%-CI (lower)" = 0,
                       "95%-CI (upper)" = 0)

# Fisher z transform correlation coefficients, average, re-transform
for(j in 1:5){
  index <- ((3*(j-1))+1):((3*(j-1))+3)

  rs.12 <- cor.df.12$Estimate[index]
  rs.23 <- cor.df.23$Estimate[index]

  zs.12 <- atanh(rs.12); zs.23 <- atanh(rs.23)
  z.12 <- mean(zs.12); z.23 <- mean(zs.23)
  z.lower.12 <- z.12 - qnorm(0.975)*sqrt(1/(N.hatemi-3))
  z.lower.23 <- z.23 - qnorm(0.975)*sqrt(1/(N.hatemi-3))
  z.upper.12 <- z.12 + qnorm(0.975)*sqrt(1/(N.hatemi-3))
  z.upper.23 <- z.23 + qnorm(0.975)*sqrt(1/(N.hatemi-3))
  table1$Estimate[j] <- tanh(z.12)
  table1$Estimate.1[j] <- tanh(z.23)
  table1$X95..CI..lower.[j] <- tanh(z.lower.12)
  table1$X95..CI..lower..1[j] <- tanh(z.lower.23)
  table1$X95..CI..upper.[j] <- tanh(z.upper.12)
  table1$X95..CI..upper..1[j] <- tanh(z.upper.23)
}


### Rating shifts and rating reversals

## Our data
df.12 <- df.full[df.full$WAVE == "S1S2", ]
df.23 <- df.full[df.full$WAVE == "S2S3", ]

# Calculate proportions of rating shifts by scenario
shift.12 <- setNames(do.call(data.frame,
        aggregate(SHIFT ~ SCENARIO, data = df.12,
        FUN = function(x) binconf(sum(x != 0), n = N,
        alpha = 0.05, method = "wilson"))),
        c("Scenario", "Shifts", "95%-CI (lower)", "95%-CI (upper)"))
shift.23 <- setNames(do.call(data.frame,
        aggregate(SHIFT ~ SCENARIO, data = df.23,
        FUN = function(x) binconf(sum(x != 0), n = N,
        alpha = 0.05, method = "wilson"))),
        c("Scenario", "Shifts", "95%-CI (lower)", "95%-CI (upper)"))
```

```
shift.12$Scenario <- c("Shark", "Gas", "Footbridge",
                       "Transplant", "Expedition", "Baby")
shift.23$Scenario <- c("Shark", "Gas", "Footbridge",
                       "Transplant", "Expedition", "Baby")

# Calculate proportions of rating shifts overall
shift.12 <- rbind(shift.12, c("Overall",
                    mean(shift.12$Shifts), "--", "--"))
shift.23 <- rbind(shift.23, c("Overall",
                    mean(shift.23$Shifts), "--", "--"))

# Calculate proportions of rating reversals by scenario
rev.12 <- setNames(do.call(data.frame, aggregate(REVERSAL ~ SCENARIO,
      data = df.12, FUN = function(x) binconf(sum(x != 0), n = N,
      alpha = 0.05, method = "wilson"))),
      c("Scenario", "Reversals", "95%-CI (lower)", "95%-CI (upper)"))
rev.23 <- setNames(do.call(data.frame, aggregate(REVERSAL ~ SCENARIO,
      data = df.23, FUN = function(x) binconf(sum(x != 0), n = N,
      alpha = 0.05, method = "wilson"))),
      c("Scenario", "Reversals", "95%-CI (lower)", "95%-CI (upper)"))
rev.12$Scenario <- NULL; rev.23$Scenario <- NULL

# Calculate proportions of rating reversals overall
rev.12 <- rbind(rev.12, c(mean(rev.12$Reversals), "--", "--"))
rev.23 <- rbind(rev.23, c(mean(rev.23$Reversals), "--", "--"))

# Create Table 5
table5 <- rbind(cbind(shift.12, rev.12),
                cbind(shift.23, rev.23))
table5$Wave <- rep(c("W1/W2", "W2/W3"), each = 7)



## Hatemi et al.'s data
df.12 <- data.hatemi[data.hatemi$WAVE == "12", ]
df.23 <- data.hatemi[data.hatemi$WAVE == "23", ]

#  Calculate proportions of rating shifts by item
shift.12 <- setNames(do.call(data.frame,
                  aggregate(SHIFT ~ ITEM, data = df.12,
                  FUN = function(x) binconf(sum(x != 0), n = N.hatemi,
                  alpha = 0.05, method = "wilson"))),
                  c("Item", "Shifts", "95%-CI (lower)", "95%-CI (upper)"))
shift.23 <- setNames(do.call(data.frame,
                aggregate(SHIFT ~ ITEM, data = df.23,
                FUN = function(x) binconf(sum(x != 0), n = N.hatemi,
                alpha = 0.05, method = "wilson"))),
                c("Item", "Shifts", "95%-CI (lower)", "95%-CI (upper)"))
shift.12$Item <- rown; shift.23$Item <- rown

# Calculate proportions of rating reversals by item
rev.12 <- setNames(do.call(data.frame, aggregate(REVERSAL ~ ITEM,
              data = df.12, FUN = function(x) binconf(sum(x != 0),n = N.hatemi,
              alpha = 0.05, method = "wilson"))),
              c("Item", "Reversals", "95%-CI (lower)", "95%-CI (upper)"))
rev.23 <- setNames(do.call(data.frame, aggregate(REVERSAL ~ ITEM,
              data = df.23, FUN = function(x) binconf(sum(x != 0), n = N.hatemi,
              alpha = 0.05, method = "wilson"))),
             c("Item", "Reversals", "95%-CI (lower)", "95%-CI (upper)"))
rev.12$Item <- NULL; rev.23$Item <- NULL

# Generate Table 4
```

```
table4 <- rbind(data.frame(
  "Item" = c("Harm", "Fairness", "Ingroup", "Authority", "Purity", "Overall"),
   "Waves" = "1st/2nd",
   "Shifts" = c(mean(shift.12$Shifts[1:3]), mean(shift.12$Shifts[4:6]),
                mean(shift.12$Shifts[7:9]), mean(shift.12$Shifts[10:12]),
                mean(shift.12$Shifts[13:15]), mean(shift.12$Shifts)),
   "Reversals" = c(mean(rev.12$Reversals[1:3]), mean(rev.12$Reversals[4:6]),
                   mean(rev.12$Reversals[7:9]), mean(rev.12$Reversals[10:12]),
                   mean(rev.12$Reversals[13:15]), mean(rev.12$Reversals))),
data.frame("Item" = c("Harm", "Fairness", "Ingroup", "Authority", "Purity",
"Overall"),
   "Waves" = "2nd/3rd",
   "Shifts" = c(mean(shift.23$Shifts[1:3]), mean(shift.23$Shifts[4:6]),
                mean(shift.23$Shifts[7:9]), mean(shift.23$Shifts[10:12]),
                mean(shift.23$Shifts[13:15]), mean(shift.23$Shifts)),
   "Reversals" = c(mean(rev.23$Reversals[1:3]), mean(rev.23$Reversals[4:6]),
                   mean(rev.23$Reversals[7:9]), mean(rev.23$Reversals[10:12]),
                   mean(rev.23$Reversals[13:15]), mean(rev.23$Reversals))))


### Reasons hypothesis

## Our data
# Get data into wide format
df.wide <- cbind(df.full[df.full$WAVE == "S1S2", ],
                 "SHIFT2" = df.full$SHIFT[df.full$WAVE == "S2S3"],
                 "REVERSAL2" = df.full$REVERSAL[df.full$WAVE == "S2S3"])

# Look at the persistence of changes
# Run linear mixed-effect model
m <- lmer(SHIFT2 ~ SHIFT + (1|SCENARIO) + (1|ID), data = df.wide)
s <- summary(m)
b <- s$coefficients["SHIFT", 1]

# Calculate confidence interval
cint <- confint(m)
CI95 <- c(cint["SHIFT", 1], cint["SHIFT", 2])

# Run binomial mixed-effects model with logit link
m <- glmer(abs(REVERSAL2) ~ abs(REVERSAL) + (1|SCENARIO) + (1|ID),
           data = df.wide, family = binomial(link = "logit"))
s <- summary(m)
OR <- exp(s$coefficients["abs(REVERSAL)", 1])

# Calculate confidence interval
cint <- confint(m)
CI95 <- exp(c(cint["abs(REVERSAL)", 1], cint["abs(REVERSAL)", 2]))

# Look at whether AOT, REI, and CRT predict shifts/reversals
# Run linear mixed-effect model
m <- lmer(abs(SHIFT) ~ WAVE + AOT + NFC + FI + CRT +
             (1|SCENARIO), data = df.full)
s <- summary(m)
b.aot <- s$coefficients["AOT", 1]
b.nfc <- s$coefficients["NFC", 1]
b.fi <- s$coefficients["FI", 1]
b.crt <- s$coefficients["CRT", 1]

# Calculate confidence intervals
cint <- confint(m)
CI95.aot <- c(cint["AOT", 1], cint["AOT", 2])
CI95.nfc <- c(cint["NFC", 1], cint["NFC", 2])
CI95.fi <- c(cint["FI", 1], cint["FI", 2])
```

```r
CI95.crt <- c(cint["CRT", 1], cint["CRT", 2])

# Run binomial mixed-effects model with logit link
m <- glmer(abs(REVERSAL) ~ WAVE + AOT + NFC + FI + CRT +
             (1|SCENARIO), data = df.full,
           family = binomial(link = "logit"))
s <- summary(m)
OR.aot <- exp(s$coefficients["AOT", 1])
OR.nfc <- exp(s$coefficients["NFC", 1])
OR.fi <- exp(s$coefficients["FI", 1])
OR.crt <- exp(s$coefficients["CRT", 1])

# Calculate confidence intervals
cint <- confint(m)
CI95.aot <- exp(c(cint["AOT", 1], cint["AOT", 2]))
CI95.nfc <- exp(c(cint["NFC", 1], cint["NFC", 2]))
CI95.fi <- exp(c(cint["FI", 1], cint["FI", 2]))
CI95.crt <- exp(c(cint["CRT", 1], cint["CRT", 2]))


## Hatemi et al.'s data
# Get data into wide format
df.wide <- cbind(data.hatemi[data.hatemi$WAVE == 12, ],
       "SHIFT2" = data.hatemi$SHIFT[data.hatemi$WAVE == 23],
       "REVERSAL2" = data.hatemi$REVERSAL[data.hatemi$WAVE == 23])

# Run linear mixed-effect model
m <- lmer(SHIFT2 ~ SHIFT + (1|ITEM) + (1|ID), data = df.wide)
s <- summary(m)
b <- s$coefficients["SHIFT", 1]

# Calculate confidence interval
cint <- confint(m)
CI95 <- c(cint["SHIFT", 1], cint["SHIFT", 2])

# Run binomial mixed-effects model with logit link
m <- glmer(abs(REVERSAL2) ~ abs(REVERSAL) + (1|ID),
           data = df.wide, family = binomial(link = "logit"))
s <- summary(m)
OR <- exp(s$coefficients["abs(REVERSAL)", 1])

# Calculate confidence interval
cint <- confint(m)
CI95 <- exp(c(cint["abs(REVERSAL)", 1], cint["abs(REVERSAL)", 2]))


## Figure 1
df <- data.frame("R" = c(c(0.71, 0.68, 0.69, 0.71, 0.82,
                           0.69, 0.71, 0.72, 0.68, 0.73,
                           0.51, 0.46, 0.62, 0.59, 0.75),
                         c(0.87, 0.74, 0.71, 0.78, 0.71, 0.66, 0.71),
                         0.67, c(0.81, 0.73, 0.76, 0.81, 0.83), 0.74),
       "WAVE" = rep(c("1st wave/2nd wave", "2nd wave/3rd wave"),
                    c(5*3+7+1, 5+1)),
       "WHAT" =  rep(c("MFQ-30", "MAC-Q", "Ours", "MFQ-30", "Ours"),
                     c(3*5, 7, 1, 5, 1)))

ggplot(df, aes(x = R, fill = WHAT, col = WHAT)) + geom_dotplot() +
  theme_classic() + theme(line = element_blank(),
              axis.text.y = element_blank(),
              legend.title = element_blank(),
              strip.background = element_blank(),
              axis.text.x = element_text(size = 9, color = "black")) +
```

```
  scale_fill_manual(values = c("darkgray", "black", "red")) +
  scale_color_manual(values = c("darkgray", "black", "red")) +
  geom_hline(yintercept = 0.03, colour = "black", size = 0.4) +
  labs(x = "r", y = "") + facet_grid(WAVE ~ .)


## Figure 2
df <- data.frame("PROP" = c(c(0.44, 0.46, 0.48, 0.49, 0.51),
                            c(0.13, 0.16, 0.13, 0.17, 0.20),
                            c(0.48, 0.14),
                            c(0.34, 0.40, 0.43, 0.38, 0.45),
                            c(0.10, 0.14, 0.14, 0.11, 0.15),
                            c(0.43, 0.12)),
 "WAVE" = rep(c("1st wave/2nd wave", "2nd wave/3rd wave"),
              c(5*2+2, 5*2+2)),
 "MEASURE" = rep(rep(c("MFQ", "Ours"),2), c(2*5, 2, 2*5, 2)),
 "WHAT" = rep(rep(rep(c("Shift", "Reversal"),2), c(5,5,1,1)), 2))

ggplot(df, aes(x = PROP, fill = MEASURE, col = MEASURE)) + geom_dotplot() +
  theme_classic() + theme(line = element_blank(),
                axis.text.y = element_blank(),
                legend.title = element_blank(),
                strip.background = element_blank(),
                axis.text.x = element_text(size = 9, color = "black")) +
  scale_fill_manual(values = c("black", "red")) +
  scale_color_manual(values = c("black", "red")) +
  geom_hline(yintercept = 0.055, colour = "black", size = 0.4) +
  labs(x = "Proportion", y = "") + facet_grid(WAVE + WHAT ~ .)
```

References

Aczel, Balazs, Aba Szollosi, and Bence Bago. 2018. “The Effect of Transparency on Framing Effects in Within-Subject Designs.” *Journal of Behavioral Decision Making* 31 (1): 25–39.

Agresti, Alan, and Brent A. Coull. 1998. “Approximate Is Better than ‘Exact’ for Interval Estimation of Binomial Proportions.” *The American Statistician* 52 (2): 119–26.

Alexander, Joshua, Ronald Mallon, and Jonathan M. Weinberg. (2010) 2014. “Accentuate the Negative.” In *Experimental Philosophy*, edited by Joshua Knobe and Shaun Nichols, 2:31–50. Oxford University Press.

Andow, James. 2016. “Reliable but Not Home Free? What Framing Effects Mean for Moral Intuitions.” *Philosophical Psychology* 29 (6): 904–11.

Ansani, Alessandro, Francesca D’Errico, and Isabella Poggi. 2017. “‘It Sounds Wrong…’ Does Music Affect Moral Judgement?” In *Computational Science and Its Applications – ICCSA 2017*, edited by Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Giuseppe Borruso, Carmelo M. Torre, Ana Maria A.C. Rocha, David Taniar, Bernady O. Apduhan, Elena Stankova, and Alfredo Cuzzocrea, 10409:753–60. Springer International Publishing.

Audi, Robert. 2008. “Intuition, Inference, and Rational Disagreement in Ethics.” *Ethical Theory and Moral Practice* 11 (5): 475–92.

Baayen, Harald, Doug Davidson, and Douglas Bates. 2008. “Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items.” *Journal of Memory and Language* 59 (4): 390–412.

Baron, Jonathan. 1993. “Why Teach Thinking?--An Essay.” *Applied Psychology* 42 (3): 191–214.

———. 2019. “Actively Open-Minded Thinking in Politics.” *Cognition* 188 (July): 8–18.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using Lme4.” *Journal of Statistical Software* 67 (1).

Bauman, Christopher W., A. Peter McGraw, Daniel M. Bartels, and Caleb Warren. 2014. “Revisiting External Validity: Concerns about Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology: External Validity in Moral Psychology.” *Social and Personality Psychology Compass* 8 (9): 536–54.

Beebe, James R., and David Sackris. 2016. “Moral Objectivism across the Lifespan.” *Philosophical Psychology* 29 (6): 912–29.

Bloom, Paul. 2010. “How Do Morals Change?” *Nature* 464 (7288): 490–490.

Buchanan, Erin M., and John E. Scofield. 2018. “Methods to Detect Low Quality Data and Its Implication for Psychological Research.” *Behavior Research Methods* 50 (6): 2586–96.

Campbell, Richmond. 2007. “What Is Moral Judgment?” *Journal of Philosophy* 104 (7): 321–49.

———. 2017. “Learning from Moral Inconsistency.” *Cognition* 167 (October): 46–57.

Campbell, Richmond, and Victor Kumar. 2012. “Moral Reasoning on the Ground.” *Ethics* 122 (2): 273–312.

Chituc, Vladimir, and Walter Sinnott-Armstrong. 2020. “Moral Conformity and Its Philosophical Lessons.” *Philosophical Psychology* 33 (2): 262–82.

Christensen, Julia F., Albert Flexas, Margareta Calabrese, Nadine K. Gut, and Antoni Gomila. 2014. “Moral Judgment Reloaded: A Moral Dilemma Validation Study.” *Frontiers in Psychology* 5 (July).

Christensen, Julia F., and Antoni Gomila. 2012. “Moral Dilemmas in Cognitive Neuroscience of Moral Decision-Making: A Principled Review.” *Neuroscience & Biobehavioral Reviews* 36 (4): 1249–64.

Cullity, Garrett. 2016. “Moral Judgement.” In *Routledge Encyclopedia of Philosophy*. Routledge.

Curry, Oliver Scott. 2016. “Morality as Cooperation: A Problem-Centred Approach.” In *The Evolution of Morality*, edited by Todd K. Shackelford and Ranald D. Hansen, 27–51. Evolutionary Psychology. Springer International Publishing.

Curry, Oliver Scott, Matthew Jones Chesters, and Caspar J. Van Lissa. 2019. “Mapping Morality with a Compass: Testing the Theory of ‘Morality-as-Cooperation’ with a New Questionnaire.” *Journal of Research in Personality* 78 (February): 106–24.

Daniels, Norman. 2020. “Reflective Equilibrium.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab.

Darwall, Stephen. 1998. *Philosophical Ethics: An Historical And Contemporary Introduction*. Routledge.

Demaree-Cotton, Joanna. 2016. “Do Framing Effects Make Moral Intuitions Unreliable?” *Philosophical Psychology* 29 (1): 1–22.

Deutsch, Max. 2010. “Intuitions, Counter-Examples, and Experimental Philosophy.” *Review of Philosophy and Psychology* 1 (3): 447–60.

Duke, Aaron A., and Laurent Bègue. 2015. “The Drunk Utilitarian: Blood Alcohol Concentration Predicts Utilitarian Responses in Moral Dilemmas.” *Cognition* 134 (January): 121–27.

Epstein, Seymour, Rosemary Pacini, Veronika Denes-Raj, and Harriet Heier. 1996. “Individual Differences in Intuitive-Experiential and Analytical-Rational Thinking Styles.” *Journal of Personality and Social Psychology* 71 (2): 390–405.

Foot, Philippa. 1967. “The Problem of Abortion and the Doctrine of Double Effect.” *Oxford Review* 5: 5–15.

Frederick, Shane. 2005. “Cognitive Reflection and Decision Making.” *Journal of Economic Perspectives* 19 (4): 25–42.

Giner-Sorolla, Roger. 2020. “Decision on Submission to Journal of Experimental Social Psychology - Reject without Review,” October 21, 2020.

Goodwin, Geoffrey P., and John M. Darley. 2008. “The Psychology of Meta-Ethics: Exploring Objectivism.” *Cognition* 106 (3): 1339–66.

Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. 2011. “Mapping the Moral Domain.” *Journal of Personality and Social Psychology* 101 (2): 366–85.

Greene, Joshua D., Leigh E. Nystrom, Andrew D. Engell, John M. Darley, and Jonathan D. Cohen. 2004. “The Neural Bases of Cognitive Conflict and Control in Moral Judgment.” *Neuron* 44 (2): 389–400.

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. “An FMRI Investigation of Emotional Engagement in Moral Judgment.” *Science* 293 (5537): 2105–8.

Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage.

Haidt, Jonathan, and Jesse Graham. 2007. “When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize.” *Social Justice Research* 20 (1): 98–116.

Haran, Uriel, Ilana Ritov, and Barbara A Mellers. 2013. “The Role of Actively Open-Minded Thinking in Information Acquisition, Accuracy, and Calibration.” *Judgment and Decision Making* 8 (3): 14.

Hatemi, Peter K., Charles Crabtree, and Kevin B. Smith. 2019. “Ideology Justifies Morality: Political Beliefs Predict Moral Foundations.” *American Journal of Political Science* 63 (4): 788–806.

Helzer, Erik G., and David A. Pizarro. 2011. “Dirty Liberals!: Reminders of Physical Cleanliness Influence Moral and Political Attitudes.” *Psychological Science* 22 (4): 517–22.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The Weirdest People in the World?” *Behavioral and Brain Sciences* 33 (2–3): 61–83.

Horne, Zachary, Derek Powell, and John Hummel. 2015. “A Single Counterexample Leads to Moral Belief Revision.” *Cognitive Science* 39 (8): 1950–64.

Huang, Jason L., Paul G. Curran, Jessica Keeney, Elizabeth M. Poposki, and Richard P. DeShon. 2012. “Detecting and Deterring Insufficient Effort Responding to Surveys.” *Journal of Business and Psychology* 27 (1): 99–114.

Huemer, Michael. 2005. *Ethical Intuitionism*. Palgrave Macmillan.

———. 2009. “Précis of ‘Ethical Intuitionism.’” *Philosophy and Phenomenological Research* 78 (1): 192–96.

Kahane, Guy, Katja Wiech, Nicholas Shackel, Miguel Farias, Julian Savulescu, and Irene Tracey. 2012. “The Neural Basis of Intuitive and Counterintuitive Moral Judgment.” *Social Cognitive and Affective Neuroscience* 7 (4): 393–402.

Kamm, Frances M. 1993. *Morality, Mortality: Death and Whom to Save from It*. Vol. 1. Oxford University Press.

Killgore, William D.S., Desiree B. Killgore, Lisa M. Day, Christopher Li, Gary H. Kamimori, and Thomas J. Balkin. 2007. “The Effects of 53 Hours of Sleep Deprivation on Moral Judgment.” *Sleep* 30 (3): 345–52.

Kreps, Tamar A., Kristin Laurin, and Anna C. Merritt. 2017. “Hypocritical Flip-Flop, or Courageous Evolution? When Leaders Change Their Moral Minds.” *Journal of Personality and Social Psychology* 113 (5): 730–52.

Kreps, Tamar A., and Benoît Monin. 2014. “Core Values Versus Common Sense: Consequentialist Views Appear Less Rooted in Morality.” *Personality and Social Psychology Bulletin* 40 (11): 1529–42.

Li, Hongxia, Xue Wang, Yafei Guo, Zhansheng Chen, and Fei Teng. 2019. “Air Pollution Predicts Harsh Moral Judgment.” *International Journal of Environmental Research and Public Health* 16 (13): 2276.

Luttrell, Andrew, Richard E. Petty, Pablo Briñol, and Benjamin C. Wagner. 2016.

"Making It Moral: Merely Labeling an Attitude as Moral Increases Its Strength." *Journal of Experimental Social Psychology* 65 (July): 82–93.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. Penguin Books.

May, Joshua. 2019. "Précis of 'Regard for Reason in the Moral Mind.'" *Behavioral and Brain Sciences* 42 (e146): 1–60.

McConnell, Terrance. 2018. "Moral Dilemmas." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab.

McIntyre, Alison. 2019. "Doctrine of Double Effect." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019. Metaphysics Research Lab, Stanford University.

Mudrack, Peter E., and E. Sharon Mason. 2013. "Dilemmas, Conspiracies, and Sophie's Choice: Vignette Themes and Ethical Judgments." *Journal of Business Ethics* 118 (3): 639–53.

Nadelhoffer, Thomas, and Adam Feltz. 2008. "The Actor–Observer Bias and Moral Intuitions: Adding Fuel to Sinnott-Armstrong's Fire." *Neuroethics* 1 (2): 133–44.

Nakamura, Hiroko, Yuichi Ito, Yoshiko Honma, Takuya Mori, and Jun Kawaguchi. 2014. "Cold-Hearted or Cool-Headed: Physical Coldness Promotes Utilitarian Moral Judgment." *Frontiers in Psychology* 5 (October).

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko. 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45 (4): 867–72.

Paulo, Norbert. 2020. "The Unreliable Intuitions Objection Against Reflective Equilibrium." *The Journal of Ethics* 24 (3): 333–53.

Paxton, Joseph M., Leo Ungar, and Joshua D. Greene. 2012. "Reflection and Reasoning in Moral Judgment." *Cognitive Science* 36 (1): 163–77.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research." *Journal of Experimental Social Psychology* 70 (May): 153–63.

Pennycook, Gordon, James Allan Cheyne, Derek J. Koehler, and Jonathan Albert Fugelsang. 2019. "On the Belief That Beliefs Should Change According to Evidence: Implications for Conspiratorial, Moral, Paranormal, Political, Religious, and Science Beliefs." Preprint. PsyArXiv.

Pizarro, David A., and Paul Bloom. 2003. "The Intelligence of the Moral Intuitions: A

Comment on Haidt (2001).” *Psychological Review* 110 (1): 193–96.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rawls, John. 1971. *A Theory of Justice*. Harvard University Press.

Rehren, Paul, and Walter Sinnott-Armstrong. Under review. “Moral Framing Effects Within Subject.”

Ross, David. (1930) 2002. *The Right and the Good*. Edited by Philip Stratton-Lake. Clarendon Press.

Sá, Walter C., Carol N. Kelley, Caroline Ho, and Keith E. Stanovich. 2005. “Thinking about Personal Theories: Individual Differences in the Coordination of Theory and Evidence.” *Personality and Individual Differences* 38 (5): 1149–61.

Sauer, Hanno. 2012. “Educated Intuitions. Automaticity and Rationality in Moral Judgement.” *Philosophical Explorations* 15 (3): 255–75.

———. 2018. *Debunking Arguments in Ethics*. Cambridge University Press.

Seidel, Angelika, and Jesse Prinz. 2013. “Sound Morality: Irritating and Icky Noises Amplify Judgments in Divergent Moral Domains.” *Cognition* 127 (1): 1–5.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence*. Oxford University Press.

Silver, N. Clayton, and William P. Dunlap. 1987. “Averaging Correlation Coefficients: Should Fisher’s z Transformation Be Used?” *Journal of Applied Psychology* 72 (1): 146–48.

Sinnott-Armstrong, Walter. 2008. “Framing Moral Intuitions.” In *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity*, edited by Walter Sinnott-Armstrong, 2:47–76. MIT Press.

Skitka, Linda J., Anthony N. Washburn, and Timothy S. Carsel. 2015. “The Psychological Foundations and Consequences of Moral Conviction.” *Current Opinion in Psychology* 6 (December): 41–44.

Smith, Michael. 1994. *The Moral Problem*. Wiley-Blackwell.

Stanovich, Keith E., and Richard F. West. 1997. “Reasoning Independently of Prior Belief and Individual Differences in Actively Open-Minded Thinking.” *Journal of Educational Psychology* 89 (2): 342–57.

Toplak, Maggie E., Richard F. West, and Keith E. Stanovich. 2014. “Assessing Miserly Information Processing: An Expansion of the Cognitive Reflection

Test.” *Thinking & Reasoning* 20 (2): 147–68.

Tropman, Elizabeth. 2011. “Non-Inferential Moral Knowledge.” *Acta Analytica* 26 (4): 355–66.

Turiel, Elliot. 1983. *The Development of Social Knowledge. Morality and Convention*. Cambridge University Press.

Vallejo, Adriana, Ana Muniesa, Chelo Ferreira, and Ignacio de Blas. 2013. “New Method to Estimate the Sample Size for Calculation of a Proportion Assuming Binomial Distribution.” *Research in Veterinary Science* 95 (2): 405–9.

Waldmann, Michael R., Jonas Nagel, and Alex Wiegmann. 2012. “Moral Judgment.” In *The Oxford Handbook of Thinking and Reasoning*, edited by Keith J. Holyoak and Robert G. Morrison. Oxford University Press.

Williamson, Timothy. 2007. “Knowledge of Metaphysical Modality.” In *The Philosophy of Philosophy*, 134–78. The Blackwell/Brown Lectures in Philosophy 2. Blackwell Publishing.

Woollard, Fiona, and Frances Howard-Snyder. 2016. “Doing vs. Allowing Harm.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab.