How to show that a cruel prank is worse than a war crime: Shifting scales and missing benchmarks in the study of moral judgment

Vladimir Chituc[1], M. J. Crockett[2], & Brian J. Scholl[1]

[1] Department of Psychology, Yale University

[2] Department of Psychology and University Center for Human Values, Princeton University

Running Head : Shifting scales and missing benchmarks

Address for correspondence : Vladimir Chituc
Department of Psychology
Yale University
Box 208047
New Haven, CT 06520-8047

Email : vladimir.chituc@yale.edu
Phone : 607-857-4166

Word Count : 9575
Version : 8/31/25 — In press, *Cognition*

Abstract (223 words)

Moral judgment is central to both everyday life and cognitive science, but how can it be studied with quantitative precision? By far the most direct and ubiquitous method is to simply ask people for their judgments, in the form of ratings on a labeled scale (e.g. Likert or Visual Analog Scales). As has long been recognized in sensory psychophysics, however, such responses are meaningful only in a relative sense. (Is your dog "big"? Perhaps yes in the context of house pets, but not in the context of all mammals?) Here we illustrate the nature and extremity of this problem using two case studies. First, to explore this theme in principle, we show in a series of nine experiments that this problem can readily lead subjects to (seemingly) judge a cruel prank (involving humiliation) to be just as immoral as (or even worse than) an internationally recognized war crime (involving murder). In contrast, such nonsensical results disappear when using magnitude estimation — a psychophysical method employing an explicit moral benchmark. Second, to demonstrate the importance of this theme in practice, we show that the use of magnitude estimation (vs. Likert scales) radically changes the proper interpretation of a recent study of 'moral luck', fueling essentially the opposite conclusion. Taken together, this work illustrates how insights from psychophysics can help improve measurement in contemporary moral psychology.

Keywords

Moral psychology, moral judgment, between-subjects design, magnitude estimation, psychophysics

The LORD detests dishonest scales, but accurate weights find favor with him.
(Proverbs, 11:1)

1. Introduction

Morality is central not only to everyday life, but also to the study of how the mind works. Studying morality can reveal the nature of various underlying cognitive mechanisms (e.g. Cushman & Greene, 2012), and can serve as a hub which links together the central methods and concerns of many different parts of cognitive science and beyond — from philosophy and psychology, to economics and evolutionary biology (Heiphetz & Cushman, 2021). Moreover, morality influences many mental processes: it "predominates" social judgment (Goodwin et al., 2014; cf. Melnikoff & Bailey, 2018), is central to our sense of identity (Strohminger & Nichols, 2014), and affects judgments of beauty (Dion et al., 1972), causality (Kominsky et al., 2015), possibility (Acierno et al., 2022), intentionality (Knobe, 2003), agency (Khamitov et al., 2016), and emotion (Prinzing et al., 2023). And *immorality* seems particularly consequential for cognition: it stokes curiosity (Wylie & Gantman, 2023), grabs attention (Vanneste et al., 2007), distorts memory (Carlson et al., 2020; Pizarro et al., 2006), constrains imagination (Liao et al., 2014), and dilates subjective time (Jia et al., 2021) — and even *pretending* to act immorally can increase blood pressure (Cushman et al., 2012).

1.1 Moral Measurement

One factor that may help to explain the current renaissance in explorations of moral psychology (Malle, 2021) is that moral intuitions seem relatively easy to study. If you want to know how motion is extracted from visual stimuli, you need to conduct subtle psychophysical experiments; but if you want to know whether people find something to be immoral, you can just ask them. And how can moral judgment be studied with quantitative precision? This also seems straightforward: you can still just ask them, but use numbers. As such, one simple experimental method predominates the study of moral judgment: subjects read a short vignette

(about a person, an event, or a dilemma, etc.; for libraries of examples, see Clifford et al., 2015; Knutson et al., 2010), and then they record their judgment using a scale that is labeled with the relevant moral concept.

Perhaps the two most ubiquitous sorts of scales used for this purpose are the Likert scale (Likert, 1932) and the Visual Analog Scale (Aitken, 1969). Likert scales require subjects to choose among a limited number of discrete response categories — e.g. a 7-point scale that ranges from *not at all wrong* to *very wrong* (e.g. Gray & Keeney, 2015), *extremely immoral* to *extremely moral* (e.g. Minson & Monin, 2012), *not at all blameworthy* to *very blameworthy* (e.g. Young et al., 2010), *forbidden* to *obligatory* (e.g. Cushman et al., 2006), etc. Visual Analog Scales, in contrast, allow subjects to report a judgment that falls anywhere along a continuum that is anchored with the same sorts of labels (e.g. Siegel et al., 2017; Sosa et al., 2021) — e.g. clicking on a line with a left endpoint labeled *not at all wrong*, and a right endpoint labeled *very wrong*. These two measures are often collectively referred to as *labeled scales* (Bartoshuk et al., 2003).

Labeled scales can differ dramatically from study to study in terms of their endpoints, number of labels, and wording, but in a general sense they are utterly ubiquitous in the study of moral psychology. For example, out of the 50 most recent research articles published in *Cognition* that contained the word “moral” in the title or as a keyword at the time of this writing: 47 used a labeled scale to test a key hypothesis: 31 used a Likert scale, 6 used a Visual Analog Scale, and 10 used both. (Of the remaining articles, one was an infant study using dichotomous choice as a dependent measure, another used cheating behavior as a dependent measure, and the third was a meta-analysis. A complete list is included in the Supplementary Data file.)

This ubiquity is partially understandable, since of course such scales seem like the most direct possible ways of asking about moral judgment, and are easy to implement in online studies which have come to dominate social psychology over the past decade (Anderson et al., 2019). But this ubiquity is also surprising, since such scales have an especially deep problem:

these scales are always used within a certain frame of reference (Parducci, 1965), and unless that frame of reference is in some way made explicit, then it is not possible to compare one judgment to another. Put differently, unless these scales are anchored to some shared baseline, then responses on them are inevitably *relative*. This problem is extremely familiar in the context of everyday adjectives. Consider "big", for example: 9 seems intuitively like a big number (since it implicitly evokes the range of 1-10), but 221 seems intuitively like a small number (since it implicitly evokes the range of 1-1000; Birnbaum, 1999; Leong et al., 2019). And similarly, you might naturally say that you have a big dog, but a small house — despite the obvious fact that the latter still dwarfs the former (but see Bridwell, 1963). Such observations seem obvious and pedestrian, but they can pose substantial problems for many experiments — especially those with a between-subjects design. Even in a uniform population, one group might give an average response of 5 out of 7 when asked "How big is your dog?" on a Likert scale, and another group might give a 3 out of 7 when asked "How big is your house?" — but that doesn't provide experimental support for the notion that this population thinks they have houses that are smaller than their dogs. And to foreshadow, this problem — what has been called "one of the most difficult problems concerning intersubjectivity" (Borg, 2001) — is just as salient when the relevant adjective is "bad" instead of "big".

This challenge — both its conceptual foundation and the trouble it can cause for between-subjects experiments — has been widely appreciated in the study of psychophysics, but in an odd manner. The key insight has been recognized for a very long time. More than 65 years ago, for example, this same point was made by S. S. Stevens in his seminal studies of auditory psychophysics (Stevens, 1958), as when exploring how loud sounds appear to be (Stevens, 1956). Yet more than 40 years later, this same point had to be re-discovered in the context of taste psychophysics, e.g. to show how certain studies with labeled scales (say, of how strong some tastes are) missed large differences in actual sensory experience (e.g. Bartoshuk, 2000; Bartoshuk et al., 2003). And in general, this point seems to be strangely both intuitive yet

subtle, such that it needs to keep being rediscovered, especially in different subfields (Borg, 2001). (For other examples, see Biernat et al., 1991; Biernat & Manis, 1994; Ubel et al., 2005). As Linda Bartoshuk once noted, this insight "hasn't penetrated anywhere, because this mistake is being made all over the place" (as cited in Borg, 2001).

Nevertheless, foundational work in psychophysics provides a surprisingly straightforward solution to this challenge: when making measurements, use an explicit benchmark. This is what is done in the standard psychophysical method of *magnitude estimation* (e.g. Stevens, 1956, 1966b; for short primers, see: Lodge & Tursky, 1981; Moskowitz, 1977). This method provides subjects with a benchmark stimulus with a certain pre-assigned value, and then asks raters to make all responses relative to that baseline, in a particular manner — e.g. assigning the brightness of a benchmark light as a 10, and then asking them to rate a light that seemed half as bright as a 5, a light that seemed twice as bright as a 20, etc. In contrast to the vast number of moral psychology studies that have used labeled scales, we are aware of only a handful of prior studies that have ever employed magnitude estimation in this domain (e.g. D. J. Cohen & Ahn, 2016; Sellin & Wolfgang, 1964), or indeed used any sort of explicit benchmark (DeScioli et al., 2011; Tannenbaum et al., 2011)

1.2 The Current Studies

Here we suggest that one of the places where such challenges are still salient is in the study of moral judgment. Even more, we think these issues may be particularly relevant in this domain, for two intertwined reasons. First, as noted above, the use of labeled scales has utterly dominated this field. Second, this field has been particularly sensitive to the dangers of within-subjects designs. Indeed, much of the work in this field employs different versions of the same scenario, carefully holding all factors constant except for one — e.g. whether harm was intentional or not (e.g. Cushman, 2008), whether personal force was used (e.g. Greene et al., 2009), or whether an outcome was framed in terms of the proportion that died or the proportion that lived (e.g. McDonald et al., 2021). In such cases, a within-subjects design might

impose task demands by far too readily highlighting the relevant factor (which might not otherwise be salient; see also Hsee, 1996) — and as a result, such studies have frequently relied on between-subjects comparisons (a logic that we return to below in Experiments 4a and 4b).

These two problems — both the shifting meaning of labeled scales in between-subjects designs, and the danger of task demands imposed by within-subjects designs — can be notoriously subtle and dangerous, but they can also be readily solved. We illustrate this here using two case studies of moral judgment.

In an initial exploration, to show that these issues matter in principle, we employed these manipulations and measures in a series of 9 experiments exploring a somewhat fanciful theoretical example — in which a prank is judged to be just as bad as a war crime. We first demonstrated that this seemingly nonsensical pattern of results holds when tested between-subjects (but not within-subjects) using both a Likert scale (Experiment 1a) and a Visual Analog Scale (Experiment 1b), but not with magnitude estimation (Experiment 1c). We then replicated this pattern (for all three scale types) when the moral context was established by an explicit contrast stimulus, rather than being evoked implicitly by a scenario itself. This was implemented in both a relatively common way (using explicit contrasts of sending a prank email vs. multiple murder; Experiments 2a-2c), and in an even more extreme way (using explicit contrasts of jaywalking vs. terrorism targeting a preschool; Experiments 3a-3c).

Next, to show that these issues matter in practice, we explored an empirical example taken from the actual contemporary literature in moral psychology. We focused in particular on the notion of 'moral luck', in which two agents who have performed the same action (say, driving drunk) are nonetheless evaluated very differently based only on differential outcomes beyond their control (say, hitting a tree vs. hitting a child). A prominent recent study (Kneer & Machery, 2019) calls the very existence of this puzzle into question, finding that the traditional effect of moral luck (obtained between-subjects) is drastically reduced (if not entirely eliminated) when scenarios are jointly presented within-subjects. We replicate this finding

when using a Likert scale (in Experiment 4a), but we also show that this result disappears when using magnitude estimation (in Experiment 4b). As such, this example shows how the use of more sensitive measures can produce radically different conclusions even in well-explored cases.

2. Experiments 1a – 1c: Cruel Prank vs. War Crime (Raw Scenarios)

In the first case study, we developed a vignette describing an act that seemed highly immoral in the context of everyday behavior: a *Prank* scenario, in which someone is publicly humiliated for her weight. We then contrasted this with another vignette describing an act that also seemed highly immoral but in a far more extreme context: a *War Crime* scenario, in which a soldier kills a defenseless prisoner of war. (These vignettes were meant to evoke different moral contexts, just as one could evoke different numerical contexts when judging whether a number is "big"; Birnbaum, 1999; Leong et al., 2019.) Across several variations, we show that when participants evaluate such scenarios using labeled scales, between-subjects, they seemingly rate a cruel prank to be just as immoral as (or even worse than) an internationally recognized war crime.

We first tested whether the key nonsensical result (where a war crime is seemingly rated as no worse than a prank, in between-subjects conditions) would occur with the raw scenarios presented without any additional contrast stimuli, with a Likert scale (Experiment 1a), a Visual Analog Scale (Experiment 1b), and magnitude estimation (Experiment 1c). To implement magnitude estimation in this context, we assigned a value of 10 to the immorality of stealing a wallet. Subjects then made judgments of immorality in reference to that benchmark, such that they would assign a value of 20 to something that seemed twice as immoral, and a 5 to something that seemed half as immoral, etc. (Each of these measures is depicted in a screenshot in the Supplementary Data file.)

2.1 General Method for Experiments 1-3

Since the first three experiments shared the same basic structure, we first provide a general methodological overview. All hypotheses, analyses, and sample sizes were preregistered, and the Supplementary Data file provides all raw data and links to preregistrations.

2.1.1 Subjects. Online subjects (all from the US, who had completed at least 10 prior online studies with at least a 90% approval rate) were recruited using Prolific (Palan & Schitter, 2018). No subject participated in more than one experiment, and all testing was conducted using the Qualtrics survey platform.

2.1.2 Measures. The nine initial experiments collectively used three different measures. Subjects in Experiments 1a, 2a, and 3a made their judgments using a 7-point Likert scale (implemented using the "modern" Qualtrics layout and default font) that ranged from 1 ("not at all immoral") to 7 ("very immoral") — with the numbers arrayed horizontally, and with the two labels presented just above the most extreme values. Subjects responded by clicking on one of the seven visible numbers.

Subjects in Experiments 1b, 2b, and 3b made their judgments using a 100-point Visual Analog Scale that ranged from 0 to 100 — with these values placed just above the endpoints of a visible horizontal line, and with the same two labels presented just above their respective values. Subjects responded by clicking and dragging a slider (implemented in Qualtrics as a small blue disc) arrayed along the visible line (and initially placed at the line's center).

Subjects in Experiment 1c, 2c, and 3c made their judgments using *magnitude estimation*, instructed as follows:

> Please use a 0 to mean "neither moral nor immoral." (Imagine something like playing with a pen. It's neither morally bad nor morally good, it just is.)
>
> As a 10, we want you to think about the morality of the following event: stealing a wallet. This event is called your benchmark.
>
> In this study, we will ask you to rate other events relative to this benchmark. If the event describes something just as immoral as the benchmark, you should also rate it a 10. If it

> describes something half as immoral as the benchmark, you should rate it a 5. If it is twice as immoral as the benchmark, rate it a 20, and so on. You can use decimals (e.g. "7.5") if you feel like you need to.
>
> Please try to keep the benchmark and these instructions carefully in mind when you give your responses. If you read about something that seems extremely immoral for example, please don't respond with an arbitrarily high number. Instead, really think about just how much more immoral it is relative to the benchmark, and rate it accordingly.
>
> There's no right or wrong answer to this question, so just do your best and try to be as careful and thoughtful as possible.

Subjects entered their rating in a free-response text box which accepted only a non-negative number. (Since subjects were free to enter a number of any size, we preregistered 10,000 as an arbitrary cutoff point for exclusions, since our results might otherwise be skewed by implausibly large responses. In practice, this resulted in the exclusion of 3 subjects in Experiment 1c, 2 subjects in Experiment 2c, and 7 subjects in Experiment 3c (or 1%, 0.7%, and 4.7% of subjects for each study, respectively). Following standard practices for magnitude estimation (Butler et al., 1987), data were log-transformed prior to analysis (with raw ratings all first incremented by one, to avoid undefined values).[1]

2.1.3 Stimuli. The primary goal of each experiment was to compare how subjects rated the morality of the event described in each of two scenarios. The first scenario (*Prank*) was inspired by a story the first author read years ago online:

> A college freshman has a crush on a boy in her class. He invites her to a party, but it turns out to be a prank hosted by a fraternity to see which pledge can invite the heaviest date. The boy she has a crush on oinks at her until she cries and leaves.

The second scenario (*War Crime*) was inspired by the real story of Eddie Gallagher, a navy SEAL who was accused and partially convicted of war crimes:

> A soldier in an enemy army is receiving medical treatment after surrendering. A marine sees this and kills the prisoner of war with a knife, an action which is internationally recognized as a war crime.

[1]In fact, as reported in additional analyses in the Supplementary Data file, the key results reported below all held regardless of whether the data were raw or log-transformed, for all three measures.

2.1.4 Design and Analyses. Each experiment tested the scenarios described in the previous section in two conditions: a between-subjects condition (in which independent groups of subjects completed either the Prank and War Crime scenarios) and a within-subjects condition (in which the same group of subjects completed both the Prank and the War Crime scenarios). For Experiments 1a-1c, subjects in the within-subjects condition saw both scenarios concurrently on the same page, presented in a random order. Since Experiments 2a-2c and 3a-3c paired each scenario with its own specific contrast stimulus (presented first), subjects in the within-subjects condition always saw the Prank scenario presented above the War Crime scenario (still all on the same page). During recruitment for each experiment, subjects were assigned sequentially to each of these three possibilities (Prank only, War Crime only, Both).

Per our preregistration, all data were primarily analyzed using t-tests for the Prank vs. War Crime scenarios (paired samples for within-subjects data, independent samples for between-subjects data), with additional interactions as described below. Per our preregistered hypotheses, we predicted that the War Crime scenario would be rated as more immoral than the Prank scenario for all three measures in the Within-subjects condition, but only for magnitude estimation (and *not* for Likert and Visual Analog Scales) in the Between-subjects condition.

2.2 Methods Specific to Experiments 1a-1c

Separate groups of 300 subjects completed each of the three experiments (with a single subject excluded with replacement from Experiment 1c) — 100 subjects each for the Prank scenario, the War Crime scenario, and both. This preregistered sample size was chosen to be roughly in line with past experiments in this domain (e.g. Johnson & Ahn, 2021; Kneer & Machery, 2019).

2.3 Results and Discussion

The mean immorality ratings for each scenario and condition are depicted in Figures 1a-1c (for the three measures respectively). The key comparisons in this study always involved a prank (green bar) vs. a war crime (orange bar) — with both common sense and ethical

considerations suggesting that the latter should be rated as more immoral than the former (i.e. that the orange bar should be higher than the green bar). Inspection of these figures suggests three primary patterns: First, this sensible result (with orange higher than green) was obtained for all three measures in the within-subjects condition. Second, this sensible result was also obtained for the between-subjects condition with magnitude estimation. But third, this sensible result was *not* obtained for the between-subjects conditions with either the Likert scale or the Visual Analog Scale.

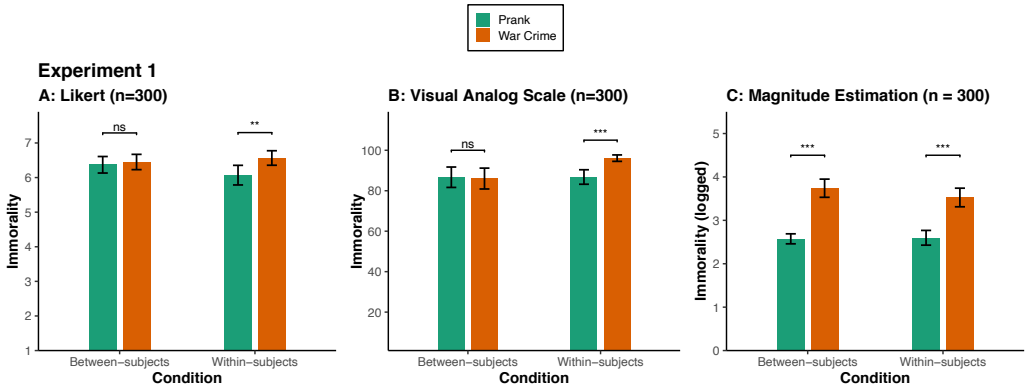

Figure 1. Mean ratings of the *Prank* and *War Crime* scenarios for (a) the Likert scale, (b) Visual Analog Scale, and (c) magnitude estimation. Error bars represent 95% confidence intervals. *** indicates differences of $p < .001$. ** indicates differences of $p < .01$. *ns* indicates non-significant differences.

These impressions were all verified by the t-tests reported in Table 1 — which for the between-subjects condition indicate a reliable difference for magnitude estimation, but null effects for the Likert scale and Visual Analog Scale (and of course with reliable differences for all three measures for the within-subjects comparisons).[2] That these measures actually differed from each other is also clear from Figures 1a-1c, but to verify this we also conducted a

[2]In our preregistered analysis plan, we also noted that we would also compute primary t-tests on only that subset of the subjects who (1) actually adjusted the scale (as opposed to simply clicking the 'next' button without interacting with the scale at all), and (2) took at least 1 second to register their response — but in fact there were only three subjects who did not meet these criteria, each of which was already excluded on the basis of the other preregistered exclusion criteria.

supplementary analysis in which we first z-scored the between-subjects ratings in each experiment (to make them directly comparable) and then computed two mixed ANOVA interactions between scenario (Prank vs. War Crime) and experiment (1a vs. 1c, and 1b vs. 1c); this confirmed that the between-subjects difference with magnitude estimation differed from the null effect for both the Likert scale ($F(1, 396)=33.44$, $p<.001$, $\eta^2=.08$) and the Visual Analog Scale ($F(1, 396)=39.66$, $p<.001$, $\eta^2=.09$). (And for completeness, we also computed the mixed ANOVA interactions between scenario [Prank vs. War Crime] and condition [Within-subjects vs. Between-subjects] for each measure. This interaction was reliable for the Visual Analog Scale [$F(1, 396)=5.82$, $p=.016$, $\eta^2=.01$], but not for either the Likert scale [$F(1, 395)=2.95$, $p=.09$] or magnitude estimation [$F(1, 396)=1.69$, $p=.20$].) These results provide an initial suggestion of the problematic nature of the labeled scales in this domain, along with the advantages of magnitude estimation.

Table 1. Means and t-test results for Experiments 1–3

				Means					
Experiment	Contrast	Measure	Design	Prank	War Crime	*t*	*df*	*p*	*d*
1a	None	Likert	Between	6.37	6.45	0.49	196.8	.63	
			Within	6.03	6.57	3.17	98	.002	0.32
1b	None	VAS	Between	86.68	86.02	0.18	197.9	.86	
			Within	86.79	96.11	5.20	99	<.001	0.52
1c	None	ME	Between	13.12	42.11	9.68	153.9	<.001	1.37
			Within	13.44	34.01	9.70	99	<.001	0.97
2a	Moderate	Likert	Between	6.48	6.51	0.20	197.9	.84	
			Within	6.09	6.41	2.44	99	.016	0.24
2b	Moderate	VAS	Between	89.42	87.07	0.70	190.2	.48	
			Within	87.24	93.11	3.78	99	<.001	0.38
2c	Moderate	ME	Between	10.80	41.21	11.81	140.9	<.001	1.67
			Within	12.38	26.26	7.69	99	<.001	0.77
3a	Extreme	Likert	Between	6.41	6.25	1.34	330.2	.18	
			Within	6.01	6.34	3.17	174	.002	0.24
3b	Extreme	VAS	Between	91.55	91.44	0.06	318.0	.95	
			Within	86.03	87.92	0.97	173	.33	
3c	Extreme	ME	Between	10.33	35.57	5.96	72.0	<.001	1.19
			Within	14.08	30.58	4.36	49	<.001	0.62

Note. We report the mean immorality ratings for the Prank and War Crime scenarios under their respective columns for the Likert scale, Visual Analog Scale (VAS), and magnitude estimation (ME) —

with raw ratings for Experiments 1a, 1b, 2a, 2b, 3a, and 3b, and geometric means for Experiments 1c, 2c, and 3c (to make the log-transformed values comparable). Paired t-tests were conducted for within-subjects comparisons, while Welch's Two Sample t-tests were conducted for between-subjects comparisons. The contrast column indicates whether a contrast scenario was provided, and if so how extreme. For Experiments 2a – 2c, the moderate contrast scenarios were sending spam mail (for the *Prank* scenario) and arson resulting in the death of a family (for the *War Crime* scenario). For Experiments 3a – 3c, the extreme contrast scenarios were jaywalking (for the *Prank* scenario) and bombing a preschool (for the *War Crime* scenario).

3. Experiments 2a – 2c: Cruel Prank vs. War Crime (Explicit Contrast Stimuli)

In Experiments 1a-1c, the relative nature of the ratings obtained with labeled scales was made clear by using scenarios that themselves implicitly triggered different contexts (everyday pranks vs. war crimes), but it is also possible to make such contexts even more explicit, by having subjects rate additional vignettes (before the key scenarios) which "set the stage". Here, we replicated Experiment 1 while using an especially mild explicit contrast for the Prank scenario (signing your boss up for spam mail), and an especially extreme explicit contrast for the War Crime scenario (multiple murder).

3.1 Method

These experiments were identical to Experiments 1a-1c (testing independent groups of subjects), with the only difference being an addition of two new explicit contrast stimuli which always preceded the Prank or War Crime scenario. We paired the Prank scenario with the following mild contrast ('Spam'), and the War Crime scenario with the following extreme contrast ('Arson'):

> *Spam:* After receiving a bad performance review at work, an employee signs their boss up to receive more junk mail.
>
> *Arson:* After receiving a bad performance review at work, an employee sets the boss's house on fire late at night, killing the boss along with the boss's spouse and three young children.

3.2 Results and Discussion

The mean immorality ratings for each scenario and condition are depicted in Figures 2a-2c (for the three measures respectively) — again with the key comparisons involving a prank

(green bar) vs. a war crime (orange bar), and now with the added ratings of the two explicit contrast stimuli also included as the faded bars for their respective scenarios. Inspection of these figures suggests that the key results fully replicated the pattern observed in Experiments 1a-1c — with the War Crime being rated as more immoral than the Prank in all Within-subjects conditions, but only for magnitude estimation when tested Between-subjects.

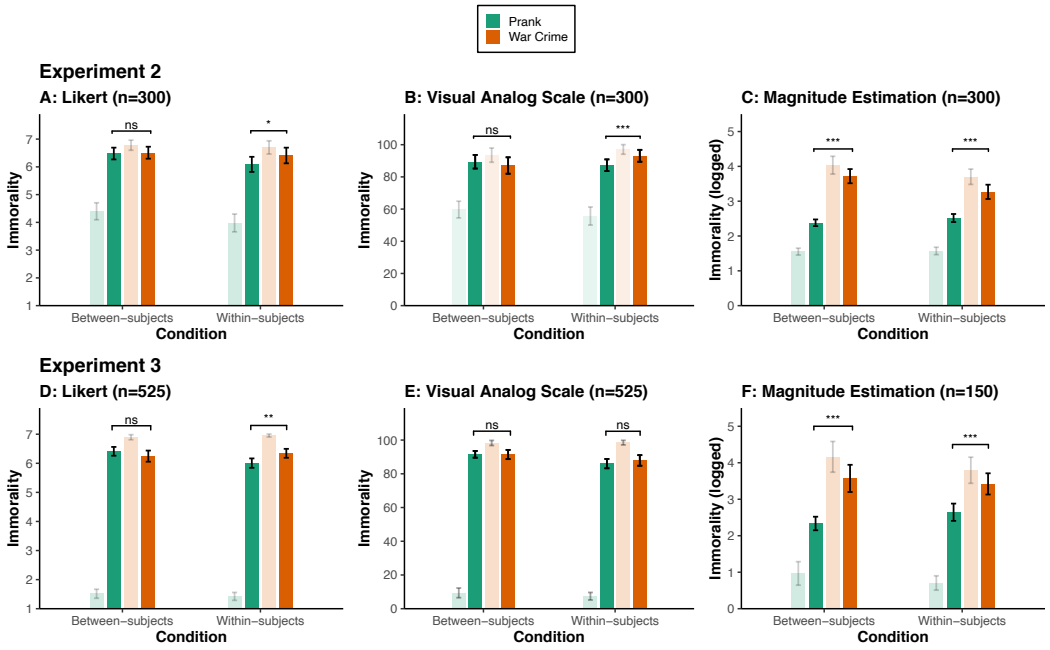

Figure 2. Mean ratings of the *Prank* and *War Crime* scenarios (solid bars) for each of the three measures in both Experiment 2 (a-c) and Experiment 3 (d-f). Explicit contrast scenarios are presented in faded bars, with the *Prank* scenario contrasted with spam mail or jaywalking and the *War Crime* scenario contrasted with arson or bombing a preschool (for Experiments 2 and 3, respectively). Error bars represent 95% confidence intervals. *** indicates differences of $p < .001$. ** indicates differences of $p < .01$. * indicates differences of $p < .05$. *ns* indicates non-significant differences.

These impressions were all verified by the t-tests reported in Table 1 — which for the Between-subjects condition again indicated a reliable difference for magnitude estimation, but null effects for the Likert scale and Visual Analog Scale (and again of course with reliable differences for all three measures for the Within-subjects comparisons). That these measures actually differed from each other is also again apparent from Figures 2a-2c, but to verify this we again conducted a supplementary analysis in which we first z-scored the between-subjects

ratings in each experiment and then computed two mixed ANOVA interactions between scenario (Prank vs. War Crime) and experiment (2a vs. 2c, and 2b vs. 2c); this confirmed that the between-subjects difference with magnitude estimation differed from the null effect for both the Likert scale ($F(1, 396)=49.33$, $p<.001$, $\eta^2=.11$) and the Visual Analog Scale ($F(1, 395)=59.87$, $p<.001$, $\eta^2=.13$) . (And for completeness, we also again computed the mixed ANOVA interactions between scenario [Prank vs. War Crime] and condition [Within-subjects vs. Between-subjects] for each measure. This interaction was not reliable for the Likert scale [$F(1, 396)=1.35$, $p=.25$] or the Visual Analog Scale [$F(1, 395)=3.76$, $p=.053$], but was reliable for magnitude estimation [$F(1, 396)=12.80$, $p<.001$, $\eta^2<.03$].[3])

These results thus provide a conceptual replication of Experiment 1 (further reinforcing the problematic nature of the labeled scales in this domain), along with even stronger evidence for the advantages of magnitude estimation (which still produced sensible results even with the additional explicit contrast stimuli).

4. Experiments 3a – 3c: Cruel Prank vs. War Crime (Extreme Contrast Stimuli)

The next three studies replicated Experiments 2a-2c in an even larger sample, while also using explicit contrast stimuli that were even more extreme (to our knowledge providing a wider range of immorality than used in almost any past study) — jaywalking in a residential neighborhood (as an explicit contrast for the Prank scenario), and terrorism targeting a preschool (as an explicit contrast for the War Crime scenario).

4.1 Method

These experiments were identical to Experiments 2a-2c, with the only differences being (a) larger sample sizes, (b) a tweak in the wording of the War Crime scenario, and (c) the

[3]Though unexpected, this interaction disappeared when the data were reanalyzed without excluding two subjects who provided responses greater than 10,000; $F(1, 399)=0.35$, $p=.55$, $\eta^2=.003$.

addition of two even more extreme explicit contrast stimuli. We recruited a preregistered 525 subjects for each of Experiments 3a and 3b, as a power analysis conducted using g*Power software (Faul et al., 2007) found that this was sufficient to detect the smallest within-subjects effect obtained in Experiments 1a or 1b (d=.32) with 95% power and an alpha of .05. We recruited a preregistered 150 subjects for Experiment 3c, as a power analysis conducted using g*Power software found that this was sufficient to detect the within-subjects effect obtained for magnitude estimation in Experiment 1c (d=.97) with 95% power and an alpha of .05. We also tweaked the wording of the War Crime scenario to address possible ceiling effects, and we paired the Prank scenario with an even milder mild contrast ('Jaywalk'), and the War Crime scenario with an even more extreme contrast ('Terrorism'):

> *War Crime (Experiment 3c)*: A soldier in an enemy army is receiving medical treatment after surrendering. A marine sees this and kills the enemy soldier with a knife.
>
> *Jaywalk*: A teenager jaywalks across a quiet, residential street.
>
> *Terrorism*: A member of a radical paramilitary group detonates a bomb at a preschool, targeting the two young children of a political opponent. All fifteen children are killed in the explosion, as well as their 23-year-old teacher.

4.2 Results and Discussion

The mean immorality ratings for each scenario and condition are depicted in Figures 2d-2f (for the three measures respectively) — again with the key comparisons involving a prank (green bar) vs. a war crime (orange bar), and with the added ratings of the two explicit contrast stimuli also included as the faded bars for their respective scenarios. Inspection of these figures suggests that for the Between-subjects condition, the key results fully replicated the pattern observed in Experiments 1a-1c and 2a-2c — with the War Crime being rated as more immoral than the Prank only for magnitude estimation. For the within-subjects condition, however, these figures suggest that the sensible pattern (with the War Crime rated as more immoral than the Prank) was still obtained for the Likert scale and magnitude estimation, but now (in contrast to Experiments 1b and 2b) not for the Visual Analog Scale.

These impressions were all verified by the t-tests reported in Table 3. That these measures actually differed from each other is also again apparent from Figures 2d-2f, but to verify this we again conducted a supplementary analysis in which we first z-scored the between-subjects ratings in each experiment and then computed two mixed ANOVA interactions between scenario (Prank vs. War Crime) and experiment (3a vs. 3c, and 3b vs. 3c); this confirmed that the between-subjects difference with magnitude estimation differed from the null effect for both the Likert scale ($F(1, 446)=28.29$, $p<.001$, $\eta^2=.06$) and the Visual Analog Scale ($F(1, 446)=21.98$, $p<.001$, $\eta^2=.05$) . (And for completeness, we also computed the mixed ANOVA interactions between scenario [Prank vs. War Crime] and condition [Within-subjects vs. Between-subjects] for each measure. This interaction was reliable for the Likert scale [$F(1, 696)=9.00$, $p=.003$, $\eta^2=.01$] but not for the Visual Analog Scale [$F(1, 695)=0.53$, $p=.47$] or magnitude estimation [$F(1, 196)=2.72$, $p=.10$]). When looking at the p-values in Table 3, one might be initially worried that the difference between the two scenarios for the Likert scale is also trending to some degree, but note that these values are in the *opposite* direction from the sensible pattern — with the Prank now being rated numerically as even worse than the War Crime. (6 of these 350 subjects gave nonsensical responses even for the explicit contrast stimuli vs. the relevant condition: 3 of 175 subjects rated jaywalking as more immoral than the Prank, and 3 of 175 subjects rated the preschool bombing as less immoral than the War Crime. We note in passing that if these 6 subjects are excluded, then the remaining 344 subjects' Likert scale data actually demonstrate a reliable effect in the opposite direction, with the Prank being rated worse than the War Crime in a Welch Two Sample t-test; $t(286.5)=2.04$, $p=.042$, $d=.22$).

These results, while providing another conceptual replication of the primary effects, show that the problematic nature of labeled scales — and the advantages of magnitude estimation — hold even in the face of the explicit moral contrasts that are especially extreme.

5. Experiments 4a and 4b: More Luck for 'Moral Luck'

The previous experiments illustrate the dangers of labeled scales — and the corresponding advantages of magnitude estimation — in the context of theoretical examples developed for the current project. But as noted above, such scales are ubiquitous in the actual experimental literature on moral psychology. So might such problems (and their solutions) also occur in practice, for an actual published result? We next present a case study of exactly this type, drawn from a prominent paper published in this same journal which focused on the so-called problem of 'moral luck' (first coined by Williams, 1981). This problem concerns an obvious tension between two widely-held commonsense views relating to moral responsibility. First, we *are* responsible for the consequences of our actions; second, we are *not* responsible for things over which we have no control. Consider the following contrast: (a) a truck driver who skids out while driving recklessly in heavy rain, hitting a tree; vs. (b) a truck driver who takes the same actions but who happens to do so at the exact moment an unseen child is crossing the street in front of that tree. On one hand, people are inclined to judge the second driver more harshly. (A child was injured, or worse!) On the other hand, the two drivers took exactly the same actions, with exactly the same knowledge and degree of control (with neither expecting a child).

This puzzle of moral luck has been studied extensively, especially in philosophy (for an overview, see Zimmerman, 2006). But a recent study suggested that this problem is in fact illusory: people do not make differential judgments in the two cases, and the apparent fact that they do is driven simply by the use of methods (in particular, between-subjects designs) which discourage careful thought (Kneer & Machery, 2019). Instead, presenting such cases *together* (in a within-subjects design) is more "favorable to reflective deliberation" (p. 331) — leading such differences to disappear.

We were attracted to this particular result for two related reasons. First, it has been highly cited (with more than 130 citations in just five years), and was published in this same journal. Second, we found the conclusion to be somewhat incredible: we, at least, do not need any subtle statistical analyses to reveal the tension at the heart of the puzzle of moral luck, since we can viscerally *feel* it ourselves when considering such cases. We thus suspect that the relevant effects disappear when tested within-subjects simply because this highlights the brute similarity of the relevant cases, which provides an extrinsic task demand to answer each case similarly (see Charness et al., 2012; Orne, 1962).

Regardless of whether this intuition is correct, this result should seem suspect in the present context — since it is precisely a case where an incredible outcome results from the use of Likert scales. Accordingly, a more careful test of this hypothesis would simply use magnitude estimation to ask the same question, and the current experiments tested this directly. First, in Experiment 4a, we simply replicated the original result using a Likert scale (Kneer & Machery, 2019). Then, in Experiment 4b, we employed magnitude estimation to test the two cases — still within-subjects, but now using a different benchmark stimulus for each case. If the disappearance of the key contrast is due to the heightened likelihood of "reflective deliberation", then Experiment 4b should effectively replicate Experiment 4a — since the within-subjects design still allows for that. But if the disappearance was simply due to a task demand, then the empirical contrast at the heart of the puzzle of moral luck should reappear in Experiment 4b, despite the within-subjects test.

5.1 Method

Experiments 4a and 4b were identical except for the measure used, as described below. All hypotheses, analyses, and sample sizes were preregistered, and the Supplementary Data file provides all raw data and links to preregistrations.

5.1.1 Subjects. Separate groups of 300 subjects completed each of the two experiments: in each case, 100 subjects were tested on (only) the Lucky Outcome scenario (as part of the

between-subjects test), 100 subjects were tested on (only) the Unlucky Outcome scenario (as part of the same between-subjects test), and 100 subjects were tested on both (in a within-subjects design). This preregistered sample size was chosen to be roughly in line with that used by Kneer and Machery (2019) — who e.g. tested 95 subjects in their (within-subjects) Experiment 1b. Per our preregistered criteria, we excluded and replaced any subject who did not report that killing an innocent stranger for fun is worse than stealing a wallet (leading to 40 exclusions from Experiment 4b).

5.1.2 Stimuli. Subjects read the same vignettes used in Experiments 1a and 1b from Kneer and Machery (2019):

> *Unlucky Outcome*: Anna is at home, giving her 2-year-old son a bath. She fills the bath, while her son stands near the tub. The phone rings in the next room. Anna tells her son to stand near the tub while she answers the phone. Anna believes her son will stand near the tub for a few minutes and wait for her to return. Anna leaves the room for 5 min. When Anna returns, her son is in the tub, dead, face down in the water.
>
> *Lucky Outcome*: Beth is at home, giving her 2-year-old son a bath. She fills the bath, while her son stands near the tub. The phone rings in the next room. Beth tells her son to stand near the tub while she answers the phone. Beth believes her son will stand near the tub for a few minutes and wait for her to return. Beth leaves the room for 5 min. When Beth returns, her son is still standing near the tub, where she left him. The boy then enjoys his bath.

5.1.3 Measures. In the original experiment, subjects provided several different kinds of moral judgment — including wrongness, blameworthiness, permissibility, and how much punishment is deserved. To simplify our replication, we considered only judgments of wrongness, measured as follows. Subjects in Experiment 4a made their judgments using a 7-point Likert scale, ranging from 1 (“not at all morally wrong”) to 7 (“very morally wrong”), in response to the question “How wrong was Anna/Beth to leave her son alone in the above scenario?”. Subjects in Experiment 4b made their judgments in reference to a benchmark stimulus which was unique to each scenario:

> *Unlucky Outcome:* If killing an innocent stranger for fun is a 100, how morally wrong was Anna to leave her son alone in the above scenario? (A 50 would mean "half as immoral," 200 would mean “twice as immoral,” etc. You can use any number you'd like in your response, including fractions, decimals, and numbers over 100).

> *Lucky Outcome*: If stealing a wallet is a 10, how morally wrong was Beth to leave her son alone in the above scenario? (A 5 would mean "half as immoral," 20 would mean “twice as immoral,” etc. You can use any number you'd like in your response, including fractions, decimals, and numbers over 10).

Magnitude estimates were transformed following the standard procedure previously described. At the end of the experiment, subjects in the within-subjects condition of Experiment 4b rated the two benchmark stimuli in reference to one another, allowing direct comparison, while avoiding any possible task demands to provide identical ratings to the (nearly identical) scenarios:

> *Wallet-to-Murder Conversion*: If killing an innocent stranger for fun is a 100, how morally wrong is it to steal a wallet? (Remember: a 50 would mean “half as immoral,” 200 would mean “twice as immoral,” etc. You can use any number you'd like in your response, including fractions, decimals, and numbers over 100).

5.1.4 Design and Analyses. Each experiment tested the scenarios described in the previous section in two conditions: a between-subjects condition (in which independent groups of subjects completed the *Unlucky Outcome* and the *Lucky Outcome* scenarios) and a within-subjects condition (in which the same group of subjects completed both scenarios). (Following the original experiment, subjects in the within-subjects condition always saw the *Unlucky Outcome* scenario presented above the *Lucky Outcome* scenario and on the same page.)

To analyze the results, we conducted the same analyses reported in Studies 1a and 1b of Kneer and Machery (2019), with the only addition being a preregistered series of 2 (Scenario: Unlucky Outcome vs. Lucky Outcome) x 2 (Experimental Design: Between-Subjects vs. Within-Subjects) mixed-design ANOVAs. To mirror the analysis conducted by Kneer and Machery, which compared the number of within-subjects judgments that were identical across the *Unlucky Outcome* and *Lucky Outcome* scenarios, we simply converted the magnitude estimates from Experiment 4b (in which each scenario was judged in reference to a different benchmark stimulus) to a common scale, by dividing the raw ratings for the *Lucky Outcome* scenario by 10

(the value assigned to the wallet benchmark), with the result then multiplied by the rating provided by each subject for the Wallet-to-Murder Conversion.

5.2 Results

The mean moral wrongness ratings for each scenario and experimental design are depicted in Figures 3a-3b, with the subject-wise difference between the two scenarios in the within-subjects design depicted in Figures 3c-3d (for Experiments 4a and 4b, respectively). Here, the key comparisons involve two scenarios depicting identical behaviors (leaving a toddler unattended in a bath) which nonetheless lead to a morally lucky outcome (the toddler is fine; gray bar) vs. a morally unlucky one (the toddler drowns; white bar).

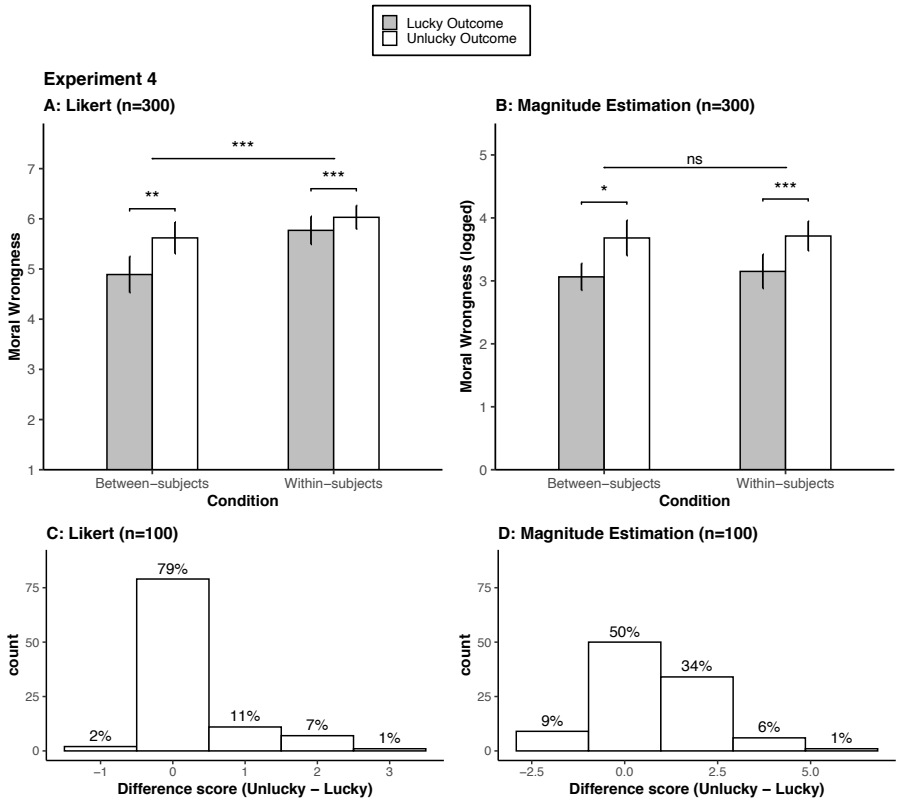

Figure 3. Mean ratings from the *Lucky Outcome* and *Unlucky Outcome* conditions grouped by experimental design (Between- vs. Within-subjects), for both (a) Likert ratings and (b) magnitude estimates. Paired t-tests were conducted for Within-subjects comparisons, and Welch's Two Sample t-tests were conducted for Between-subjects comparisons. Error bars represent 95% confidence intervals. The within-subjects differences between the two scenarios are plotted for (c) Likert ratings and (d) magnitude estimates (which were binned into the same number of discrete categories as present in the Likert data). *** indicates differences of $p < .001$. ** indicates differences of $p < .01$. *ns* indicates non-significant differences.

Inspection of these figures suggests two main conclusions. First, the Likert ratings (from Experiment 4a) qualitatively replicated the effects first reported by Kneer and Machery (2019): judgments collected between subjects reflected the problem of moral luck (in that the Unlucky Outcome was judged more harshly than the Lucky Outcome), but this affect was attenuated in the within-subjects condition (Figure 3a), with a large majority of subjects providing numerically identical responses to the two scenarios (Figure 3c). Second, magnitude estimates collected using independent benchmarks in a within-subjects design (from Experiment 4b) were virtually indistinguishable from magnitude estimates collected using those same benchmarks in a between-subjects design (Figure 3b) — thus reflecting the problem of moral luck in *both* experimental designs. Furthermore, individual subjects were much less likely to provide numerically identical responses (Figure 3d).

These impressions were all verified by two primary analyses. First, as can be seen from the mixed ANOVAs reported in Table 2, the Likert ratings collected in Experiment 4a revealed a main effect of scenario (such that an action leading to an *Unlucky Outcome* was judged to be morally worse than that same action leading to a *Lucky Outcome*) as well as a main effect of experimental design (such that the scenarios were judged to be more morally wrong when ratings were collected within-subjects). In contrast, the magnitude estimates collected in Experiment 4b revealed only a main effect of scenario, and nothing else.[4] Second, as can be seen from the post-hoc tests reported in Table 3, the difference found for the *Lucky Outcome* scenario in Experiment 4a (in which Likert ratings collected within-subjects were more severe

[4]We failed to find any Design x Scenario interaction for the Likert ratings collected in Experiment 4a, as predicted in our preregistration. However, this may stem from presence of a main effect of experimental design (which was absent in the findings reported in Studies 1a and 1b of Kneer and Machery, 2019). Nonetheless, an exploratory mixed effects model which included subjects as a random intercept revealed (using Satterthwaite's method to calculate p-values) a significant interaction between experimental design and scenario, $F(1,351)=4.08$, $p=0.044$. This same exploratory analysis was not significant for the magnitude estimates collected in Experiment 4b, $F(1,370.0)=0.14$, $p=0.71$.

than Likert ratings collected between-subjects) was entirely absent from the magnitude estimates collected in Experiment 4b — in which the two experimental designs produced qualitatively identical and statistically indistinguishable results.

Table 2. Mixed ANOVA results for Experiments 4a–4b

Experiment	Measure	Predictor	df	Sum of Squares	*F*	*p*	η^2
4a	Likert	Design	1	41.6	18.1	<.001	0.04
		Scenario	1	24.5	10.6	.001	0.03
		Design x Scenario	1	5.5	2.4	.12	
		Residuals	396	912.0			
4b	ME	Design	1	0.2	0.1	.75	
		Scenario	1	23.8	12.2	.001	0.03
		Design x Scenario	1	0.1	0.1	.82	
		Residuals	396	776.7			

Note. Since subjects in Experiment 4b rated the *Lucky Outcome* scenario and the *Unlucky Outcome* scenario in reference to different benchmark stimuli (stealing a wallet and murder, respectively), ratings in the *Lucky Outcome* scenario were divided by 10 (the value given for stealing a wallet) and multiplied by the average Wallet-to-Murder Conversion factor. This transformation produces qualitatively identical results, but provides an effect size that is more directly comparable to the Likert judgments made in Experiment 4a (which are naturally provided along a common scale).

Table 3. Means and t-test results for Experiments 4a–4b

Experiment	Measure	Scenario Outcome	Between-Subjects	Within-Subjects	*t*	*df*	*p*	*d*
4a	Likert	Lucky	4.89	5.77	3.83	186.2	<.001	0.54
		Unlucky	5.62	6.03	2.07	183.1	.08	0.29
4b	ME	Lucky	9.45	10.53	0.54	198.0	1	
		Unlucky	45.40	46.03	0.07	192.5	1	

Note. Mean moral wrongness ratings for both the *Lucky* and *Unlucky* scenarios are reported in their respective rows, with the between-subjects ratings and within-subjects ratings under their respective columns. The means for Experiment 4a represent the raw Likert ratings, while the means for Experiment 4b represent the geometric means. This allows each scenario to be interpretable in reference to its respective benchmark stimulus: stealing a wallet (assigned the value 10) and killing an innocent stranger for fun (assigned the value 100) for the Lucky and Unlucky scenarios, respectively. Welch's Two Sample t-tests were conducted for all comparisons, which are Bonferroni-corrected within each experiment. The uncorrected p-values for the Lucky and Unlucky scenarios in Experiment 4b are .59 and .94, respectively.

5.3 Discussion

These results again illustrate the advantage of magnitude estimation over simple labeled scales. And critically, whereas the demonstrations in Experiments 1-3 did so using hypothetical examples created specifically for the present study, the current experiments did so in the context of a prominent recent result from the extant moral psychology literature. Just as with Experiments 1-3, Likert scales produced unlikely or nonsensical patterns of results, which then disappeared when using magnitude estimation. This result was not pre-ordained, since magnitude estimation can readily reveal within-subjects differences when they truly exist — as indeed was true for Experiments 1-3. (In fact, Likert scales have notorious problems in which they may sometimes *fail* to reveal real differences for other reasons, which become salient when measured in other ways; e.g. Bartoshuk et al., 2003.) In fact, as a proof of concept with our exact implementation of magnitude estimation (including the two-benchmark procedure), we showed that it can readily reveal a genuine within-subjects difference in a separate test of moral judgment.[5] This case study illustrates the theoretical importance of such measurement challenges and their solutions, since here the different types of measurements fueled entirely opposing conclusions: Likert scale tests suggested that the well-known puzzle of moral luck does not exist in within-subject designs; but the more careful, anchored magnitude estimation procedure resuscitated the standard contrast — suggesting that the puzzle of moral luck is alive

[5]In particular, we replicated the primary finding from Lombrozo (2009), using the infamous "switch" and "push" variations of the trolley problem (in which one life is sacrificed to save five lives, either by flipping a switch to divert a trolley to a different track, or by pushing a man in front of a trolley, respectively). We recruited 300 subjects from Prolific (n=290 after the same exclusions described above). A 2x2 mixed ANOVA revealed a significant Design x Scenario interaction ($F(1, 397)=9.66$, $p=.002$, $\eta 2=.02$), with Bonferroni-corrected post hoc tests revealing that the "push" scenario was rated as morally better when judgments were made within-subjects (i.e. jointly presented with the "switch" scenario; $M=53.89$), compared to when the same judgment was made in isolation, between-subjects ($M=82.21$; with both geometric means relative to the benchmark of killing an innocent stranger for fun, which was assigned the value 100; $t(191.7)=2.28$, $p=.048$, $d=0.30$). In contrast the "switch" scenario was not rated as significantly morally worse when judgments were made within-subjects (following the jointly-presented "push" scenario, $M=9.96$) compared to when the same judgment was made in isolation, between-subjects ($M=6.38$; with both geometric means relative to the benchmark of stealing a wallet, which was assigned the value 10; Welch's two-sample $t(187.7)=2.12$, $p=.07$, $d=0.30$). Raw data can be found in the Supplementary Data File.

and well, even in within-subjects designs. This shows how such subtle issues of measurement can have profound consequences in practice for theorizing in moral psychology.

6. General Discussion

This project explored a potentially insidious problem (and a simple and straightforward solution) that may be widely applicable in moral psychology. The problem is that the most ubiquitous type of measurement in this field — labeled (e.g. Likert) rating scales — may often be unreliable, since different people may use the same scale in different ways, in different contexts. In short, such scales may often lead us astray because they are *unanchored*. Depending on the details, this problem can manifest in terms of either (a) illusory differences (which do not really exist in a substantive way in the underlying population), or (b) illusory null effects (when real underlying differences are masked). The straightforward solution, then, is to simply provide an anchor — which is what the technique of *magnitude estimation* does. Here we presented two case studies of this dynamic — one 'in principle' (with new theoretical examples), and one 'in practice' (drawn from the contemporary moral psychology literature).

6.1 The Problem in Principle (Cruel Pranks)

Our first case study employed novel materials to demonstrate the problem and its solution in a context in which the use of Likert scales can give rise to illusory differences. It contrasted the immorality of a prank (resulting in humiliation) with the immorality of a war crime (resulting in murder), with the assumption that it would be a desirable quality for our measures of moral judgment to indicate that the latter is obviously morally worse than the former.[6] This is precisely what was found by the psychophysically-inspired method of

[6]Must this always be true? We suppose that the devil is in the details: in some odd or rare circumstances, the prankster might indeed be more morally worse than the killer — e.g. if it was clear that the prankster was acting out of pure sadism, with great forethought, while the soldier was still in the 'fog of war', etc. Recall that our actual vignette attempted to minimize such possibilities, e.g. by specifying explicitly that

magnitude estimation (involving an explicit benchmark stimulus, as in Experiments 1c, 2c, and 3c). Indeed, the results of these experiments were entirely pedestrian, and even boring: the war crime was indeed rated as more immoral than the prank, just as one might hope — and this was true regardless of several sources of experimental variation (e.g. when tested both with and without an explicit contrast stimulus, and when tested both within-subjects and between-subjects).

But when this same simple contrast was tested using labeled scales, tested between-subjects, everything went haywire (with the trolley going off the rails, so to speak). In six separate experiments (with both Likert scales as in Experiments 1a, 2a, and 3a; and with Visual Analog Scales as in Experiments 1b, 2b, and 3b) we consistently obtained results in which the prank was rated as just as immoral as (or even worse than) the war crime. This nonsensical result was especially striking in the current project, since the difference that these measures failed to capture was anything but subtle: when tested between-subjects with magnitude estimation, the war crime wasn't just rated as more immoral, but as *definitively* more immoral — with Cohen's *d* effect sizes of 1.37, 1.67, and 1.19, for Experiments 1c, 2c, and 3c, respectively. Effects of this size are on par with the strongest effects in psychology (e.g. Table 1 from Lovakov & Agadullina, 2021), making it especially concerning when common methods cannot capture them.

While these common methods did, unsurprisingly, capture the central intuitive result when tested within-subjects, we still do not recommend the use of labeled scales for within-subjects designs. Even though such measures could detect this obvious difference, the effects were nonetheless much weaker with labeled scales, in at least three ways. First, the average effect sizes for the within-subjects effects when tested with labeled scales [Likert Scale: d = .27; Visual Analog Scale: d = .32] were notably smaller than those with magnitude estimation [d =

---

the relevant murder occurred after the victim's surrender, and while they were receiving medical treatment (so that they clearly didn't pose any current threat, etc.).

.79]. Second, this contrast was not statistically reliable for the Visual Analog Scale in Experiment 3b. And third, as depicted in the nonparametric within-subjects data from Figure 4: while the vast majority of individual subjects [on average 84%] rated the war crime as worse than the prank when tested with magnitude estimation, far fewer did when tested with the Likert scale [37%] or Visual Analog Scale [47%].

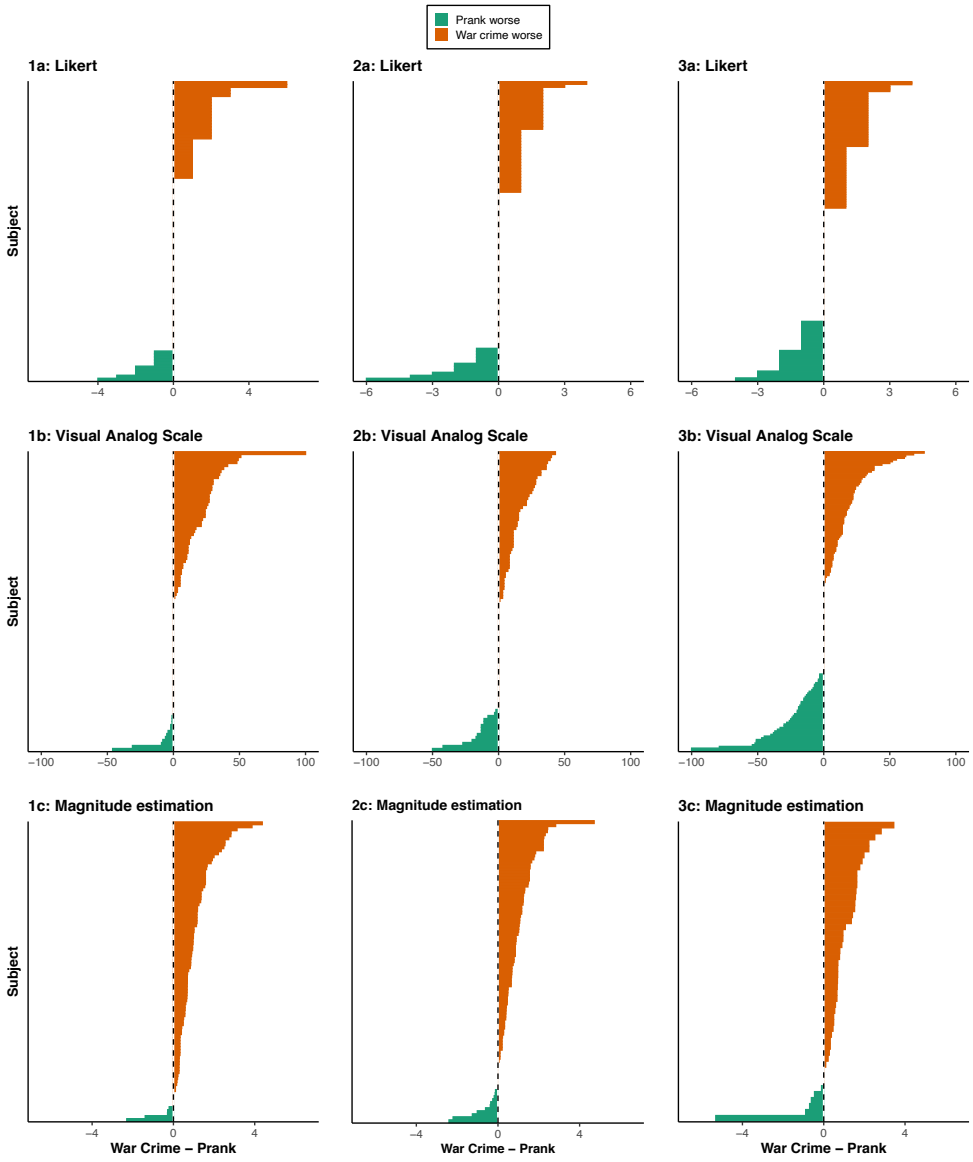

Figure 4: Within-subject difference between *War Crime* and *Prank* scenarios. Each bar represents an individual subject. Positive bars (in orange) represent subjects who rated the War Crime to be more immoral than the Prank. Negative bars (in green) represent subjects who rated the Prank to be more immoral than the War Crime.

6.2 The Problem in Practice (Moral Luck)

The 'cruel prank' experiments constitute a 'toy' example in which the use of Likert scales can yield misleading (and deeply implausible) conclusions. But we think such problems may exist not only in principle, but also in practice. Accordingly, our second case study used a prominent example from the contemporary moral psychology literature to demonstrate the problem and its solution in a context in which the use of Likert scales can yield illusory null effects (effectively masking real underlying differences).

The recent study in question attempted to deny the psychological reality of the puzzle of 'moral luck' (Kneer & Machery, 2019). This seemed ambitious and exciting, given that nearly everyone intuitively agrees that some moral weight is carried by the outcomes of one's actions (such that accidentally hitting a child with your car is morally worse than accidentally hitting a tree). This earlier study wanted to maximize the chances that subjects could consider such cases in ways that were "favorable to reflective deliberation", and they were surely correct that providing the two cases within-subjects does exactly that. (After all, you cannot thoughtfully compare two cases if you only ever encounter one of them.) And sure enough, these authors then found that the contrast at the core of moral luck was attenuated (if not entirely eliminated): now people no longer rated one case as morally worse based only on its consequences.

We have no disagreement with this result itself; and indeed, we essentially replicated it ourselves (in Experiment 4a). However, because they used Likert scales to measure moral judgment within-subjects, this also allowed for pernicious task demands to play a role (Charness et al., 2012; Orne, 1962). ("Those two vignettes are almost exactly the same, so I guess I should give them similar ratings.") And given the powerful and visceral intuitions embodied by such moral-luck cases, we suspected that the null result was due simply to such task demands. In short, we suspected that this is a case wherein the use of Likert scales allowed a real underlying difference to be masked.

Conveniently, the solution in this context is the very same one from our first case study: we simply need to test the same effect while using a method that (a) still allows for within-subjects testing (and thus plenty of reflection and deliberation), while (b) being resistant to task demands (by using a method in which subjects cannot simply give the same numerical answer twice). We did just this in Experiment 4b, and readily obtained the opposite result from the previous study (and from our own Experiment 4a): now the puzzle of moral luck reappeared, and again moral judgments were influenced by uncontrollable outcomes in the familiar way. Indeed, the results we obtained in the within-subjects condition of Experiment 4b were nearly *identical* to those obtained between-subjects.

6.3 Lessons for Moral Psychology

The central lesson of this project is thus that the most commonly used measures in moral psychology are deeply problematic, both in principle and in practice, yielding both illusory differences (as in the 'cruel prank' example) and illusory null effects (as in the 'moral luck' example). Of course, not every use of such Likert scales in moral psychology will yield such problematic results; the findings of many (or even most) such studies may emerge unscathed. But the use of such measures in essence introduces a sort of unnecessary vulnerability to such studies.

The fix for this vulnerability is to use magnitude estimation instead of labeled scales to measure moral judgment.[7] Magnitude estimation provides at least five related advantages in this context, with no appreciable disadvantages:

First, and most directly, magnitude estimation makes moral ratings meaningful on an absolute scale — since different measurements can be meaningfully related to each other. In particular, as long as there is no systematic deviation in the ratings of the (ideally non-

[7]Many moral psychology studies are conducted using the Qualtrics survey platform, which can readily accommodate the use of magnitude estimation. A template for using magnitude estimation on this platform is available at: https://vladchituc.com/s/magnitude-estimation.zip

controversial) benchmark stimulus across the different groups, one could meaningfully compare their relative moral judgments — whereas this is simply not possible with labeled scales. Indeed, the entire purpose of magnitude estimation as applied to judgment is to make such comparisons meaningful (e.g. Stevens, 1966). This could even be philosophically relevant, since positions such as *contrastivism* suggest that moral reasons (and in fact all reasons) can only ever be expressed relative to a contrast clause (Sinnott-Armstrong, 2008).

Second, magnitude estimation thus allows for meaningful comparison across different experiments (even by different research groups), as long as the studies employ comparable benchmark stimuli. This seems of potentially great value in the study of moral psychology — which currently features hundreds of individual studies, few of which can be directly compared with each other in terms of their absolute ratings (despite the fact that doing so is required by nearly all meta-analyses). If this research community instead pivoted to the use of magnitude estimation with a set of common shared benchmark stimuli, our science might become more cumulative, such that the results of different studies could be directly combined and compared. (Good benchmarks are often familiar and of moderate extremity, with minimal variance and controversy, and no intrinsic numerical content; cf. 'stealing a wallet' vs. 'having an abortion' (too controversial) vs. 'killing 20 people' (too numerical). For discussion of the sometimes-nuanced factors involved in choosing benchmark stimuli (e.g. taking care to choose benchmarks that are sufficiently different from the key test case, and choosing the appropriate anchor values to use) we refer readers to standard magnitude estimation primers such as Lodge, 1981; Moskowitz, 1977; and Stevens, 1956).

Third, as a result, magnitude estimation allows for the valid comparison of moral judgments *across different groups*, including different demographic groups whose moral judgments may be directly compared with one another — e.g. when comparing how character judgments might vary as a function of religious identification (A. B. Cohen & Rozin, 2001); or whether judgments of intentional vs. accidental harms might differ for psychopaths (Young et

al., 2012) or across cultures (Barrett et al., 2016; McNamara et al., 2019); or whether judgments involving sacrificial moral dilemmas change as a function of age (Hannikainen et al., 2018), sex (Capraro & Sippel, 2017), culture (Xu et al., 2024), or some combination of such factors (Arutyunova et al., 2016). Such claims *may* be possible, but current methods (using labeled scales) cannot rule out a mundane alternative explanation: such results may merely reflect group differences in how the scale is interpreted (rather than differences in moral judgment, *per se*).

Fourth, magnitude estimation also has direct pragmatic advantages for researchers in this domain, due to its ability to more efficiently reveal underlying patterns. This was especially apparent in Experiments 1-3 of the current project in the comparison of the effect sizes across the different measures. And in practice, this means that magnitude estimation can reveal effects using far fewer subjects than are required by studies using labeled scales — in a way that can be quantified. For example, if we aimed to conduct a replication of the Within-subjects conditions of Experiments 1a-1c with 95% power and an alpha of .05, we would need a total of 129 subjects using a Likert scale, 51 using Visual Analog Scale, but only 16 using magnitude estimation.

Finally, the use of magnitude estimation may help to methodologically integrate the study of moral psychology with those of other subfields of cognitive science. As noted above, the problems with labeled scales (and the corresponding advantages of magnitude estimation) are quite familiar in other areas, especially sensory psychophysics. In recent years, there have been several attempts to synthesize research on morality and perception (for a review see Schein et al., 2016), though it has been suggested that some such efforts go too far (e.g. Firestone & Scholl, 2016). The current project suggests that this useful crosstalk may also extend to how the methods of different fields can enrich each other — and in a sense, this would be something of a return to form. Not too long ago, psychophysical measurement and higher-level topics such as morality were studied in tandem (e.g. Ekman, 1962; Sellin & Wolfgang, 1964; for an overview, see Stevens, 1966a), and we think that both could profit from doing so again.

References

Acierno, J., Mischel, S., & Phillips, J. (2022). Moral judgements reflect default representations of possibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1866), 20210341. https://doi.org/10.1098/rstb.2021.0341

Aitken, R. C. B. (1969). A growing edge of measurement of feelings [Abridged]: Measurement of feelings using Visual Analogue Scales. *Proceedings of the Royal Society of Medicine*, *62*(10), 989–993. https://doi.org/10.1177/003591576906201005

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of Social and Personality Psychology. *Personality and Social Psychology Bulletin*, *45*(6), 842–850. https://doi.org/10.1177/0146167218798821

Arutyunova, K. R., Alexandrov, Y. I., & Hauser, M. D. (2016). Sociocultural influences on moral judgments: East–West, male–female, and young–old. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01334

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., Scelza, B. A., Stich, S., von Rueden, C., Zhao, W., & Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113*(17), 4688–4693. https://doi.org/10.1073/pnas.1522070113

Bartoshuk, L. M. (2000). Comparing sensory experiences across individuals: Recent psychophysical advances illuminate genetic variation in taste perception. *Chemical Senses*, *25*(4), 447–460. https://doi.org/10.1093/chemse/25.4.447

Bartoshuk, L. M., Duffy, V. B., Fast, K., Green, B. G., Prutkin, J., & Snyder, D. J. (2003). Labeled scales (e.g., category, Likert, VAS) and invalid across-group comparisons: What we have learned from genetic variation in taste. *Food Quality and Preference*, 14.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66*(1), 5–20. https://doi.org/10.1037/0022-3514.66.1.5

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, *60*(4), 485–499. https://doi.org/10.1037/0022-3514.60.4.485

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods*, *4*(3), 243.

Borg, G. (2001). Are we subjected to a "long-standing measurement oversight?" In R. Kompass (Ed.), *Fechner Day 2001. Proceedings of the Seventeenth Annual Meeting of the International Society for Psychophysics* (pp. 1–44).

Bridwell, N. (1963). *Clifford the Big Red Dog*. Scholastic US.

Capraro, V., & Sippel, J. (2017). Gender differences in moral judgment and the evaluation of gender-specified moral agents. *Cognitive Processing*, *18*(4), 399–405. https://doi.org/10.1007/s10339-017-0822-9

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, *11*(1), 2100. https://doi.org/10.1038/s41467-020-15602-4

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, *47*(4), 1178–1198. https://doi.org/10.3758/s13428-014-0551-2

Cohen, A. B., & Rozin, P. (2001). Religion and the morality of mentality. *Journal of Personality and Social Psychology*, *81*(4), 697–710. https://doi.org/10.1037/0022-3514.81.4.697

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359–1381. https://doi.org/10.1037/xge0000210

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*(1), 2–7. https://doi.org/10.1037/a0025071

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience*, *7*(3), 269–279. https://doi.org/10.1080/17470919.2011.614000

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089. https://doi.org/10.1111/j.1467-9280.2006.01834.x

DeScioli, P., Bruening, R., & Kurzban, R. (2011). The omission effect in moral cognition: Toward a functional explanation. *Evolution and Human Behavior*, *32*(3), 204–215. https://doi.org/10.1016/j.evolhumbehav.2011.01.003

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*(3), 285–290. https://doi.org/10.1037/h0033731

Ekman, G. (1962). Measurement of moral judgment: A comparison of scaling Methods. *Perceptual and Motor Skills*, *15*(1), 3–9. https://doi.org/10.2466/pms.1962.15.1.3

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Firestone, C., & Scholl, B. J. (2016). 'Moral perception' reflects neither morality nor perception. *Trends in Cognitive Sciences*, *20*(2), 75–76. https://doi.org/10.1016/j.tics.2015.10.006

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science*, *6*(8), 859–868. https://doi.org/10.1177/1948550615592241

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001

Hannikainen, I. R., Machery, E., & Cushman, F. A. (2018). Is utilitarian sacrifice becoming more morally permissible? *Cognition*, *170*, 95–101. https://doi.org/10.1016/j.cognition.2017.09.013

Heiphetz, L., & Cushman, F. (2021). Introduction to morality as a hub: Connections within and beyond social cognition. *Social Cognition*, *39*(1), 1–3. https://doi.org/10.1521/soco.2021.39.1.1

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *67*(3), 247–257. https://doi.org/10.1006/obhd.1996.0077

Jia, L., Shao, B., Wang, X., & Shi, Z. (2021). Phrase depicting immoral behavior dilates its subjective time judgment. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.784752

Johnson, S. G. B., & Ahn, J. (2021). Principles of moral accounting: How our intuitive moral sense balances rights and wrongs. *Cognition*, *206*, 104467. https://doi.org/10.1016/j.cognition.2020.104467

Khamitov, M., Rotman, J. D., & Piazza, J. (2016). Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition*, *146*, 33–47. https://doi.org/10.1016/j.cognition.2015.09.009

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*(2), 309–324. https://doi.org/10.1080/09515080307771

Knutson, K. M., Krueger, F., Koenigs, M., Hawley, A., Escobedo, J. R., Vasudeva, V., Adolphs, R., & Grafman, J. (2010). Behavioral norms for condensed moral vignettes. *Social Cognitive and Affective Neuroscience*, *5*(4), 378–384. https://doi.org/10.1093/scan/nsq005

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209. https://doi.org/10.1016/j.cognition.2015.01.013

Leong, L. M., McKenzie, C. R. M., Sher, S., & Müller-Trede, J. (2019). Illusory inconsistencies in judgment: Stimulus-evoked reference sets and between-subjects designs. *Psychonomic Bulletin & Review*, *26*(2), 647–653. https://doi.org/10.3758/s13423-019-01585-x

Liao, S., Strohminger, N., & Sripada, C. S. (2014). Empirically investigating imaginative resistance. *The British Journal of Aesthetics*, *54*(3), 339–355. https://doi.org/10.1093/aesthj/ayu027

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*, 55–55.

Lodge, M., & Tursky, B. (1981). On the Magnitude Scaling of Political Opinion in Survey Research. *American Journal of Political Science*, *25*(2), 376. https://doi.org/10.2307/2110859

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, *33*(2), 273–286. https://doi.org/10.1111/j.1551-6709.2009.01013.x

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*(1), 293–318.

McDonald, K., Graves, R., Yin, S., Weese, T., & Sinnott-Armstrong, W. (2021). Valence framing effects on moral judgments: A meta-analysis. *Cognition*, *212*, 104703. https://doi.org/10.1016/j.cognition.2021.104703

McNamara, R. A., Willard, A. K., Norenzayan, A., & Henrich, J. (2019). Weighing outcome vs. intent across societies: How cultural models of mind shape moral reasoning. *Cognition*, *182*, 95–108. https://doi.org/10.1016/j.cognition.2018.09.008

Melnikoff, D. E., & Bailey, A. H. (2018). Preferences for moral vs. immoral traits in others are conditional. *Proceedings of the National Academy of Sciences*, *115*(4), E592–E600. https://doi.org/10.1073/pnas.1714945115

Minson, J. A., & Monin, B. (2012). Do-gooder derogation: Disparaging morally motivated minorities to defuse anticipated reproach. *Social Psychological and Personality Science*, *3*(2), 200–207. https://doi.org/10.1177/1948550611415695

Moskowitz, H. R. (1977). Magnitude estimation: Notes on what, how, when, and why to use it. *Journal of Food Quality*, *1*(3), 195–227. https://doi.org/10.1111/j.1745-4557.1977.tb00942.x

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783. https://doi.org/10.1037/h0043424

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

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*(6), 407–418. https://doi.org/10.1037/h0022602

Pizarro, D. A., Laney, C., Morris, E. K., & Loftus, E. F. (2006). Ripple effects in memory: Judgments of moral blame can distort memory for events. *Memory & Cognition*, *34*(3), 550–555. https://doi.org/10.3758/BF03193578

Prinzing, M., Earp, B. D., & Knobe, J. (2023). Why do evaluative judgments affect emotion attributions? The roles of judgments about fittingness and the true self. *Cognition*, *239*, 105579. https://doi.org/10.1016/j.cognition.2023.105579

Schein, C., Hester, N., & Gray, K. (2016). The visual guide to morality: Vision as an integrative analogy for moral experience, variability and mechanism. *Social and Personality Psychology Compass*, *10*(4), 231–251. https://doi.org/10.1111/spc3.12247

Sellin, J. T., & Wolfgang, M. E. (1964). *The measurement of delinquency*. Patterson Smith.

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211. https://doi.org/10.1016/j.cognition.2017.05.004

Sinnott-Armstrong, W. (2008). A Contrastivist Manifesto. *Social Epistemology*, *22*(3), 257–270. https://doi.org/10.1080/02691720802546120

Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, *217*, 104890. https://doi.org/10.1016/j.cognition.2021.104890

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *The American Journal of Psychology*, *69*(1), 1. https://doi.org/10.2307/1418112

Stevens, S. S. (1958). Adaptation-level vs. the relativity of judgment. *The American Journal of Psychology*, *71*(4), 633–646. https://doi.org/10.2307/1420322

Stevens, S. S. (1966a). A metric for the social consensus: Methods of sensory psychophysics have been used to gauge the intensity of opinions and attitudes. *Science*, *151*(3710), 530–541. https://doi.org/10.1126/science.151.3710.530

Stevens, S. S. (1966b). Matching functions between loudness and ten other continua. *Perception & Psychophysics*, *1*(1), 5–8. https://doi.org/10.3758/BF03207813

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, *47*(6), 1249–1254. https://doi.org/10.1016/j.jesp.2011.05.010

Ubel, P. A., Jankovic, A., Smith, D., Langa, K. M., & Fagerlin, A. (2005). What Is perfect health to an 85-year-old?: Evidence for scale recalibration in subjective health ratings. *Medical Care*, *43*(10), 1054–1057. https://doi.org/10.1097/01.mlr.0000178193.38413.70

Vanneste, S., Verplaetse, J., Van Hiel, A., & Braeckman, J. (2007). Attention bias toward noncooperative people. A dot probe classification study in cheating detection. *Evolution and Human Behavior*, *28*(4), 272–276. https://doi.org/10.1016/j.evolhumbehav.2007.02.005

Williams, B. (1981). *Moral luck: Philosophical papers 1973-1980*. Cambridge University Press.

Wylie, J., & Gantman, A. (2023). People are curious about immoral and morally ambiguous others. *Scientific Reports*, *13*(1), 7355. https://doi.org/10.1038/s41598-023-30312-9

Xu, X., Bostyn, D., Ren, X., & Roets, A. (2024). An Eastern look at a Western dilemma: Cross-cultural differences in action-balanced trolley dilemmas. *Social Psychological and Personality Science*, 19485506241289459. https://doi.org/10.1177/19485506241289459

Young, L., Koenigs, M., Kruepke, M., & Newman, J. P. (2012). Psychopathy increases perceived moral permissibility of accidents. *Journal of Abnormal Psychology*, *121*(3), 659–667. https://doi.org/10.1037/a0027489

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, *1*(3), 333–349. https://doi.org/10.1007/s13164-010-0027-y

Zimmerman, M. J. (2006). Moral luck: A partial map. *Canadian Journal of Philosophy*, *36*(4), 585–608. https://doi.org/10.1353/cjp.2007.0006

Declarations of interest: None.

Funding: This work was supported by the National Science Foundation Graduate Research Fellowship [grant number: DGE1752134].

Acknowledgement: For helpful conversation and/or comments on previous drafts, we thank Charlotte Austin, Ryan Carlson, Megha Chawla, Brianna Nguyen, Jessie Sun, and Amisha Vyas; the members of the Crockett lab and the Yale Perception and Cognition lab; Liane Young and the Boston College Morality lab; and Fiery Cushman, Joshua Greene, and the members of the Harvard Moral Psychology Research Lab.

CRediT authorship contribution statement:

Vladimir Chituc: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review and editing, Visualization

M.J. Crockett: Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition

B.J. Scholl: Methodology, Writing – original draft, Writing – review & editing, Supervision