Judgment and Decision Making, Vol. 5, No. 7, December 2010, pp. 547–554

Wording effects in moral judgments

Ross E. O'Hara* Dartmouth College
Walter Sinnott-Armstrong Duke University
Nicholas A. Sinnott-Armstrong Brown University

Abstract

As the study of moral judgments grows, it becomes imperative to compare results across studies in order to create unified theories within the field. These efforts are potentially undermined, however, by variations in wording used by different researchers. The current study sought to determine whether, when, and how variations in wording influence moral judgments. Online participants responded to 15 different moral vignettes (e.g., the trolley problem) using 1 of 4 adjectives: "wrong", "inappropriate", "forbidden", or "blameworthy". For half of the sample, these adjectives were preceded by the adverb "morally". Results indicated that people were more apt to judge an act as wrong or inappropriate than forbidden or blameworthy, and that disgusting acts were rated as more acceptable when "morally" was included. Although some wording differences emerged, effects sizes were small and suggest that studies of moral judgment with different wordings can legitimately be compared.

Keywords: morality, wording, trolley problem, language.

1 Introduction

More and more psychological and neuroscientific research on moral judgments appears each year. As fascinating results accumulate, the question arises of whether and how individual studies fit together to form a larger picture. In order to connect various studies and guide future work in this field, researchers need to determine which studies conflict, which support each other, and which are simply talking past each other.

Unfortunately, precise comparisons are hampered by the use of different moral terms across studies. Different researchers ask whether acts are wrong (e.g., Cushman, 2008; Haidt et al., 1993; Schaich Borg et al., 2006; Wheatley & Haidt, 2005), forbidden versus permitted (e.g., Cushman, 2008; Cushman et al., 2006), (in)appropriate (e.g., Greene et al., 2004; Moore et al., 2008; Valdesolo & DeSteno, 2006) or deserve blame (e.g., Cushman, 2008; Pizzaro et al., 2003). Some researchers include the adverb "morally" before these terms (e.g., Moore et al., 2008; Schaich Borg et al., 2006; Wheatley & Haidt, 2005), whereas others do not (e.g., Cushman, 2008; Cushman et al., 2006; Greene et al., 2004; Pizzaro et al., 2003; Valdesolo & DeSteno, 2006). It is unclear whether judgments of what is morally wrong vary in response to the same factors as do judgments of what is forbidden, inappropriate, or blameworthy. Some evidence comes from a meta-analysis on the asymmetry between "forbid" versus "allow" in attitudes research. This study demonstrated that people are reluctant to forbid but will readily not allow, even though these judgments are conceptually equivalent (Holleman, 1999). This asymmetry suggests that moral judgments, as well, may be influenced by subtle variations in wording. Further evidence from Cushman (2008, Experiment 1) showed that harmless acts were judged as more wrong than blameworthy *only* when the act was intended and believed to cause harm. This finding also suggests that people, in some circumstances, will draw fine distinctions between moral terms.

Until the effects of wording variations are understood, we cannot tell whether studies on similar moral issues couched in different terms really agree or disagree. In addition, some wording effects on moral judgments would undermine the search for a moral faculty. Some researchers suggest that moral judgments result from innate psychological mechanisms, or even a moral module that conforms to a universal moral grammar (Dwyer, 1999; Harman, 1999; Hauser et al., 2008; Mikhail, 2007). Others propose dual-process models that build emotions or beliefs, desires, and consequences into the processes that form moral judgments (Cushman et al., 2010; Greene et al., 2004). These theories and many more would be challenged if people judge acts in very different ways based on the moral terms used, because psychologically real mechanisms would be unlikely to vary markedly

*Portions of this research were presented at the 2010 Association for Psychological Science annual convention. The authors thank Jay Hull, Bertram Malle, and the Moral Psychology Research Group for their helpful comments. This paper is dedicated to Nicola Knight, whose untimely death saddened us all. Nicola contributed much inspiration and hard work during the design phase of this study. Address: Ross O'Hara, Department of Psychological and Brain Sciences, Hinman Box 6207, Dartmouth College, Hanover, NH, USA, 03755. Email: ross.ohara@dartmouth.edu.

with such fine differences in wording within a particular natural language. Conversely, if certain patterns of moral judgments are robust enough to persist through non-substantial variations in wording, it would help defend the assumption that these studies are investigating distinctive psychological mechanisms.

In order to bring this research together into a coherent field and determine whether there are distinctive psychological mechanisms to be studied in moral psychology, we need to know whether, when, and how much phrasing questions in different terms may lead to different moral judgments. As an exploratory first step toward answering these questions, we tested the effects of four different moral adjectives across six different types of moral judgments.

2 Methods

2.1 Participants

Adult participants who had an internet protocol address within the United States were recruited through Amazon Mechanical Turk. This online participant recruitment system has been shown to produce quality data (Hsueh et al., 2009; Paolacci et al., 2010; see also Kittur et al., 2008, for a description of this system). A total of 845 participants received $3 for completing the study. Ninety-seven participants were removed for insufficient responding (withdrawal before completing the second block), seven for suspicious responding (predominantly entering the first available response across measures), and one for being younger than 18 years, leaving a final sample of 740 participants (716 participants provided complete data; see Table 1 for demographics). The sample was majority female (60.3%), White (70.1%), and ranged in age from 18 to 85 years (M = 33.5 years, SD = 11.38 years). To account for cohort effects, we controlled for age in all analyses.

Table 1: Demographics of the final sample (N = 740)

	n or M (SD)	% or Median
Sex		
Female	446	57.9
Male	261	33.9
Age	33.5 (11.4)	31
Race/Ethnicity		
White	540	78.1
Asian-American	68	9.8
Other	48	6.9
African-American	35	5.1
Religion		
Protestant	215	27.9
Catholic	122	15.8
None	111	14.4
Other	85	12.7
Atheist	66	9.9
Hindu	25	3.7
Jewish	15	2.2
Buddhist	14	2.1
Mormon	14	2.1
Yearly income		
< $20,000	146	19.0
$20,001 - $40,000	167	21.7
$40,001 - $60,000	119	15.5
$60,001 - $80,000	97	12.6
$80,001 - $100,000	53	6.9
> $100,000	36	4.7
Highest level of education		
High school degree	89	12.6
Some college	209	27.1
Associate's degree	52	6.8
Bachelor's degree	248	32.2
Post-graduate degree	104	13.5

2.2 Moral vignettes

Participants read 15 vignettes, each displayed on a new screen, which presented a hypothetical person's morally ambiguous behavior. After each vignette, participants responded to a statement expressing disapproval of the behavior (e.g., "turning the train was wrong") using a 9-point Likert-type scale (1 = *strongly disagree*; 5 = *neither agree nor disagree*; 9 = *strongly agree*). Low values, therefore, indicated acceptability and high values indicated unacceptability. The 15 vignettes were divided into six blocks of moral judgments[1]:

- Trolley — three vignettes in which the actor kills one person in order to save five others by either flipping a switch to divert a train (sidetrack, loop) or pushing a man in front of the train (footbridge) (Hauser et al., 2007).[2]

[1]These blocks loosely followed Haidt's (2007) distinctions between types of morality (namely harm, purity, and fairness), but were not an exhaustive list. The full vignettes are available from the first author upon request.

[2]Hauser et al. (2007) included diagrams with some of these vi-

- Victimless — three vignettes describing taboo behaviors: brother-sister incest (Haidt, 2001)[3], cannibalism, and interspecies sex.
- Harm versus offense — two vignettes comparing a private transgression, stealing money from a lost wallet (Greene et al., 2001)[4], with a public taboo, sexual intercourse.
- Deceit — two vignettes comparing deception through lying versus omission.
- Moral luck — three vignettes in which a drunk driver ignores a stoplight and either kills a pedestrian, misses a pedestrian, or there is no pedestrian present.
- Disgust — two vignettes that compare sloppily eating unconventional foods privately versus publically (Feinberg, 1985).

2.3 Design and procedure

The experiment was a between-subjects 2 (Order) x 2 (Adverb) x 4 (Adjective) randomized full factorial. Participants were randomly assigned to respond to the 15 moral vignettes with 1 of 4 adjectives: "wrong", "inappropriate", "forbidden", or "blameworthy". Whether the adjective was preceded by the adverb "morally" was also randomly assigned. The judgment made by a given participant (e.g., "morally wrong") remained constant across vignettes. Blocks were assumed to be independent and, thus, were presented in the same order across participants. Presentation order within each block, however, was randomly assigned to 1 of 2 conditions (Table 2).

A short description of the survey, including compensation, was posted online. Participants voluntarily clicked a hyperlink that directed them to our website, which served a multi-page Ruby on Rails application. After participants indicated they were at least 18 years old and provided informed consent, they made 15 moral judgments and provided demographic information.[5]

2.4 Analysis and power

To test for general effects of wording variations on moral judgments, a repeated measures analysis of covariance (ANCOVA) controlling for age was performed, with between-subjects factors of Order (2 levels), Adverb (2 levels), and Adjective (4 levels); Vignette (15 levels) was the within-subjects factor. This analysis, however, ignored the distinction between different types of moral judgments. To determine, therefore, whether effects were limited to specific types of morality, repeated measures ANCOVAs were performed separately for each block. The within-subjects factor in these six ANCOVAs contained either 2 or 3 levels, depending on the number of vignettes in that block.

Table 2: Presentation order of moral vignettes.

Block	Order 1	Order 2
Trolley	Sidetrack	Footbridge
	Loop	Loop
	Footbridge	Sidetrack
Victimless	Incest	Interspecies sex
	Cannibalism	Cannibalism
	Interspecies sex	Incest
Harm versus offense	Lost wallet	Public sex
	Public sex	Lost wallet
Deceit	Lying	Omission
	Omission	Lying
Moral luck	Pedestrian killed	Pedestrian absent
	Pedestrian missed	Pedestrian missed
	Pedestrian absent	Pedestrian killed
Disgust	Private	Public
	Public	Private

Because sufficient power is required to claim meaningful null effects (i.e., wording makes no difference), we conducted a sensitivity analysis using *G*Power 3.1* (Faul et al., 2007) to determine how small of an effect we could detect in each ANCOVA. For a repeated measures ANCOVA with a within-between factor interaction, 16 between-subjects groups, 15 repeated measures (Cronbach's α = .81), a Type II error probability of α = .05, and power equal to .80, we could find an effect size > .032 (i.e., a small effect; Cohen, 1992). We also performed this analysis for each block separately and found we could detect an effect > .06 for blocks with three judgments, and > .08 for blocks with two judgments, both small effects. After collecting the data, we confirmed that we achieved sufficient power to find these effects using

gnettes. To keep these vignettes comparable to the other blocks in the current study, we omitted these diagrams.

[3]This vignette was adapted to no longer indicate that the sexual intercourse drew the brother and sister closer together. We made this change in order to make the vignette more morally ambiguous.

[4]This vignette was adapted from the first-person to the third-person.

[5]Participants were instructed not to use the back button on their internet browser, but in such rare instances all responses were recorded. We used participants' first response to all items unless they changed a non-response to a response.

a *post-hoc* power analysis for each block. Using a conservative estimated effect size of .08 and a Type II error probability of $\alpha = .05$, power for 5 of 6 tests was > .94, the exception being the harm versus offense block, which had power of .81. These tests provided evidence that our analyses were sufficient for detecting a small effect in the data.

3 Results

3.1 Overall analysis

Main effects. The repeated measures ANCOVA[6] for all 15 moral judgments revealed significant between-subjects main effects for Order, $F(1,622) = 5.18$, $p = .023$, generalized eta-squared ($\eta_G{}^2$; Bakeman, 2005; Olejnik & Algina, 2003) = .040, Adverb, $F(1,622) = 4.75$, $p = .030$, $\eta_G{}^2 = .008$, and Adjective, $F(3,622) = 4.04$, $p = .007$, $\eta_G{}^2 = .019$, and a significant within-subjects effect of Vignette, $F(14,8708) = 30.53$, $p < .001$, $\eta_G{}^2 = .047$ (see Table 3 for all means and standard deviations). Because Order is meaningless across blocks, it is explored below in further detail. For Adverb, participants were more accepting when "morally" was present ($M = 5.57$, $SE = .06$) versus absent ($M = 5.76$, $SE = .06$). For Adjective, participants judged acts as more wrong ($M = 5.81$, $SE = .08$) or inappropriate ($M = 5.80$, $SE = .08$) than either forbidden ($M = 5.59$, $SE = .08$) or blameworthy ($M = 5.45$, $SE = .09$). Figure 1 displays means and standard errors for the eight Adverb x Adjective conditions.

Interactions. There were no significant between-subjects interactions. Because the assumption of sphericity was violated, the Greenhouse-Geisser correction (1959) was applied to all within-subjects interactions. The Vignette x Order interaction, $F(8,8708) = 13.83$, $p < .001$, $\eta_G{}^2 = .012$, and the Vignette x Adverb interaction, $F(8,8708) = 3.06$, $p = .002$, $\eta_G{}^2 = .003$, were significant. To understand these interactions, they are described below for each block in which they achieved significance.

3.2 Analysis by block

Main effects. Each block showed a significant within-subjects Vignette effect. The Order main effect was found for the Trolley block, $F(1,646) = 55.23$, $p < .001$, $\eta_G{}^2 = .079$: these behaviors were judged more acceptable when side track was presented first ($M = 3.73$, $SE = .11$) versus last ($M = 4.90$, $SE = .11$); and the Disgust block, $F(1,643) = 10.31$, $p = .001$, $\eta_G{}^2 = .016$: unconventional eating was judged more acceptable when first described publically ($M = 3.46$, $SE = .12$) versus privately ($M = 4.01$, $SE = .12$). The Adverb effect only emerged for the Disgust block, $F(1,643) = 17.68$, $p < .001$, $\eta_G{}^2 = .027$: these acts were more accepted when "morally" was included ($M = 3.37$, $SE = .12$) versus excluded ($M = 4.09$, $SE = .12$). The Adjective effect was found for the Victimless block, $F(3,639) = 3.46$, $p = .016$, $\eta_G{}^2 = .016$, and the Disgust block, $F(3,643) = 4.68$, $p = .003$, $\eta_G{}^2 = .021$. Participants judged victimless offenses as less blameworthy ($M = 5.60$, $SE = .17$) than either wrong ($M = 6.17$, $SE = .17$), inappropriate ($M = 6.31$, $SE = .16$), or forbidden ($M = 6.17$, $SE = .16$). Disgust acts were judged as more wrong ($M = 4.00$, $SE = .17$) or inappropriate ($M = 4.10$, $SE = .17$) than either forbidden ($M = 3.41$, $SE = .16$) or blameworthy ($M = 3.42$, $SE = .18$).

Interactions. Again, a Greenhouse-Geisser correction was used on all tests. The Vignette x Order interaction emerged only in the Trolley block, but was qualified by a significant Vignette x Order x Adverb x Adjective interaction that did not appear in the overall test, $F(5,1292) = 2.78$, $p = .017$, $\eta_G{}^2 = .011$. When footbridge was presented first, footbridge was rated as more unacceptable than either sidetrack or loop across moral terms. When sidetrack was presented first, however, "blameworthy" (but not "morally blameworthy") showed no significant differences between vignettes, $F(2,70) = 2.36$, $p = .102$. Finally, the Vignette x Adverb interaction emerged for the Disgust block, $F(1,643) = 12.31$, $p < .001$, $\eta_G{}^2 = .019$: there was a smaller difference between judgments of private disgust when "morally" was included ($M = 2.31$, $SE = .15$) versus excluded ($M = 2.66$, $SE = .14$) than there was for public disgust (Ms = 4.44, 5.52, SEs = .14, .13, respectively).

4 Discussion

4.1 Implications

This study suggests that wording effects do not undermine psychological studies of moral judgments. For harm versus offense, deceit, and moral luck, we found no evidence of wording effects, indicating that these types of morality are robust against linguistic variations. We did find wording effects, though, for victimless offenses, disgust, and the trolley scenario. For the Victimless and Disgust blocks, we discovered scaling effects: participants judged victimless offenses as less blameworthy than wrong, inappropriate, or forbidden, and disgust as less blameworthy and less forbidden than wrong or inappropriate. In addition, participants were more likely

[6] Age produced a significant between-subjects effect in the overall analysis, $F(1,622) = 8.32$, $p = .004$, as well as for the Trolley block ($p = .006$), Victimless block ($p = .024$), and Harm versus offense block ($p = .027$). Means indicated in all cases that older participants rated these behaviors are more unacceptable than younger participants.

Table 3: Means and standard deviations for all moral judgments.

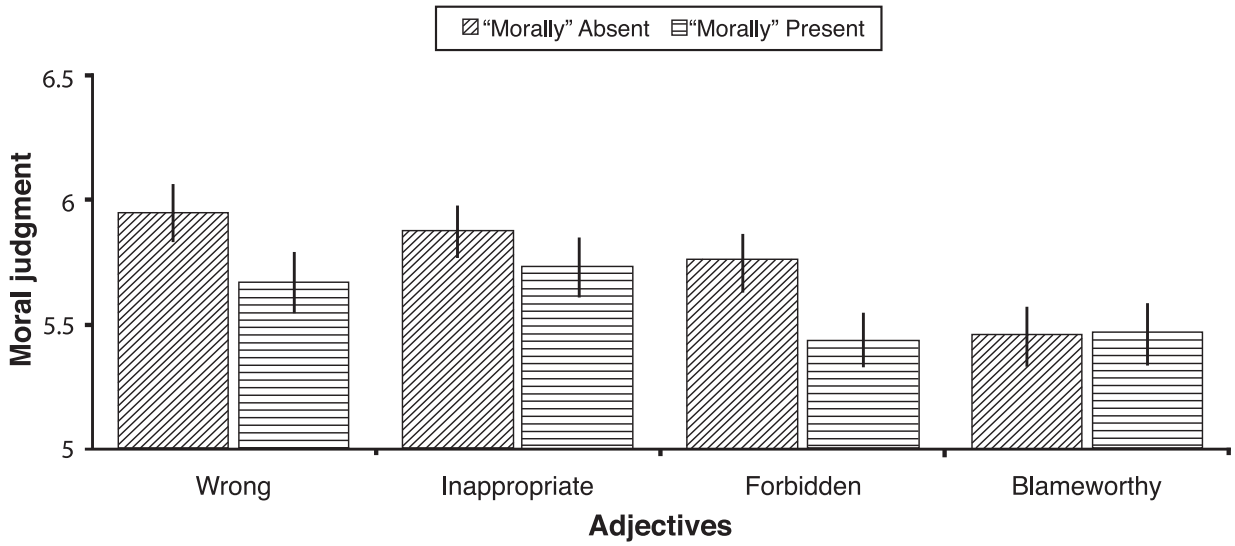
	Order 1								Order 2							
	"Morally" included				"Morally" omitted				"Morally" included				"Morally" omitted			
Vignette	W	I	F	B	W	I	F	B	W	I	F	B	W	I	F	B
Sidetrack	3.70 (2.81)	2.92 (2.78)	2.88 (2.82)	2.45 (3.10)	2.98 (2.66)	3.00 (2.87)	3.35 (2.96)	3.30 (3.02)	4.23 (2.32)	3.76 (2.68)	4.14 (2.38)	3.62 (2.80)	4.44 (2.10)	3.68 (2.81)	4.42 (2.75)	4.41 (2.53)
Loop	3.50 (2.76)	3.13 (2.42)	2.60 (2.60)	2.57 (2.80)	2.92 (2.63)	2.94 (2.65)	3.25 (2.76)	2.70 (2.52)	5.34 (2.42)	5.05 (2.67)	4.86 (2.25)	4.03 (2.97)	5.07 (2.11)	4.37 (2.76)	4.96 (2.57)	5.11 (2.50)
Footbridge	4.95 (2.49)	5.77 (2.48)	5.00 (2.76)	5.12 (2.56)	5.53 (2.68)	5.02 (2.67)	5.58 (2.72)	3.98 (2.59)	6.24 (2.11)	5.76 (2.54)	5.40 (2.27)	5.14 (2.58)	5.54 (2.18)	5.74 (2.78)	5.09 (2.80)	5.37 (2.52)
Incest	5.70 (2.89)	6.59 (2.74)	6.29 (2.68)	5.92 (2.69)	6.54 (2.55)	6.65 (2.27)	6.21 (2.71)	6.18 (2.87)	6.38 (2.54)	7.27 (1.84)	6.39 (2.60)	6.05 (2.99)	6.57 (2.25)	6.12 (2.70)	6.49 (2.62)	5.80 (3.00)
Cannibalism	5.23 (2.96)	5.23 (2.77)	5.63 (2.84)	5.30 (2.86)	5.79 (2.88)	5.71 (2.53)	5.23 (3.07)	4.71 (2.85)	6.25 (2.31)	5.71 (2.80)	5.33 (2.84)	5.19 (2.94)	5.09 (2.80)	5.34 (2.91)	5.56 (2.95)	4.24 (3.21)
Inter. sex	6.48 (1.94)	6.59 (2.60)	6.62 (2.29)	5.95 (2.76)	7.04 (1.86)	6.60 (2.37)	6.02 (2.62)	6.05 (2.31)	6.34 (2.42)	6.79 (2.06)	6.33 (2.57)	6.14 (2.72)	6.24 (2.26)	6.31 (2.75)	5.96 (2.90)	5.67 (3.11)
Lost wallet	5.58 (2.42)	6.10 (2.55)	6.10 (2.48)	5.85 (2.49)	6.04 (2.32)	6.10 (2.26)	5.90 (2.42)	5.15 (2.88)	5.80 (2.36)	5.80 (2.81)	6.22 (2.35)	5.85 (2.92)	6.41 (1.73)	5.91 (2.73)	6.05 (2.34)	5.64 (2.75)
Public sex	5.72 (1.92)	5.87 (2.40)	5.78 (2.27)	6.40 (2.18)	6.71 (2.10)	6.79 (1.89)	6.06 (2.48)	5.50 (2.87)	6.09 (2.33)	6.24 (2.18)	5.93 (2.03)	6.14 (2.36)	5.83 (2.52)	6.17 (2.49)	6.19 (2.29)	6.27 (2.58)
Lying	6.74 (1.44)	6.31 (2.04)	6.41 (1.96)	6.32 (1.91)	6.73 (1.77)	6.54 (2.15)	6.15 (2.32)	5.98 (2.36)	5.93 (1.95)	5.06 (2.46)	5.22 (2.49)	5.57 (2.62)	6.87 (1.66)	6.47 (2.28)	6.07 (2.57)	6.51 (2.27)
Omission	5.82 (1.96)	5.21 (2.71)	5.69 (2.17)	5.63 (2.18)	5.85 (2.12)	5.27 (2.44)	5.00 (2.52)	5.80 (1.81)	5.55 (2.32)	6.07 (2.28)	5.29 (1.98)	5.56 (2.47)	6.89 (1.54)	6.88 (2.04)	6.50 (1.94)	6.75 (1.93)
Ped. killed	7.74 (.83)	7.54 (1.59)	6.92 (2.14)	7.60 (1.36)	7.71 (.77)	7.44 (1.65)	7.23 (2.04)	7.42 (1.38)	7.54 (1.03)	7.14 (2.16)	6.93 (2.42)	7.00 (2.51)	7.53 (1.29)	7.38 (1.98)	7.09 (2.43)	6.98 (2.48)
Ped. missed	7.47 (1.13)	7.36 (1.44)	6.70 (2.05)	7.07 (1.62)	7.60 (.87)	7.56 (1.43)	6.92 (2.16)	7.40 (1.52)	7.04 (1.87)	7.46 (1.05)	7.02 (1.72)	7.25 (1.99)	7.38 (1.35)	7.58 (1.11)	7.05 (2.04)	7.42 (1.86)
Ped. absent	7.24 (1.46)	7.21 (1.54)	6.61 (2.13)	7.00 (1.40)	7.42 (1.37)	7.33 (1.51)	6.60 (2.44)	7.33 (1.38)	2.58 (2.46)	2.45 (2.94)	2.56 (2.62)	1.56 (2.57)	7.41 (1.38)	6.89 (2.19)	6.96 (2.12)	6.93 (2.29)
Private	3.08 (2.90)	2.26 (2.34)	2.33 (2.59)	2.80 (2.84)	3.23 (2.95)	3.81 (2.90)	2.83 (2.96)	2.60 (2.80)	2.34 (2.68)	2.23 (2.66)	2.15 (2.59)	1.71 (2.27)	7.02 (1.56)	7.20 (1.23)	6.70 (1.87)	7.14 (2.00)
Public	5.18 (2.48)	4.97 (2.33)	4.29 (2.76)	4.45 (2.75)	6.34 (2.07)	5.96 (1.89)	5.27 (2.41)	5.95 (2.08)	4.13 (2.65)	4.85 (2.44)	3.80 (2.78)	4.08 (2.26)	5.38 (2.42)	5.48 (2.61)	4.49 (2.81)	4.62 (2.49)

Note. *SD*s in parentheses. Scale: 1 = acceptable; 9 = unacceptable. W = "wrong", I = "inappropriate", F = "forbidden", B = "blameworthy".

to condone disgust when "morally" was included in their judgment. It is noteworthy, however, that the wording effects we found in both blocks did not cross the scale midpoint. Inasmuch as one can assume that responses above versus below the midpoint indicate global judgments of moral unacceptability versus acceptability, wording did not appear to change global judgments, only the strength or certainty of those judgments. In other words, people did not judge incest, for example, as "wrong" *but not* "blameworthy"; the tendency was only to judge it as *more* wrong than blameworthy.

The picture was more complicated for the trolley vignettes. When footbridge was presented first, we replicated the finding that pushing the man in footbridge is judged as less acceptable than flipping the switch in sidetrack or loop (Cushman et al., 2006), regardless of moral term used. When sidetrack was presented first, however, blameworthiness was rated similarly across all three vignettes, suggesting that wording effects do occur in some orders of trolley scenarios. Similar to the results of Cushman (2008), this finding suggests that judgments of blame are determined uniquely from global judgments of wrongness. The current finding that "morally blameworthy" demonstrated the expected order effect further supports this idea, suggesting that the inclusion of the adverb altered this judgment to be more abstract. We must ac-

Figure 1: Mean moral judgments across 15 vignettes by Adverb and Adjective.



Note. Scale: 1 = acceptable; 9 = unacceptable

knowledge, however, that these differences showed very small effect sizes ($\eta_G^2 < .05$), indicating that the influence of wording variations on moral judgments was negligible.

4.2 Limitations

Online data collection limited our control over the participant population and the testing environment. We eliminated approximately one-eighth of the original sample for incomplete or inappropriate responding, but this rate was comparable to previous studies on the utility of Amazon Mechanical Turk (Kittur et al., 2008). Additionally, reaction times suggested that as many as 100 more participants may have rushed through the experiment, but analyses without these participants did not differ from the results presented. This rate of "gaming" the system was substantially lower, however, than in previously published reports (Kittur et al., 2008). In addition, we tested moral wording across an array of moral judgments in a within-subjects design: Having participants judge such varied situations without counterbalancing may have biased our results toward null effects. Future studies may be better served by focusing on a single type of judgment (e.g., the trolley problems) and examining a more comprehensive set of moral terms.

4.3 Conclusion and future directions

Our results indicate that participants in moral psychology studies are interpreting different moral terms in a similar manner, suggesting that researchers are studying a real psychological phenomenon, not a linguistic artifact. Our findings are also compatible with the possibility of a universal moral faculty or grammar. Although we did not examine every moral term used in previous studies or in our natural language, we believe the use of eight common terms makes our results sufficiently generalizable. It seems unlikely that people would process "wrong", "inappropriate", "forbidden", and "blameworthy" in similar ways yet provide radically different responses to another related term.

Researchers should, however, still be cautious regarding the terms used in their studies. We found evidence that people are less apt to forbid (Holleman, 1999) or to lay blame (Cushman, 2008), so to the extent that the magnitude of judgments is relevant to one's research question, similar terms should be used across studies. In addition, we also found evidence that judgments of purity may be more susceptible to these wording effects than harmful acts. These effects are likely to be found in any situation, like the disgust scenarios, in which the behaviors are governed by cultural norms but not formal rules or laws. It may be prudent, therefore, that studies of moral purity only be compared when the adverb "morally" is included in participants' judgments. Finally, the trolley vignettes appeared to be interpreted differently when participants were asked to judge blameworthiness, but not moral blameworthiness: future research should take care not to treat wrongness and blameworthiness as interchangeable concepts, but acknowledge that they are likely derived from different processes (Cushman, 2008).

As the field expands, a meta-analysis on moral judgment research that examines wording as an independent

variable will be necessary. From our findings, we do not expect large wording effects to emerge, but such research is yet to be completed. In addition, neither the types of morality studied nor the moral terms used were exhaustive, limitations that require follow-up studies. Related to this point, future research should also test whether wording may shift people between utilitarian and deontological perspectives. Some of the terms used in the current study were more relevant to a deontological framework (e.g., "forbidden") than a utilitarian one, and follow-up work could compare these moral terms to judgments of whether behaviors should be done or would be best. These steps will facilitate comparison of studies in the field of moral psychology and help build a coherent picture of how people understand morality.

References

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*, 379–384.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*, 353–380.

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. Doris (Ed.), *Moral psychology handbook* (pp. 47–71). New York: Oxford University Press.

Cushman, F., Young, L. & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgments: Testing three principles of harm. *Psychological Science, 17*, 1082–1089.

Dwyer, S. (1999). Moral competence. In R. Stainton (Ed.), *Philosophy and Linguistics* (pp. 169–190). CITY: Westview Press.

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

Feinberg, J. (1985). *Offense to others: The moral limits of the criminal law, Vol. 2*. Oxford: Oxford University Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105- 2108.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95–112.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*, 998–1002.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*, 613–628.

Harman. G. (1999). Moral philosophy and linguistics. In K. Brinkmann (Ed.), *Proceedings of the 20th World Congress of Philosophy: Volume I: Ethics* (pp. 107–115) Bowling Green, OH: Philosophy Documentation Center.

Hauser, M. D., Cushman, F., Young, L., Jin, R. K., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*(1), 1–21.

Hauser, M., Young, L., & Cushman, F. (2008). Reviving Rawls' linguistic analogy: Operative principles and the causal structure of moral action. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Volume 2: The cognitive science of morality* (pp. 107–143). Cambridge, MA: MIT Press.

Holleman, B. (1999). Wording effects in survey research using meta-analysis to explain the forbid/allow asymmetry. *Journal of Quantitative Linguistics, 6*, 29–40.

Hseuh, P., Melville, P., & Sindhwani, V. (2009, June). *Data quality from crowdsourcing: A study of annotation selection criteria*. Paper presented at the 2009 NAACL HLT Workshop on Active Learning for Natural Language Processing. Retrieved December 8, 2010, from http://portal.acm.org/citation.cfm?id=1564131.1564137.

Kittur, A., Chi, E. H., & Suh, B. (2008, April 5–10). *Crowdsourcing user studies with Mechanical Turk*. Paper presented at the 2008 Conference on Human Factors in Computing Systems. Retrieved December 7, 2010, from http://portal.acm.org/citation.cfm?id=1357127.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences, 11*, 143–152.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*, 549–557.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*, 434–447.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411-419.

Pizzaro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*, 267–272.

Schaich Borg, J., Hynes, C., van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience, 18*, 803–817.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*, 476–477.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science, 16*, 780–784.