It's not right but it's permitted: Wording effects in moral judgement

Sergio Barbosa[*] William Jiménez-Leal[†]

Abstract

This study aims to provide evidence about two widely held assumptions in the experimental study of moral judgment. First, that different terms used to ask for moral judgment (e.g., blame, wrongness, permissibility. . .) can be treated as synonyms and hence used interchangeably. Second, that the moral and legal status of the judged action are independent of one another and thus moral judgment have no influence of legal or other conventional considerations. Previous research shows mixed results on these claims. We recruited 660 participants who provided moral judgment to three identical sacrificial dilemmas using seven different terms. We experimentally manipulated the explicit legal status of the judged action. Results suggest that terms that highlight the utilitarian nature of the judged action cause harsher moral judgments as a mechanism of reputation preservation. Also, the manipulation of the legal status of the judged action holds for all considered terms but is larger for impermissibility judgments. Taken as a whole, our results imply that, although subtle, different terms used to ask for moral judgment have theoretically and methodologically relevant differences which calls for further scrutiny.

Keywords: moral judgment, wording effects, social conventions, law

1 Introduction

Recent psychological studies in moral judgment follow a similar pattern: participants are presented with a number of moral dilemmas carefully designed by the researchers to test their research hypothesis. After viewing each dilemma, participants are asked to provide their moral judgments about the presented actions, often with numerical scales ranging from harsh moral condemnation to absolution or even praise. This way to observe moral judgment has the advantage of being easily adaptable to experimental conditions and quantitative research (Bauman, McGraw, Bartels & Warren, 2014); however, it can also be subject to a number of methodological and conceptual limitations. Our study deals with a specific methodological limitation: a wording effect on moral judgment.

The way moral dilemmas are usually phrased is known to affect moral judgment (Borg, Hynes, van Horn, Grafton & Sinnot-Armstrong, 2006). For instance, presenting a dilemma in a highly emotional tone has a significant, although subtle, effect on how wrong participants believe an action is and how willing they are to act in morally dire circumstances (Petrinovich & O'Neil, 1996; Sinnot-Amstrong, 2008). Likewise, negative and positive framing (i.e., 50% of people killed vs 50% of people saved) influences the way people judge a formally identical action. Additionally, different studies inquire about moral judgments using different terms, for instance asking whether a specific action is appropriate, right or wrong, permissible or simply if the participant would be willing to it carry out. Researchers interpret these different terms as if they were synonyms of one another (Greene et al. 2001; Bjorklund, 2003, Cushman, Young & Hausser, 2006; Koenings, Kruepke, Zeier & Newman, 2012). Empirical research shows mixed results on whether different terms used to ask for moral judgment are empirically indistinguishable.

First, O'Hara, Sinnott-Armstrong and Sinnott-Armstrong (2010) used a within subject design to present participants 15 moral scenarios. Scenarios were loosely based on Moral Foundation Theory (Graham, Haidt & Nosek, 2009) and included three formulations of the trolley dilemma, two 'victimless dilemmas' and two 'purity dilemmas', two 'harm versus offense' dilemmas, two deception dilemmas enacting the action-omission distinction, and two dilemmas dealing with moral luck. Every participant saw all possible dilemmas but provided moral judgment using only one random term per dilemma. Dilemmas were presented in blocks that corresponded to all formulations of a specific dilemma type. All participants saw blocks in the same order. Relevant results showed that the different terms used to ask for moral judgment had a small yet significant effect at the overall level ($\eta_g^2 = 0.019$). In addition, at the block level, disgusting and victimless dilemmas showed a significant difference between the terms used to ask for moral judgment ($\eta_g = 0.016$ and $\eta_g = 0.021$ respectively). According to author's interpretation of these results "wording effects do not undermine psychological studies of moral judgments." (O'Hara et al, 2010, p 550).

Copyright: © 2017. The authors license this article under the terms of the Creative Commons Attribution 3.0 License.

[*] Departamento de Psicología, Universidad de los Andes, Cra. 1 Nº 18ª–12, Edificio Franco Bogotá, 111711. Colombia. Email: s.barbosa11@uniandes.edu.co.

[†] Departamento de Psicología, Universidad de los Andes.

https://doi.org/10.1017/S1930297500005908 Published online by Cambridge University Press

Several objections can be raised to these results and their interpretation. First, the fact that all participants saw all possible formulations of all dilemmas may have disclosed the purposes of the study. Moreover, the fact that all blocks were presented in the same order may have caused a carryover effect. Additionally, conclusions that different terms used to ask for moral judgment are interchangeable are drawn from a small effects size rather than from a lack of significant differences between terms. Interpreting theoretically meaningful differences based on apparently small effects sizes is a delicate matter since usual effect sizes in social sciences are rarely larger than moderate. O'Hara et al's (2010) study exhibited sufficient statistical power to observe small effects sizes so it is to be expected that they provide an accurate estimate of the effect size of wording effects in moral judgment. Yet, it is possible that the observed small effect sizes reflect theoretically meaningful differences.

Cushman (2008) presented participants with dilemmas in which the agent's intention and foreknowledge as well as actual harm caused were systematically manipulated. Participants provided blame, wrongness, permissibility and punishment judgments to different dilemmas with different conditions of agent's intention, belief and caused harm. Results showed that the variance explained by manipulating the agent's mental states (i.e., intention and foreknowledge) is much larger for blame and punishment judgments than for wrongness and permissibility judgment. In contrast, the variance of wrongness and permissibility judgments is largely explained by the actual harm caused by the agent rather than by her mental states. These results suggest an important theoretical distinction between wrongness and permissibility judgments, on the one hand, and blame and punishment judgments, on the other. Wrongness and permissibility are influenced by caused harm. On the contrary, blame and punishment judgments are swayed more by the agent's mental states rather than the caused harm. This distinction between action and intent-based moral judgment is of paramount importance on the development of moral judgment. Similarly, it has been shown that moral judgment are distinguishable from willingness to carry out specific actions (Gold, Pulford & Colman, 2015; Tassy, Oullier, Mancini & Wicker, 2013). For example, in a trolley dilemma, more participants believe that the trolley should be diverted than those that deem diverting the trolley permissible (Gold et al., 2015). Given the conflicting findings on the assumption that different terms used to ask for moral judgment are interchangeable, the aim of our study is to offer empirical evidence as to whether these terms are actually synonyms of one another.

As described earlier, O'Hara et al's (2010) conclusions that wording effects in moral judgment have no theoretical significance are based on significant but quantitatively small wording effects. When analyzed at the block level wording effects are significant only when dealing with disgusting (i.e., sloppy eating in private vs in public) and victimless (i.e., incest between consenting adults) dilemmas. In the original authors' view, this is evidence in favor of wording effects being of no theoretical importance in the study of moral judgment. We suggest that these results are compatible with an alternative explanation. The original study used six types of moral dilemmas (trolley, victimless, harm vs offense, deceit, moral luck and disgusting dilemmas). We reason that among these dilemma types, trolley, deceit, moral luck and harm vs offense dilemmas have some clear *legal* implications. Indeed, deceiving someone (deceit dilemmas), drunk driving (moral luck), keeping someone else's wallet (harm vs offense dilemmas) and even killing a person in order to save five (trolley dilemmas) have some implied legal consequences. However, neither consensual incest (victimless dilemmas) nor sloppy eating (disgusting dilemmas) have clear legal consequences. Thus, we claim that at least part of these results could be driven by an uncontrolled effect of inferred legal considerations rather than by "pure" moral judgment. For instance, permissible/ impermissible actions may imply the *legal* permission to carry out a specific action (i.e., one may or may not be allowed by law to carry a concealed weapon). In this view, permissible/impermissible judgments may be heavily influenced by legal and conventional considerations and would fall in the realm of conventional, rather than moral, transgressions. On the contrary, right/wrong judgments may deal with "pure" moral considerations (i.e., one may think it is wrong to cheat on one's spouse even though cheating is not legally punishable in most western societies) and thus would be a "pure" moral transgression. We predict an effect of the explicit legal status of the judged action only for those terms that heavily rely on legal considerations (e.g., permissible/ impermissible judgments) but not when terms only reflect "true" moral considerations (e.g., right/ wrong judgments).

1.1 Overview

This study aims to provide evidence relevant to the claim that different terms used to ask for moral judgment can be used interchangeably in research on moral judgment. In view of methodological limitations of previous studies on wording and moral judgment, we randomize dilemma order. Also, every participant will offer moral judgments using all considered terms in a randomized order for each participant. This procedure will allow participants to directly compare different terms and, contrary to previous studies, may yield a more accurate estimate of the differences between them. Conversely, if we fail to find significant differences between terms used to ask for moral judgement, we will interpret results as evidence in favor of the claim that different terms can be used interchangeably. Also, we hypothesize that, at least some terms (e.g., permissible/impermissible ratings), are heavily influenced by conventional (i.e., legal) rather than "pure" moral considerations. Explicitly manipulating

https://doi.org/10.1017/S1930297500005908 Published online by Cambridge University Press

Table 1: Table 1: Fit measures for participant level analysis.

	AIC	BIC	R2	χ^2
M0: random slopes for participants	39496.5	39666.3	6.85	
M1: only Wording	39481.5	39688.2	6.89	25.01
M2: only Law	38545.1	38729.6	7.12	955.46
M3: Wording and Law, not their interaction.	38529.1	38750.5	7.73	981.46
M4: Wording, Law and interaction	38520.8	38816.1	7.04	1009.70

Chi square test compare each model to M0. All were $p < .01$.

the conventional status of the judged action will allow us to differentiate terms that refer to "pure" moral judgment from those more influenced by conventional considerations.

2 Method

2.1 Participants

We recruited 660 participants (370 women, mean age = 34.27±11.92) via Amazon Mechanical Turk.[1] Participants were paid 0.20 US dollars for their participation.

2.2 Materials

We created five dilemmas loosely based on Moral Foundation Theory. These dilemmas were presented in a random order between subjects. Every dilemma had three versions, which manipulated the explicit legal status of the judged action within participants. Thus, every participant saw the same dilemma in three conditions: an Innocent condition, in which the judged action is explicitly *legal* (the judged action does not entail any sort of legal consequence); a Guilty condition (the judged action is illegal and entails a jail sentence of four years); and a Control condition (legal consequences are not mentioned). The order in which dilemmas were presented was randomized for each participant.

2.3 Procedure

After accepting the task, participants were redirected to a Qualtrics survey to carry out data collection. Participants answered demographics question and then were randomly assigned to view a set of three similar dilemmas. All dilemmas were sacrificial dilemmas in which the agent must choose whether to perform (utilitarian choice) or not (deontological choice) a relatively small moral transgression in order to prevent a larger moral harm. These dilemmas were identical except for the manipulation of the explicit legal status of the judged actions. Participants could evaluate the dilemma with six types of judgment: wrongness (*Wrong*), blame (*Blame*), impermissibility (*Impermissible*) and unacceptability (*Unacceptable*) as well as whether the agent should choose the utilitarian choice (*Should*) and whether the best action was the utilitarian choice (*Best Action*). All moral judgments were reported on a 7-point scale where higher number represented a stronger condemnation of the utilitarian choice. The order of the terms was randomized for every participant. The entire study took an average of six minutes per participant.

3 Results

To account for individual and item-level variance and the nested nature of our results, and to ensure that dilemma's psychometric properties did not bias individual data, we ran two multilevel analysis with participant and dilemma as grouping variables with random slopes for participants and dilemmas. In both cases, we tested five models: M0, which only included random slopes for participant or dilemma; M1, which included only the Wording variable; M2, which included only the Law variable; M3, which included both Wording and Law variables but not their interaction; and M4, which included both Wording and Law variables and their interaction. Table 1 presents the models and corresponding fit measures.

As presented in Table 1, participant-level analysis suggests that, depending on the chosen fit, index M3 or M4 exhibits better fit to observed data (BIC M3=38529.1 > BIC M4= 38520.8, yet AIC M3 = 38750.5 < AIM M4 =38816.1). Given this discrepancy the effects of the Law and Wording variables were further explored by running a series of planned comparisons (t-tests) that showed the expected differences between all three conditions of the Law variable. Namely, mean moral judgment was significantly harsher in the Guilty condition (mean = 4.369±0.51) than in both Innocent (mean = 3.354 ±0.54) and Control conditions (mean = 3.904±0.54) (all p < 0.001). Also, the *Should* and *Best Action* terms yielded significantly harsher moral judgment than all other terms (all p < 0.001). As for the interaction of the

[1] Given the subtlety of the Law manipulation we recruited a larger sample of 1397 participants. This sample was filtered out by asking explicit legal ratings of the judged actions (i.e., Is [action] illegal?). Participants who failed to take the Law manipulation into account were dropped from analysis. Legal ratings were not included in any of the described analysis.

https://doi.org/10.1017/S1930297500005908 Published online by Cambridge University Press

Table 2: Fit measures for dilemma level analysis.

	AIC	BIC	R2	χ^2
M0: random slopes for participants.	48542.3	48712.1	2.14	
M1: only Wording	48547.9	48754.6	2.05	4.40
M2: only Law	48239.9	48424.5	2.14	306.38
M3: Wording and Law, not their interaction	48245.4	48466.9	2.05	310.84
M4: Wording, Law and interaction	48256.8	48552.1	2.05	319.53

Chi square test compare each model to M0. All were $p < .01$ except M1.

Law and Wording variables, planned comparisons showed that the Law manipulation held for all different words used (all p < 0.001). Eta-squares changed substantially according to considered term (η^2 *Wrong* = .015; η^2 *Blame* = .018; η^2 *Impermissible* = .031; η^2 *Unacceptable* = .018; η^2 *Should* = .018; η^2 *Best Action* = .015). Notice that the Law manipulation has an effect size almost twice as large when considering impermissibility ratings compared to any other term.

As presented in Table 2, dilemma-level analysis suggests that M2 better fitted observed data. The effect of the Law variable was further explored by a series of t-tests that showed the expected effects. Namely, we found that ratings in the Guilty condition (Mean = 4.369±1.985) were significantly higher than in the Control condition (Mean = 3.929±1.967, t(115.38) = 1.989, p = 0.049). In addition, the Innocent condition yielded more lenient moral judgments than the Control condition (Mean = 3.682±1.997 and Mean = 4.093±1.967, for the Innocent and Control conditions respectively) but difference was not significant (t(117.85) = 1.0566, p = 0.2929). Also, replicating results in the participant level analysis, *Should* and *Best Action* terms showed significantly harsher moral judgment than all other terms (all p < 0.001).

4 Discussion

The main purpose of this study is to assess the claim that different words used to ask for moral judgment can be used interchangeably. Previous studies (such as O'Hara et al., 2010) have failed to find relevant differences between different terms, even though they did observe a number of wording effects, especially when dealing with legally ambiguous transgressions (i.e., disgusting transgressions and victimless crimes). On the other hand, some studies have theoretically and empirically shown differences between kinds of moral judgments and between moral judgment and actions (Cushman, 2008; Malle, Guglielmo & Monroe, 2014). Given the discrepancies in the relevant literature, we selected several terms used to elicit moral judgment (*Wrongness*, *Blame*, *Permissibility*, *Acceptability*, *Should* and *Best Action*) and asked for moral judgments on 15 sacrificial dilemmas (Gold, Pulford & Colman, 2014; Gold et al., 2015). Both participant and dilemma level analysis suggest that asking for judgments that highlight the utilitarian nature of the action (i.e. *Should* and *Best Action* terms) causes harsher moral judgments *against* the utilitarian option compared to any other term. This is consistent with previous studies on utilitarian moral judgment that suggest that utilitarians are perceived to be callous and cold and that a number of undesirable character traits such as Machiavellianism are associated with utilitarian moral judgments (Royzman, Landy & Leeman, 2015; Uhlmann, Zhu, & Tannenbaum, 2013). Indeed, highlighting the utilitarian nature of the action may motivate participants to protect their reputation against accusations of being Machiavelian or cold by being especially harsh against utilitarian actions. Our results are coherent with these reputational biases being at play, even in the highly anonymous environment of mTurk. Hence, our study suggest that terms that may easily be linked to utilitarian considerations may be especially sensitive to reputational or self-presentation biases.

Our second claim was that at least some of the wording effects found by O'Hara et al (2010) could be explained by an uncontrolled effect of perceived *legal* rather than *moral* transgressions, that is, by whether participants believed that at least some of the presented transgressions were illegal even though they may deem them morally acceptable. As predicted, overall moral judgments were significantly harsher in the Guilty condition than in the Control condition and significantly lighter in the Innocent condition than in the Control condition across all terms. In addition, participant level analysis exhibits a significant Law and Wording interaction. This interaction corresponds to a larger effect size of Law when considering impermissibility ratings (η^2 = .031) compared to any other term (all η^2 below .018). These results suggest that permissibility judgements are more sensitive to legal or conventional considerations than all other terms.

Taken as a whole our results suggest that different terms used to ask for moral judgment are not interchangeable but rather that different terms may be subject to different and specific influences like the legal status (Impermissibility judgments) or the utilitarian nature (Should and Best Action judg-

https://doi.org/10.1017/S1930297500005908 Published online by Cambridge University Press

ments) of the judged action. Our results support previous studies pointing that blame and punishment judgments are heavily linked to agent's mental states whereas wrongness and permissibility are specifically influenced by the actual causal role of the agent with less regard for her mental states (Cushman, 2008; Malle et al., 2014).

Finally, we have to emphasize that the Law and Wording interaction was not significant in the dilemma level analysis which highlights the importance of properly validated materials in moral psychology and motivates the replication and extension of our results (Baron, Gürçay, Moore & Starcke, 2012; Christensen, Flexas, Calabrese, Gut & Gomila, 2014; Lotto, Manfrinati & Sarlo, 2014; Moore, Lee, Clark & Conway, 2011). Future studies should address the subtle yet theoretically meaningful differences between terms used to ask for moral judgment and design a theoretically and empirically informed list of types of moral judgment and their specific features in order to offer a cohesive theoretical framework for moral judgment research.

References

Baron, J., Gürçay, B., Moore, A. B., & Starcke, K. (2012). Use of a Rasch model to predict response times to utilitarian moral dilemmas. *Synthese, 189*(S1), 107–117. http://dx.doi.org/10.1007/s11229-012-0121-z.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*(1), 154–161. http://dx.doi.org/10.1016/j.cognition.2011.05.010.

Bauman, C. W., McGraw, P. A., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8/9*, 536–554. http://dx.doi.org/10.1111/spc3.12131.

Bjorklund, F. (2003). Differences in the justification of choices in moral dilemmas: Effects of gender, time pressure and dilemma seriousness. *Scand J Psychol, 44*, 459–466.

Borg, J. S., 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*(1), 803–817.

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: a moral dilemma validation study. *Frontiers in Psychology, 5*, 607. http://dx.doi.org/10.3389/fpsyg.2014.00607.

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

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*(1), 6-21. http://dx.doi.org/10.1016/j.cognition.2012.11.008.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment testing three principles of harm. *Psychol Sci, 17*(12), 1082–1089.

Gold, N., Pulford, B. D., & Colman, A. M. (2014). The outlandish, the realistic, and the real: contextual manipulation and agent role effects in trolley problems. *Front Psychol, 5*, 35. http://dx.doi.org/10.3389/fpsyg.2014.00035.

Gold, N., Pulford, B. D., & Colman, A. M. (2015). Do as I Say, Don't Do as I Do: Differences in moral judgments do not translate into differences in decisions in real-life trolley problems. *Journal of Economic Psychology, 47*, 50–61. http://dx.doi.org/10.1016/j.joep.2015.01.001.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029-1046. http://dx.doi.org/10.1037/a0015141.

Greene, J. D., Sommerville, B. R., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105 - 2018.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908–911.

Lotto, L., Manfrinati, A., & Sarlo, M. (2014). A new set of moral dilemmas: Norms for moral acceptability, decision times, and emotional salience. *Journal of Behavioral Decision Making, 27*(1), 57–65. http://dx.doi.org/10.1002/bdm.1782.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*, 147–186. http://dx.doi.org/10.1080/1047840X.2014.877340.

Moore, A. B., Lee, N. Y. A., Clark, B. A. M., & Conway, A. R. A. (2011). In defense of the personal/impersonal distinction in moral psychology research: Cross-cultural validation of the dual process model of moral judgment. *Judgment and Decision Making, 6*(1), 186–195.

O'Hara, R. E., Sinnott-Armstrong, W., & Sinnott-Armstrong, N. A. (2010). Wording effects in moral judgment. *Judgment and Decision Making, 5*(7), 547–554.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17*(3), 145–171. http://dx.doi.org/http://dx.doi.org/10.1016/0162-3095(96)00041-6.

Royzman, E. B., Landy, J. F., & Leeman, R. F. (2015). Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. *Cognitive Science, 39*, 325–352. http://dx.doi.org/10.1111/cogs.12136.

Sinnott-Armstrong, W. (2008). Framing moral intuitions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vol. 2

https://doi.org/10.1017/S1930297500005908 Published online by Cambridge University Press

- The Cognitive Science of Morality, pp. 47-82). Cambridge, Massachusetts: Bradford.

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Front Psychol, 4*, 250. http://dx.doi.org/10.3389/fpsyg.2013.00250.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition, 126*(2), 326–334. http://dx.doi.org/10.1016/j.cognition.2012.10.005.

https://doi.org/10.1017/S1930297500005908 Published online by Cambridge University Press