Cognition 212 (2021) 104703

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Valence framing effects on moral judgments: A meta-analysis

Kelsey McDonald[a,b,c], Rose Graves[d,1], Siyuan Yin[e], Tara Weese[f,g], Walter Sinnott-Armstrong[a,b,c,f,g,h,*]

[a] Duke Institute for Brain Sciences, Duke University, Durham 27710, NC, USA
[b] Center for Cognitive Neuroscience, Duke University, Durham 27710, NC, USA
[c] Department of Psychology and Neuroscience, Duke University, Durham 27708, NC, USA
[d] Department of Statistical Science, Duke University, Durham 27708, NC, USA
[e] Department of Marketing, The Wharton School, University of Pennsylvania, Philadelphia 19104, PA, USA
[f] Department of Philosophy, Duke University, Durham 27708, NC, USA
[g] Law School, Duke University, Durham 27708, NC, USA
[h] Kenan Institute for Ethics, Duke University, Durham 27708, NC, USA

ARTICLE INFO

Keywords:
Meta-analysis
Framing effects
Moral judgment
Intuition
Decision-making

ABSTRACT

Valence framing effects occur when participants make different choices or judgments depending on whether the options are described in terms of their positive outcomes (e.g. lives saved) or their negative outcomes (e.g. lives lost). When such framing effects occur in the domain of moral judgments, they have been taken to cast doubt on the reliability of moral judgments and raise questions about the extent to which these moral judgments are self-evident or justified in themselves. One important factor in this debate is the magnitude and variability of the extent to which differences in framing presentation impact moral judgments. Although moral framing effects have been studied by psychologists, the overall strength of these effects pooled across published studies is not yet known. Here we conducted a meta-analysis of 109 published articles (contributing a total of 146 unique experiments with 49,564 participants) involving valence framing effects on moral judgments and found a moderate effect ($d = 0.50$) among between-subjects designs as well as several moderator variables. While we find evidence for publication bias, statistically accounting for publication bias attenuates, but does not eliminate, this effect ($d = 0.22$). This suggests that the magnitude of valence framing effects on moral decisions is small, yet significant when accounting for publication bias.

1. Introduction

For centuries, philosophers have argued about whether and how moral judgments can be justified (Sinnott-Armstrong, 2007). One recently popular anti-skeptical position in these debates is moral intuitionism, which claims that certain moral judgments are self-evident or justified in themselves or immediately, merely on the basis on understanding them without any need for support from inference or argument from other beliefs (Audi, 2013; Hernandez, 2011; Sinnott-Armstrong, 2001; Stratton-Lake, 2020). The moral judgments that are supposed to be justified without need for inferential support[2] are often identified as those that result from a process that is reliable in the sense that it is likely to produce a high proportion of moral judgments that are true (Shafer-Landau, 2005) or at least correct in some broad sense (Blackburn, 1996), even if the person who makes the judgment does not have any reason to believe that the process is reliable. Moral intuitionism then depends on the claim that the processes that lead to moral judgments are reliable in this sense.

Recently, several philosophers (including Sinnott-Armstrong, 2008a, 2008b) have argued that framing effects on moral judgments provide evidence against their reliability and, hence, against the claim that any moral judgments are self-evident or justified in themselves. If a process

[*] Corresponding author at: Duke Institute for Brain Sciences, Duke University, Durham 27710, NC, USA.
E-mail address: walter.sinnott-armstrong@duke.edu (W. Sinnott-Armstrong).

[1] Rose Graves is co-first author.

[2] Moral judgments that are not based on inference are sometimes called *moral intuitions*, but moral intuitionists often insist that moral intuitions are *seemings*, which are not beliefs (Stratton-Lake, 2020). The psychological studies in our meta-analysis do not distinguish seemings from judgments or beliefs, so they cannot verify that participants' responses reflect seemings. That is why we write instead about moral judgments rather than moral intuitions.

https://doi.org/10.1016/j.cognition.2021.104703
Received 10 August 2020; Received in revised form 21 March 2021; Accepted 22 March 2021
Available online 6 May 2021

yields one judgment in some frames but another judgment in other equivalent frames, and if these judgments cannot both be true or correct, then the process must have produced an incorrect judgment in at least one of the frames, even if we do not know which one. Thus, a process that yields enough contrary results in different frames that are morally equivalent cannot be reliable in the specified sense. Critics sometimes debate whether semantically-equivalent frames convey equivalent information (e.g., Aczel, Szollosi, & Bago, 2018; Frisch, 1993; Mandel, 2014; McKenzie & Nelson, 2003; Sher & McKenzie, 2006), but effects of frames that are truly equivalent still do show unreliability. The next step in the argument claims that too much unreliability creates a need for some support by inference or argument. Of course, how much unreliability is *too* much is a normative issue that can depend on what is at stake. The same degree of unreliability might be acceptable when errors would do little harm, but unacceptable when mistakes could break apart families, friends, and societies, as moral judgments sometimes do. When moral judgments are not reliable enough in this normative sense, we should not trust them without independent confirmation, and then they are not justified in themselves, according to these opponents of moral intuitionism.

This study looks at the empirical evidence concerning valence framing effects on moral judgments in order to test the claims that their formation processes are reliable, as assumed by moral intuitionists and denied by their critics. Psychological research cannot show positively that a moral belief *is* justified or that the processes that yield moral intuitions *are* reliable, because those claims are normative or moral instead of scientific. Nonetheless, psychological research can show that such a process is *unreliable* to the extent that moral judgments are affected by factors that both sides of the debate agree are morally irrelevant.

Moral framing effects provide examples of such morally irrelevant factors. *Framing effects* on moral judgments occur when an individual's moral judgment is affected (a) by the circumstances of that individual (rather than those of the agent whose act is judged) or (b) by the way in which the options are presented (rather than any feature of the actions that are chosen or judged to be morally right or wrong, which could affect whether those actions are right or wrong). The *circumstantial* kind of moral framing effect (a) occurs, for example, when bad smells or cleaning products in their environment affect people's moral judgments. In a recent meta-analysis, Landy and Goodwin (2015) found that this kind of effect is not as strong as many moral psychologists claim. In contrast, one *presentational* kind of moral framing effect (b) occurs when people make different moral judgments about the same scenarios just because the scenarios are presented in a different order. This kind of framing effect has been found in several studies, starting with Petrinovich and O'Neill (1996).

Our meta-analysis will instead focus on a different kind of presentational framing effect: valence framing effects. *Valence* framing effects occur when participants make different choices or judgments depending on whether the options are described in terms of their positive outcomes (e.g. lives saved) or their negative outcomes (e.g. lives lost). A prominent example that spawned much research on this kind of framing effect is Tversky and Kahneman's famous Disease Problem (TKDP), in which individuals are asked to express a preference between two programs that are described differently in different conditions (see box below; Tversky & Kahneman, 1981).

Program A in Condition 1 has exactly the same outcome as Program C in Condition 2, since saving 200 out of 600 means that 400 die. Similarly, Program B in Condition 1 has exactly the same outcome as Program D in Condition 2, since a 1/3 probability of saving all 600 equals a 1/3 probability that none die, and a 2/3 probability that none of the 600 are saved equals a 2/3 probability that all 600 die. The descriptions do not change the outcomes. Consequently, if there are no morally relevant differences between the programs, anyone who favors Program A should also favor Program C and anyone who favors Program B should also favor Program D. Surprisingly, Tversky and Kahneman found that 72% favored Program A but only 22% favored Program C. This finding suggests that many subjects were influenced by the framing or description of the programs in negative or positive terms, so it is a valence framing effect.

These results have been replicated often with the TKDP, but it is still not clear whether similar valence framing effects on moral judgments occur consistently with other moral scenarios. Our meta-analysis tries to answer that question by comparing studies of a wide variety of moral scenarios. A moral scenario pair is suitable to show a valence framing effect only if the options are described or framed positively in one scenario but negatively in the other scenario and yet the options in the different scenarios remain equivalent in all morally relevant respects. The results then show a valence framing effect if participants show a bias in favor of an option under one frame (positive or negative) but not under the other frame.

Many studies have individually investigated the impacts of framing effects on human decision-making; moreover, other meta-analyses have been published seeking to estimate the overall effect size of framing effects across studies. Our present meta-analysis contributes to this field of research in several ways. A previous meta-analysis mainly of valence framing effects on moral judgments (Demaree-Cotton, 2016) concludes that moral intuitions are fairly reliable. However, that meta-analysis failed to include many published studies and also suffered from technical flaws (McDonald, Yin, Weese, & Sinnott-Armstrong, 2019), such as concluding that approximately 80% of people's moral intuitions subject to framing effects don't change by taking the difference between the

Shared Introduction:

Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:

Condition 1:

If Program A is adopted, 200 people will be saved.
If Program B is adopted, there is 1/3 probability that 600 people will be saved and 2/3 probability that no people will be saved.
Which of the two programs would you favor?

Condition 2:

If Program C is adopted, 400 people will die.
If Program D is adopted, there is 1/3 probability that nobody will die and 2/3 probability that 600 people will die.
Which of the two programs would you favor?

proportion of moral judgments in distinct frames. By taking the difference between frame groups to show the proportion of people whose moral intuitions are changed by the frame, this interpretation obscures differences among subjects and among types of moral intuitions in susceptibility to framing effects. Further, a difference between framing groups (i.e. 80% difference) is not statistically equivalent to a 80% chance that a randomly selected person's moral judgment is determined by the frame. This inference misconstrues how effect sizes found in between-subjects designs are generalized to new populations. Our meta-analysis involves both between- and within-subjects designs for this purpose. The present meta-analysis is more complete and rigorous, and it provides evidence that our moral intuitions are less reliable than Demaree-Cotton suggests. Other meta-analyses examining framing effects are focused primarily on risky decisions involving neuroeconomic decisions such as mixed gambles, which are not exclusively moral decisions (Steiger & Kühberger, 2018; Kühberger, 1998). Finally, a related meta-analysis of moral judgments focuses on the effect of disgust on moral judgments (Landy & Goodwin, 2015). The present meta-analysis is distinct in its exclusive focus on how framing decisions in terms of positive and negative valence impacts moral decisions in particular.

2. Methods

2.1. Overview of the meta-analysis

We conducted a meta-analysis in order to estimate the magnitude of the difference between moral judgments made in positive (e.g. saving lives) frames compared to negative (e.g. not saving lives/lives lost) frames through a survey of all relevant, published literature. This estimation is intended not only to determine the extent to which presentation affects moral judgments, but also to inform the broader theoretical debate in moral psychology regarding how reliable moral judgments are. Moral intuitions that are subject to framing effects are unreliable because then certain factors change which moral judgments people accept but do not change which moral judgments are correct (Sinnott-Armstrong, 2008a, 2008b). Robust framing effects in moral judgments, thus, would undermine the central claims by moral intuitionists that our moral intuitions are reliable and, hence, justified in themselves (Audi, 2013; Demaree-Cotton, 2016; Shafer-Landau, 2005; Stratton-Lake, 2020).

This meta-analysis is divided into three sections. First, we estimate the pooled effect size of valence framing on moral dilemmas from the published literature. We then conduct a series of subgroup analyses to examine whether certain variables moderate the valence framing effect on moral decisions. Finally, we assess the published literature for the presence of publication bias and re-estimate the pooled framing effect size after accounting for publication bias.

2.2. Literature search

We began our search for relevant studies by searching within the following online publication databases: PubMed, PsycINFO, and Web of Science. In each of these databases, we conducted an intentionally broad full-text search using terms such as "framing" or "valence" or "moral judgment" (see Supplemental Materials for the full list of search queries used for each database). Of the resulting unique articles (k = 3613), we narrowed down the search results according to a set of inclusion criteria. For the inclusion and exclusion criteria, please see Table 1. Briefly, an experimental study was included in our database if the paradigm involved altering the framing of an inherently moral scenario or dilemma in a valenced way (i.e. scenarios had both a positive frame, using phrasing such as "lives saved", and a negative frame, using phrasing such as "lives lost"). An article would be excluded, for example, if it was not original, empirical research (e.g. a discussion article), if the dependent variable was not assessing moral judgments made by subjects, or if it did not meet the inclusion criteria (see Table 1) for what was considered a moral judgment. Importantly, in order to determine that the frames were logically equivalent, we required that a scenario framed in the positive frame have the same expected value as that scenario framed in the negative frame.

Table 1
List of inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Must alter the framing of a moral scenario/ dilemma in a valenced (positive/ negative) way.	Is not truly a framing effect (i.e. if an experimental manipulation changes the facts of a moral judgment, such as 5 or 15 workers in the trolley game)
The scenario vignette is a "moral judgment". Must meet the following definition of a "moral judgment"—must consist of one or more of these topics: • Harm to Others: ○ Death: Do not kill. ○ Disability: Do not blind, paralyze, or maim. ○ Loss of property: Do not steal. ○ Physical pain: Do not torture. ○ Psychological pain: Do not insult or make people feel bad. • Justice/Fairness: ○ Retributive: Do not punish more or less than is deserved. ○ Distributive: Do not treat people unequally. ○ Procedural: Give everyone a fair hearing and a fair chance. ○ Markets: Do not price gouge (or charge unfair prices) • Dishonesty: ○ Do not lie (or deceive). ○ Do not break promises. ○ Do not cheat (e.g., in games or in marriage). • Social position: ○ Hierarchy: Do not disrespect or disobey your parents or elders. ○ Role: Do your job and duty (e.g., as employee, citizen, or club member). ○ Loyalty (to an in-group): Be patriotic. Don't rat on friends. • Purity: ○ Sexual: Do not commit incest or necrophilia. • Gustatory: No cannibalism!	Focuses on behavior instead of moral judgments (i.e. a manipulation to the Trust or Ultimatum games that affects morality of behavior, not on explicit judgments). In other words, asks subjects to engage in an action (i.e. accept or reject an offer) rather than make a moral judgment. • We specifically do not include studies that use a Trust/Ultimatum Game paradigm, since cognitive factors other than moral judgments might be at play for some/all of subjects, making this paradigm outside the scope of our meta-analysis.
	Does not contribute original, empirical data (i.e. any discussion pieces)
	Ask subjects to make a choice or express their preferences instead of making a moral judgment (i.e. making a choice isn't enough, it has to be a moral judgment)
	Asks people to make a judgment about themselves (i.e. whether or not they should have a surgery given certain survival or fatality probabilities). • If subjects are asked to make a medical decision that impacts only them (i.e. the patient makes a decision about his/her own care), this is excluded. • However, if a subject is asked to make a medical decision on behalf of someone else, then this counts as a moral judgment.
	Mood framing studies (e.g. testing effects of negative versus positive mood on moral judgments) were not included, because they impact the state of mind in the judger but not presentation of the case.

The process of narrowing down the initial article set according to the inclusion and exclusion criteria occurred in two stages. The first stage was an abstract-review stage, in which we examined the abstract of each article; the article was either rejected when it was clear that it would not be relevant, or it proceeded to the full-review stage. At the full-review stage, we examined the methods sections to determine whether any studies within the article met our inclusion criteria. In both the abstract-review and full-review stages, two of our authors needed to agree on whether a study should be included or excluded; in cases of disagreement, the team discussed the study and came to a group consensus. In cases where it was not possible to extract the required effect size statistics, we contacted the corresponding author of the paper requesting the original data. This complete literature search produced a final set of 109 articles containing a total of 146 experiments/studies with a total N of 49,564 participants, with a mean sample size of approximately 178 participants per effect size. Of note, some studies yielded several effect sizes in cases where framing effects on multiple different types of moral judgments or framing effects on multiple subgroups of participants (i.e. women and men) were reported.

2.3. Obtaining effect sizes

This meta-analysis sought to estimate the effect size of valence framing effects in published experimental studies. We converted all results from the published relevant literature in the meta-analysis to standardized mean difference scores, or Cohen's *d* (Cook et al., 1992; Lipsey & Wilson, 2001). In most studies, *d* was calculated from reported proportions, means and standard deviations or standard errors, or *t* or *F* tests. When the necessary statistical information was not reported in an article or manuscript, we contacted the corresponding author and requested the statistical information or the raw data necessary to calculate *d*.

By and large, most experimental designs of psychology studies investigating framing effects employ between-subjects designs, in which experimenters assign one group of subjects to be exposed to a vignette's positive frame and another group of subjects to be exposed to a vignette's negative frame (Rehren and Sinnott-Armstrong, 2021). The effect of the experimental manipulation is assessed in a between-subjects design by comparing the responses of group A and group B. Alternatively, a within-subject experimental design involves each participant serving as his/her own control group, in which each subject is exposed to both frames over the course of the experiment. Only 19 articles in our dataset contributed effect sizes that were from within-subjects experimental designs, compared with 86 studies with a between-subjects design (in addition to 4 studies whose experimental design we were not able to obtain). Since effect sizes from different experimental designs cannot be directly compared or aggregated, we chose to analyze effect sizes separately based on experimental design type. The subgroup statistical analyses examining the role various moderators play in effect size magnitude are restricted to only the between-subjects design, due to the small sample size of effect sizes from within-subjects designs on framing effects.

Since our research question was in relation to the original Tversky and Kahneman (1981) study using the TKDP, we scored all effect sizes such that positive numbers indicate support for the Tversky & Kahneman framing pattern which observed risk-seeking for losses (negative context) and risk-aversion for gains (positive context). Moreover, negatively-valued effect sizes indicate an anti-Tversky & Kahneman framing pattern, such that a study would observe the opposite pattern (risk-seeking in gain contexts and risk-aversion in loss contexts). For studies that employed a valence framing manipulation of Trolley Problem scenarios (e.g. Cao et al., 2017), neither option involved uncertain outcomes; subsequently, we coded the utilitarian or consequentialist option the same as the "safe" choice (in the T&K scenario) and the non-consequentialist option as the "risky" choice (in the T&K scenario) when determining the direction of the extracted effect sizes (see Cao et al., 2017). This coding was based on Petrinovich and O'Neill (1996), finding that more participants agreed to act when the Trolley Problem options were presented in the Save frame, rather than the Kill frame; thus, our coding reflects the dominant direction of framing effects in both sets.

We conducted meta-analytic computations in R, using packages such as meta and metafor (Schwarzer, Carpenter, & Rücker, 2015; Viechtbauer, 2010). We conducted the statistical analyses of the extracted effect sizes from between-subjects studies in our dataset using a three-level mixed-effects meta-analysis model with restricted maximum likelihood estimation, in order to accurately account for some articles contributing multiple studies to our analysis, and also having some studies contributing multiple effect sizes from the same set of subjects (Harrer, Cuijpers, Furukawa, & Ebert, 2019; López-López, Page, Lipsey, & Higgins, 2018). This statistical model thus captures the hierarchical structure of our data.

3. Results

3.1. Valence framing effect point estimate

We calculated a point estimate for the size of the valence framing effect in moral decisions as well as the 95% confidence interval. Specifically, we calculated a weighted mean of the effect sizes in which each effect size was weighted by the inverse of the sampling variance of the study from which it was derived (Borenstein, Hedges, Higgins, & Rothstein, 2011; Harrer et al., 2019). Using a mixed-effects model, we found evidence supporting the existence and robustness of framing effects, with a pooled effect size estimate for between-subjects studies of $d = 0.50$ ($p < 0.0001$, 95% CI [0.44, 0.57]) and $d = 0.67$ ($p < 0.0001$, 95% CI [0.45, 0.89]) for within-subjects studies. For a forest plot displaying the results from the mixed-effects model for both between- and within-subjects studies, please see Supplementary Figs. 1 and 2. Using the Cohen's *d* conventional interpretation (Cohen, 2013), a value of *d* for small, medium, and large effects, respectively, are 0.2, 0.5, and 0.8. Thus, we find a moderate effect of valence framing for vignettes that require subjects to make moral judgments. Note that one should not directly compare these separate effect sizes and conclude there to be more of a framing effect in within-subjects designs compared to between-subjects designs. This is because a between-subject design's effect size (dIG, the difference in means between conditions divided by the pooled standard deviation of the sample) and a within-subject design's effect size (dRM, the mean within-subjects difference score, divided by the standard deviation of difference scores) are fundamentally different effect size statistics and therefore must be interpreted differently (Morris & DeShon, 2002).

We also found evidence for significant heterogeneity among the extracted effect sizes (between-subjects: Q(237) = 1720.49, $p < 0.0001$; within-subjects: Q(36) = 344.81, $p < 0.0001$). Since generalizations of I^2 for more complex models like multilevel models are still an ongoing area of research (Borenstein et al., 2011), for estimates of variability, we report statistics from a single-level random effects model (between-subjects: $I^2 = 80.28\%$, $tau^2 = 0.10$; within-subjects: $I^2 = 92.17\%$, $tau^2 = 0.22$). This led us to assess the degree to which a set of moderating variables (see below) influences the overall observed valence framing effect.

3.2. Moderator analyses

We investigated the extent to which a set of moderating variables influences the effect size of framing effects on moral dilemmas. We included variables that were suggested in the framing effects literature to be potentially important moderators (Rehren & Sinnott-Armstrong, 2021; Kühberger, 1998*)*. A priori, we selected the following variables to test for moderating effects: 1) experimental design type (between-subjects vs within-subjects designs), 2) scenario type (involving mortality vs no mortality, see Hayashi & Sasaki, 2013), 3) student vs non-student subjects, and 4) proportion of male to female subjects. During the data analysis, we determined we were unable to assess the effect that gender plays in valence framing effects since 39.2% of our extracted effect sizes belonged to studies that did not report the gender proportions of their subject pool, although this is an important area of research that should be investigated for future research. During data analysis, as an exploratory follow-up analysis to the student vs non-student moderator, we added online vs in-person studies as a moderator to test whether online studies differed in valence framing effects compared with in-person data collection. As mentioned above, statistical analysis of these moderating variables will only be applied to effect sizes from between-subjects designs due to the small number of effect sizes from within-subjects designs. For statistical tests of whether a moderating variable significantly impacts the aggregated effect size, we report Q, a test statistic for the omnibus test of coefficients used by the metafor package in R (Viechtbauer, 2010).

3.2.1. *Mortal vs non-mortal*

A large proportion of the scenarios included were similar to the original TKDP scenario, and this involved choices of life or death. Thus, we wanted to examine how the effect size of scenarios that involved risk to human lives ($k = 214$) differed from scenarios that did not involve risk to human lives ($k = 25$). Effect sizes were excluded from this subgroup analysis if they involved multiple scenarios consisting of mortal and non-mortal scenarios, and a study involving alien lives was also excluded. We found that scenarios involving risk of human lives ($d = 0.51$) had an effect size slightly larger than scenarios that did not involve risk of human lives ($d = 0.43$), however this difference was not significantly different (Q = 2.03, $p = 0.154$). Thus, valence framing effects on moral judgments are still as prominent when the scenario is not a matter of life and death.

3.2.2. *Scenario types*

One potential concern is that the TKDP scenario is an isolated case, and other scenarios as well as modifications of the TKDP will not yield as large framing effects. Thus, we carried out multiple multilevel models to determine the differences between various types of scenarios.

Initially, we looked at the effect sizes of specifically TKDP scenarios with no modifications ($k = 116$), compared to all other scenarios ($k = 127$). There was a significant difference between the scenarios where the original TKDP scenario had a higher effect size ($d = 0.55$) than all other scenarios ($d = 0.46$) (Q = 5.62, $p = 0.0178$).

Next, we looked at the original TKDP scenario (k = 116) and TKDP-modified scenarios (k = 91). A scenario was considered to be TKDP-modified, if it had the same set up as the original problem but there was a modified aspect, such as the number of lives at stake or changing the original disease to another catastrophe. There was no significant difference in effect size between the two types of scenarios (Q = 0.0056, $p = 0.940$; TKDP $d = 0.54$, TKDP-modified $d = 0.53$). This indicates that the framing effect still holds for modified versions of the TKDP scenario.

One theoretically important aspect of the scenario used is whether, in a given binary decision, one or both options contain risk (or if they are "certain" options where an outcome is known for sure). We were unable to determine whether this aspect of scenario is significant in effecting the magnitude of framing effect because the majority of effect sizes were TKDP-variants that included one risky and one certain option. Nevertheless, this is a potentially important moderating variable of framing effects that can be investigated in future studies.

3.2.3. *Student vs non-student*

One common and valid criticism of the field of psychology as a whole is the propensity for researchers to recruit predominantly American undergraduates. Not only do these research participants tend to fall into a very restricted age range (i.e. 18–22 years), but also these participants tend to be "WEIRD" (Western, educated, industrialized, rich and democratic) (Henrich, Heine, & Norenzayan, 2010). We first wanted to investigate the extent to which being a student impacted the estimated effect size of a valence framing experiment. We defined "student" to be either an undergraduate or graduate student at either a domestic or foreign university. We found the classification of a participant group as either student or non-student to be non-significant (Q = 0.53, $p = 0.467$; non-student $d = 0.46$, student $d = 0.52$), suggesting that there is no statistically significant difference in framing effects of moral decisions between student and non-student populations.

3.2.4. *Online vs in-person data collection*

As an exploratory follow-up to studying whether aspects of the participant population (student vs non-student in particular) moderate the magnitude of the valence framing effect, we also tested whether conducting the experiment online vs in-person had a significant difference on framing effects. While there was a smaller amount of effect sizes derived from online studies ($k = 31$) compared to in-lab studies ($k = 202$) (in part due to experimental norms and in part due to publication year), we found that there was no significant difference of effect size magnitude between online vs in-person studies (Q = 0.0022, $p = 0.963$; in-person $d = 0.499$, online $d = 0.503$).

3.3. *Assessing publication bias*

Finally, we wanted to quantify the level of publication bias present in our effect size dataset. One nefarious issue in empirical research is the *file-drawer* or *publication bias* problem, in which a study that reports a higher effect size is more likely to be published than a study that reports a smaller effect size (Borenstein et al., 2011; Dickersin, 2005; Rothstein, Sutton, & Borenstein, 2006). This leads to meta-analyses analyzing published studies estimating an inflated effect size for the variable of interest compared to what the average effect size would be if no publication bias was present, since likely weaker effects are not present in the published studies. Further, meta-analyses are inherently uncertain as to how many unpublished, relevant studies with likely smaller effect sizes are missing from the final, extracted dataset. We assessed the level of publication bias present in our framing effects data both visually with funnel plots and statistically using Egger's test.

First, in order to visualize the possible presence of publication bias in our data, we created a funnel plot (see Fig. 1), frequently used in meta-analyses to plot each study's standard error (which is a function of the number of subjects in each study) and effect size (Peters, Sutton, Jones, Abrams, & Rushton, 2008). If there was no publication bias present in the data, then studies would appear symmetrically around the pooled (between-subjects) effect size mean (d = 0.50) within the funnel denoted with a dashed line. In particular, studies with high sampling error (i.e. studies with fewer numbers of participants) near the bottom of the plot would appear symmetrically distributed around the pooled effect size mean, since small studies reporting small effect sizes would have an equal chance of being published as small studies with large effect sizes, absent publication bias. In our funnel plot in Fig. 1, we observe a lack of studies in the bottom-left area of the plot, meaning that in our data of published effect sizes, few studies have both high standard error (fewer participants) and a lower effect size. Interestingly, we observe a large cluster of studies lying right along the $p < 0.05$ contour. We suspect this straight line (that would be very unlikely to appear by chance) might arise due to systemic factors such as potential bias in journals deciding what research to publish or reject, or researchers (consciously or otherwise) manipulating experimenter degrees of freedom in order to reach statistical significance at $p < 0.05$ level (Simmons, Nelson, & Simonsohn, 2011).

The visual results from the contour-enhanced funnel plot suggest small-sample publication bias, in which studies with higher standard

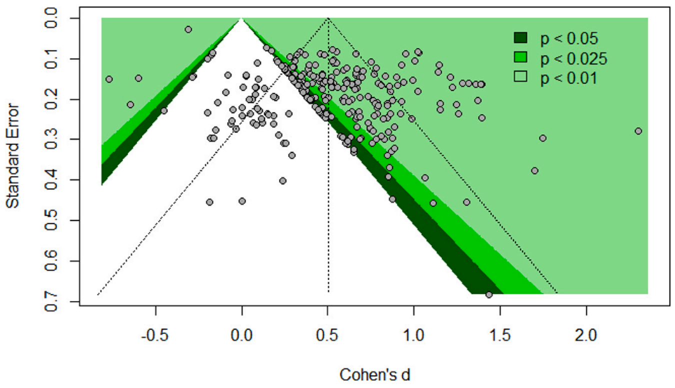

Fig. 1. Contour-enhanced funnel plot of published moral valence framing effect sizes. Each dot represents a study; X-axis is the magnitude of the effect size (Cohen's d), and the y-axis is the standard error of the effect estimate. Shaded regions denote the level of statistical significance of the studies in our dataset. The vertical dashed line denotes the pooled effect size.

error and smaller effect sizes likely missing from the publication data. We then wanted to quantify whether there was statistically significant publication bias present in the data. Consistent with the visual results, Egger's test was significant (Intercept = 2.97, $t = 7.51$, $p < 0.0001$), indicating that more studies with higher sampling error are associated with larger effect sizes more often than would be expected by chance (Egger, Smith, Schneider, & Minder, 1997). This is consistent with the hypothesis that small-sample publication bias is present in our moral valence framing effects dataset.

Due to the suggestive evidence of publication bias in the literature on valence framing of moral decisions, we next wanted to determine the magnitude of the pooled effect size of valence framing studies when this publication bias is statistically accounted for. Since Egger's test was found to be significant, we used Duval & Tweedie's trim-and-fill procedure to estimate what the actual effect size would be if these hypothesized small missing studies had been published by imputing the missing studies into the funnel plot until symmetry is achieved (Duval & Tweedie, 2000). After imputing 93 added studies (27.7% of the resulting studies for this analysis) to produce the expected symmetric funnel plot, the estimated effect size was attenuated, yet still remained significant ($d = 0.22$, 95% CI = [0.16; 0.28], $p < 0.0001$). Thus, while we find suggestive evidence for the presence of publication bias in published studies on valence framing in dilemmas that require people to make moral decisions, we can conclude that this only lessens the previously moderate effect size, but does not refute the robust presence of valence framing effects on moral judgments.

4. Discussion

How reliable are our moral intuitions and judgments? Our meta-analytic review of the published literature on valence framing effects on moral dilemmas shows a moderate ($d = 0.50$) framing effect on moral judgments when otherwise equivalent vignettes are framed either positively (e.g. lives saved) or negatively (e.g. lives lost). While we find that small-sample publication bias is present in our dataset, such that small-sample studies with weaker effect sizes are missing from the literature, we find that statistically accounting for this publication bias attenuates ($d = 0.22$), but does not eliminate, the presence of valence framing effects on moral judgments. This indicates there is a robust and consistent valence framing effect in moral decision-making that is roughly one-fourth to half of a standard deviation.

We also identified variables of interest that potentially moderated the strength of the valence framing effect. In particular, we found that studies involving original TKDP scenarios reported a significantly higher effect size ($d = 0.55$) than all other scenario types ($d = 0.46$), with no significant difference found between original TKDP scenarios and TKDP-modified vignettes. This suggests that part of the framing effect size observed in a particular experiment is driven by the chosen vignette. This is something researchers seeking to elaborate on framing effects research should be aware of when designing experiments.

The overwhelming preference of the psychology field to utilize between-subjects designs to measure framing effects limits the field's ability to probe important questions regarding the nature of valence framing effects within the context of making moral judgments. For example, for individuals who are subject to valence framing effects, is the resulting shift in moral judgment uniform across the population, or are some people highly susceptible to framing effects (i.e. experiencing large shifts in their moral judgments as a function of frame) while others are only mildly susceptible to framing effects (Rehren & Sinnott-Armstrong, 2021)? Are individuals equally susceptible to all framing effect types (including valence and order presentation effects as well as cleanliness circumstantial effects) or do individual differences determine one's framing effect "susceptibility profile" (Demaree-Cotton, 2016; Landy & Goodwin, 2015; Petrinovich & O'Neill, 1996)? The nature of these questions can be sufficiently answered only with well-designed within-subject paradigms that specifically probe how individual differences moderate the strength of individual valence framing effects.

It has become increasingly clear in recent years that restricting empirical study (particularly in psychology research) to predominantly student populations is not only non-inclusive, but also skews results and the collective literature's understanding of the research question at hand (Henrich et al., 2010; Jones, 2010). Advances in online data collection, such as more widespread use of online platforms such as Amazon Mechanical Turk have the promise of making data-collection more representative of the general population compared to the current practice of favoring undergraduate samples for data collection (McCredie & Morey, 2019). Furthermore, we found that there was no significant difference between valence framing effect sizes reported from in-person vs online studies. This validates previous research showing no statistically significant difference between more controlled in-lab studies and online experimental platforms such as Amazon Mechanical Turk (Crump, McDonnell, & Gureckis, 2013).

Since Tversky & Kahneman published the first study noting the phenomenon of valence framing effects in 1981, valence framing in moral dilemmas such as the famous disease problem have raised questions regarding the reliability of our moral intuitions and judgments. Does the moral choice regarding medical care for a loved one or a public policy that impacts strangers align with the decision-maker's true moral values, or is a given decision swayed by how the information about the options was presented? Our meta-analytic evidence lends more support to the questioning of the reliability of moral judgments.

While the evidence we present in this study suggests that the average person's moral judgments are likely to be subject to valence framing effects, we want to underscore the point that the extent to which individual differences play a role in one's own susceptibility to framing is still unclear. Our finding of a moderate effect size of valence framing on moral decisions is a population measure, so the extent to which individual differences such as personality, cognitive measures, and demographic factors might make one more or less susceptible to framing effects is unclear. It is also unclear whether other types of framing effects (e.g. order effects) present additive or even multiplicative effects on one's propensity to make judgments that are influenced by frame. More research examining the role that these sorts of individual differences play in framing susceptibility would be crucial in answering these questions, so individuals making important decisions such as medical or legal decisions can assess the reliability of their judgments.

The current study has important implications for moral philosophers debating the justification of individual moral judgments. Moral intuitionists typically argue that that moral judgments must be formed through a reliable process in order to be justified in themselves. Valence framing effects, however, are only one of many different biases that undermine the reliability of non-inferential processes of forming moral judgments. Without a reliable process by which to form moral judgments, the arguments of moral intuitionists fail to establish that any moral judgment can be justified without inferential support. This does not mean that none of our moral judgments are ever justified; they just need to be justified through inferential processes rather than by mere intuition alone.

Moral intuitionists sometimes reply that supposedly equivalent frames actually convey different information to survey participants, so they are not truly equivalent (e.g., Aczel et al., 2018; Frisch, 1993; Mandel, 2014; McKenzie & Nelson, 2003; Sher & McKenzie, 2006). This response is controversial and would need to be applied case by case, so it cannot be debated here for all of the studies we analyzed. In any case, it would not affect our estimates of the effect sizes of the frames, and framing effects would still show unreliability when frames really are equivalent.

Another reply by moral intuitionists is that the reported effect sizes are not large *enough* to require confirmation by inference at least when very little is at stake, as in many hypothetical scenarios. However, critics of moral intuitionism can argue that moral mistakes are sometimes very costly in real life, such as when moral disagreements turn friends into enemies or exacerbate political polarization; and then a smaller amount of unreliability might be enough to show that we should not trust our moral intuitions without independent confirmation. This normative dispute cannot be settled by science alone, but it also cannot be settled without determining empirically how much moral judgments are subject to framing effects, which was our goal here.

The implications of this study reach beyond abstract debates in moral philosophy about belief justification. Assume that a doctor is considering different treatment options for a patient with a serious, life-threatening condition (perhaps like COVID-19). The doctor's framing of that treatment to the patient in terms of survival rate versus mortality rate could change the patient's decision on whether to accept the treatment. Since we measured framing effects as a population measure, thinking about how to frame important policy measures for public health such as hand-washing or wearing masks could potentially lead to thousands of lives being positively or negatively impacted. Furthermore, the way that treatments are presented to doctors themselves by actors like pharmaceutical companies could influence whether or not the doctors even put the treatment options forward to their patients. Now assume that you are an attorney presenting information to a jury at a trial. Framing the evidence positively or negatively could make a difference in whether or not the jury finds a defendant guilty even though the information presented is the same. In effect, the jury's moral decision to condemn or not condemn a defendant is based on morally irrelevant information like the way the material is framed.

Acknowledgments

We are incredibly thankful to Megan Van Noord for collaborating with the authors in conducting the literature search, as well as Paul Rehren, Dr. Harris Cooper, and JJ Moncus for insightful discussions that strengthened this manuscript. Research reported in this publication was supported by a National Science Foundation Graduate Research Fellowship under NSF GRFP DGE-1644868 and a Duke Dean's Graduate Fellowship to K.M. For additional financial support, we also thank the Kenan Institute of Ethics at Duke University and the John Templeton Foundation (grant ID #48365). Of course, these funders are not responsible for the content.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104703.

References

Aczel, B., Szollosi, A., & Bago, B. (2018). The effect of transparency on framing effects in within-subject designs. *Journal of Behavioral Decision Making, 31*(1), 25–39.

Audi, R. (2013). *Moral perception*. Princeton: Princeton University Press.

Blackburn, S. (1996). Securing the nots. In W. Sinnott-Armstrong, & M. Timmons (Eds.), *Moral knowledge? New readings in moral epistemology*. New York: Oxford University Press.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Cao, F., Zhang, J., Song, L., Wang, S., Miao, D., & Peng, J. (2017). Framing effect in the trolley problem and footbridge dilemma: Number of saved lives matters. *Psychological Reports, 120*(1), 88–101.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., & Light, R. J. (Eds.). (1992). *Meta-analysis for explanation: A casebook*. Russell Sage Foundation.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), Article e57410.

Demaree-Cotton, J. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology, 29*(1), 1–22.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments* (pp. 11–33).

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj, 315*(7109), 629–634.

Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes, 54*(3), 399–429.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing Meta-analysis in R: A Hands-on Guide*. https://doi.org/10.5281/zenodo.2551803. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/.

Hayashi, Y., & Sasaki, H. (2013). Situational and Dispositional Factors Moderating Three Types of Framing Effects: Mortality Salience and Regulatory Focus. *Tohoku psychologica folia, 71*, 42–56.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302), 29.

Hernandez, J. G. (Ed.). (2011). *The new intuitionism*. A&C Black.

Jones, D. (2010). *A WEIRD view of human nature skews psychologists' studies*.

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes, 75*(1), 23–55.

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science, 10*(4), 518–536.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.

López-López, J. A., Page, M. J., Lipsey, M. W., & Higgins, J. P. (2018). Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Research Synthesis Methods, 9*(3), 336–351.

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General, 143*(3), 1185.

McCredie, M. N., & Morey, L. C. (2019). Who are the Turkers? A characterization of MTurk workers using the personality assessment inventory. *Assessment, 26*(5), 759–766.

McDonald, K., Yin, S., Weese, T., & Sinnott-Armstrong, W. (2019). Do framing effects debunk moral beliefs? *Behavioral and Brain Sciences, 42*.

McKenzie, C. R., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin & Review, 10*(3), 596–602.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological methods, 7*(1), 105.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology, 61*(10), 991–996.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17*(3), 145–171.

Rehren, P., & Sinnott-Armstrong, W. (2021). Moral Framing Effects Within Subjects. *Philosophical Psychology*. https://doi.org/10.1080/09515089.2021.1914328.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley and Sons [Google Scholar].

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R* (Vol. 4784). Cham: Springer.

Shafer-Landau, R. (2005). *Moral realism: A defense*. Oxford: Clarendon Press.

Sher, S., & McKenzie, C. R. (2006). Information leakage from logically equivalent frames. *Cognition, 101*(3), 467–494.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Sinnott-Armstrong, W. (2001). Intuitionism. In C. Becker (Ed.) (Second edition,*, Vol. II. Encyclopedia of Ethics* (pp. 879–882). London and New York: Routledge.

Sinnott-Armstrong, W. (2007). *Moral skepticisms*. New York: Oxford University Press.

Sinnott-Armstrong, W. (2008a). Framing moral intuitions. In W. Sinnott-Armstrong (Ed.), *The cognitive science of morality: Intuition and diversity: Vol. 2. Moral psychology* (pp. 47–76). Cambridge, MA: MIT Press.

Sinnott-Armstrong, W. (2008b). How to apply generalities: Reply to Tolhurst and Shafer-Landau. In W. Sinnott-Armstrong (Ed.), *The cognitive science of morality: Intuition and diversity: Vol. 2. Moral psychology* (pp. 97–105). Cambridge, MA: MIT Press.

Steiger, A., & Kühberger, A. (2018). A meta-analytic re-appraisal of the framing effect. *Zeitschrift für Psychologie, 226*(1), 45.

Stratton-Lake, P. (2020). In Zalta Edward N. (Ed.), *Intuitionism in Ethics. The Stanford Encyclopedia of Philosophy (Summer 2020 Edition)*. https://nam03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fplato.stanford.edu%2Farchives%2Fsum2020%2Fentries%2Fintuitionism-ethics%2F&amp;data=04%7C01%7Ce.mishael%40elsevier.com%7C182026e5a2b744fd150108d909c01cc7%7C9274ee3f94254109a27f9fb15c10675d%7C0%7C0%7C637551542993846572%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&amp;sdata=EUaoAqvBFIdgN0IG7TjzSzuVCRkUXArpXfD9rNX442Q%3D&amp;reserved=0.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.