Reasons or Rationalisations: The Role of Principles in the Moral Dumbfounding Paradigm

Cillian McHugh[1], Marek McGann[2], Eric R. Igou[1], & Elaine L. Kinsella[1]

[1] University of Limerick

[2] Mary Immaculate College ~ University of Limerick

Author Note

All procedures performed in studies involving human participants were approved by institutional research ethics committee and conducted in accordance with the Code of Professional Ethics of the Psychological Society of Ireland, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. All authors consented to the submission of this manuscript.

Correspondence concerning this article should be addressed to Cillian McHugh, Department of Psychology, University of Limerick, Limerick, Ireland, V94 T9PX. E-mail: cillian.mchugh@.ul.ie

Abstract

Moral dumbfounding occurs when people maintain a moral judgment even though they cannot provide reasons for it. Recently, questions have been raised about whether dumbfounding is a real phenomenon. Two reasons have been proposed as guiding the judgments of dumbfounded participants: harm-based reasons (believing an action may cause harm) or norm-based reasons (breaking a moral norm is inherently wrong). Participants who endorsed either reason were excluded from analysis, and instances of moral dumbfounding seemingly reduced to non-significance. We argue that endorsing a reason is not sufficient evidence that a judgment is grounded in that reason. Stronger evidence should additionally account for (a) articulating a given reason, and (b) consistently applying the reason in different situations. Building on this, we develop revised exclusion criteria across 2 studies. Study 1 included an open-ended response option immediately after the presentation of a moral scenario. Responses were coded for mention of harm-based or norm-based reasons. Participants were excluded from analysis if they both articulated and endorsed a given reason. Using these revised criteria for exclusion, we found evidence for dumbfounding, as measured by the selecting of an admission of not having reasons. Study 2 included a further three questions relating to harm-based reasons specifically, assessing the consistency with which people apply harm-based reasons across differing contexts. As predicted, few participants consistently applied, articulated, and endorsed harm-based reasons, and evidence for dumbfounding was found.

Keywords: morality, moral judgment, dumbfounding, intuition, rationalism, reasons

Word count: 11,298

Reasons or Rationalisations: The Role of Principles in the Moral Dumbfounding Paradigm

List of Tables

List of Figures

1 | Introduction

Moral dumbfounding occurs when people maintain a moral judgment even though they cannot provide a reason in support of this judgment (Haidt, 2001; Haidt, Björklund, & Murphy, 2000). It is typically evoked when people encounter taboo behaviors that do not result in any harm (Haidt, 2001; Haidt et al., 2000; see also McHugh et al., 2017). One example of such a behavior can be found in the widely discussed *Incest* scenario, which reads as follows:

> Julie and Mark, who are brother and sister are traveling together in France. They are both on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy it, but they decide not to do it again. They keep that night as a special secret between them, which makes them feel even closer to each other. (Haidt et al., 2000, p. 22)

Incest is considered taboo in most cultures, and in violating this taboo, Julie and Mark's actions are typically judged as wrong. However, the consensual and harmless nature of their actions means that the reasons people generally provide do not apply in this case. People who maintain their judgment in the absence of reasons are identified as morally dumbfounded. McHugh et al. (2017), building on the original work by Haidt et al. (2000), identified two measurable responses that may be taken as indicators of moral dumbfounding. Firstly, people may explicitly admit to not having reasons for their judgment. Secondly, people may use unsupported declarations ("it's just wrong") or tautological reasons ("because it's incest") as justifications for a judgment.

1.1 | The Influence of Moral Dumbfounding

The discovery of moral dumbfounding (Haidt et al., 2000; see also Haidt, Koller, & Dias, 1993) coincided with, and arguably contributed to, some of the key developments in moral psychology over the past two decades. It had a clear influence on the development of Haidt's social intuitionist model of moral judgment (SIM, Haidt, 2001), and by extension may be seen as contributing to the growth of intuitionist theories of moral judgment that followed (e.g., Cushman, Young, & Greene, 2010; Haidt, 2001; Prinz, 2005).

Haidt proposed the SIM in opposition to the perceived dominance of rationalist approaches (Kohlberg, 1969, 1971; Narvaez, 2005; Topolski, Weaver, Martin, & McCoy, 2013). According to rationalist approaches our moral judgments are grounded in reason, informed by discernible moral principles (Fine, 2006; Kennett & Fine, 2009; Kohlberg, 1969, 1971; Royzman, Kim, & Leeman, 2015); Haidt (2001). Moral dumbfounding is presented by Haidt (2001) and by Prinz (2005) as evidence against this rationalist perspective, in that, if moral judgments were grounded in reason, people would be able to provide reasons for their judgments (and moral dumbfounding would not occur). Intuitionist theorists propose that moral judgments are grounded in an emotional or intuitive automatic response rather than slow deliberate reasoning (Cameron, Payne, & Doris, 2013; Haidt, 2001; Prinz, 2005). In recent years the joint role of reason/deliberation and intuition in the making of moral judgments has been emphasised in dual-process theories (Brand, 2016; Crockett, 2013; Cushman, 2013a; Cushman et al., 2010; Greene, 2008). The dumbfounding paradigm may be useful in developing and extending these theories; developing an understanding of moral dumbfounding and the processes that lead to it, may inform the further development of theories of moral judgment, leading to a greater understanding of the processes that underlie moral judgment more generally.

The influence of dumbfounding may be observed in everyday discourse, particularly in relation to highly sensitive and divisive social issues. Real-world interactions differ from a

laboratory study designed to elicit a dumbfounded response, and as such, in the absence of explicit and consistent refuting of arguments, it is unlikely that people in everyday life would admit to not having reasons for their moral judgments. Despite this, it is not uncommon to hear unsupported declarations/tautological statements as arguments in support of a position with no further justification (e.g., Mustonen, Paakkonen, Ryökäs, & Nieminen, 2017; Stepniak, 1995). Similarly, moral positions are often justified by appealing to emotions (e.g., Mustonen et al., 2017; Stepniak, 1995; see also Rozin, Haidt, MacCauley, McKay, & Olatunji, 2008; Rozin, Lowery, Imada, & Haidt, 1999). This type of appeal to emotion has previously been discussed as similar/equivalent to dumbfounding (see Prinz, 2005, p. 101; see also Haidt & Hersh, 2001). These responses may not clearly demonstrate dumbfounding, however they illustrate the way in which discussions of reasons for moral positions are occasionally absent from the public debate.

That people may defend a judgment in the absence of articulated reasons, and maintain it even in the knowledge of their own inconsistencies poses a challenge for the type of rational debate that is supposed to form the basis of public discourse and inform the development of public policy. The study of moral dumbfounding, as an extreme case, may lead to a better understanding of the underlying cognitive processes that lead to these types of problematic practices that have no place in public debate. Identifying these processes and explaining moral dumbfounding is beyond the scope of the current research. Here, in light of recent critiques, here we test whether or not dumbfounding is a real phenomenon, worthy of further study.

1.2 | Challenging the Dumbfounding Paradigm

A key concern regarding the dumbfounding paradigm is that the eliciting scenarios have been artificially construed to remove potentially harmful consequences to the point that they become unrealistic or otherwise not credible (e.g., Jacobson, 2012). It could be argued that studying such idiosyncratic scenarios does little to inform our understanding of

everyday moral decision making; similar criticisms have been made regarding the widely used trolley-type sacrificial dilemmas (e.g., Bauman, McGraw, Bartels, & Warren, 2014; Bostyn, Sevenhant, & Roets, 2018). However, responses to hypothetical trolley dilemmas have been found to predict behaviour in a money burning game with real pay-off consequences (Dickinson & Masclet, 2018), and the study of trolley-type dilemmas arguably contributed to key theoretical advancements of the past two decades (e.g., Plunkett & Greene, 2019; see also Greene, 2008; Christensen, Flexas, Calabrese, Gut, & Gomila, 2014; Christensen & Gomila, 2012; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). If moral dumbfounding is a real phenomenon it may prove a useful paradigm to further advance theories of moral judgment, and examine the mechanisms and cognitive processes that underlie the making of moral judgments (e.g., the relative roles of emotion versus deliberation). It may be possible to identify specific contextual features that may lead people to change their mind rather than provide a dumbfounded response (or vice-versa). Experimental manipulations that may increase dumbfounded responding (e.g., cognitive load) or reduce dumbfounded responding (e.g., distancing) could be investigated. There may also be individual difference variables that predict susceptibility to dumbfounding.

In defending the claim that moral judgments are not caused by reasoning, Haidt (2001) presents moral dumbfounding as a demonstration of inconsistency between judgment and reasons available. The implicit alternative to this argument is that the absence of reasons would lead a moral judgment to change or to be revised; i.e., the presence or absence of reasons can cause a judgment to change. Haidt does not clearly distinguish between *reasoning* as a cause versus *reasons* as a cause of judgments (e.g., 2001, p. 822). Despite being inconsistent with approaches beyond the moral domain (e.g., Mercier, 2016; Johnson-Laird, 2006; Mercier & Sperber, 2011, 2017; Todd & Gigerenzer, 2012), this ambiguity can still be seen in discussions of moral judgment (and moral dumbfounding), such that, for the rationalist perspective (see Haidt, 2001), reasons appear to play a causal role, (e.g., Jacobson, 2012, p. 17; Flanagan, Sarkissian, & Wong, 2008, p. 7; Triskiel, 2016, p.

93). Furthermore, this assumption is implicit in challenges to the dumbfounding narrative, whereby these challenges attempt to demonstrate that people do have "warrantable reasons" for their judgments (Royzman et al., 2015, p. 309). Here we identify and address methodological limitations of one example of this type of challenge to the dumbfounding paradigm (Royzman et al., 2015).

Gray, Schein, and Ward (2014) argue that people's moral judgments are grounded in harm-based reasons, suggesting that when judging moral scenarios, people implicitly perceive harm even in scenarios that are construed as objectively harmless. If people perceive harm in the scenarios, then, even when the experimenter claims that they are harm free, this perception of harm still serves as a reason to condemn the behavior. They conducted a series of experiments demonstrating that people do implicitly perceive harm in supposedly victim-less scenarios; e.g., "masturbating to a picture of one's dead sister, watching animals have sex to become sexually aroused, having sex with a corpse, covering a Bible with feces" (Gray et al., 2014, p. 1063). This suggests that in studies of moral dumbfounding people may also be making judgments based on an implicit perception of harm.

Jacobson (2012) makes specific reference to the scenarios used in the study of moral dumbfounding, and presents a number of plausible reasons why a person may condemn the actions of the characters in these scenarios. In the case of the *Incest* scenario, he suggests that the behavior of Julie and Mark was risky, "reckless and licentious" (Jacobson, 2012, p. 25). Jacobson also discusses another scenario, *Cannibal*, that has been used in studies of moral dumbfounding. This scenario describes an act of cannibalism by a researcher in a pathology lab (Jennifer) on a cadaver from the lab. Jacobson argues that if Jennifer's behavior became known, people would be less willing to donate their bodies to the lab. In addition to providing reasons that may explain the judgments of participants, Jacobson suggests that when participants appear to be dumbfounded they have simply given up on the argument and conceded to the experimenter who is in a position of authority. While this

claim is not directly tested empirically by Jacobson, it has been studied by Royzman et al. (2015), as discussed in the following section.

1.3 | Evidence for Judgments Based on Reasons or Principles

A recent series of studies by Royzman et al. (2015), investigating the *Incest* scenario specifically, aimed to identify if participants presenting as dumbfounded genuinely had no reasons to support their judgments. In line with Jacobson (2012), they claim that dumbfounding occurs as a result of social pressure to adhere to conversational norms, arguing that dumbfounded participants do have reasons for their judgments and that these reasons are incorrectly dismissed as invalid by the experimenter. They argue that dumbfounded responding occurs as a result of social pressure to avoid appearing "uncooperative" (Royzman et al., 2015, p. 299), "inattentive" or "stubborn" (2015, p. 300). In addition to this claim, Royzman et al. (2015) identify two justifying principles that may be guiding participants' judgments: the harm principle and the norm principle. They argue that when excluding from analysis participants who endorse either of these principles, incidences of dumbfounding are negligible.

In identifying the *harm principle*, Royzman et al. (2015) draw on the work of Gray et al. (2014). They hypothesised that participants may not believe the scenario to be harm free even in the face of repeated assurances from the experimenter that it is harm free. If a participant does not believe that an act is truly harm free then this provides them with a perfectly valid reason to judge it as morally wrong (Gray et al., 2014; Royzman et al., 2015). They devised two questions which served as a "credulity check" (Royzman et al., 2015, p. 309), to assess whether or not participants believed that the *Incest* scenario was harm-free. The questions read as follows: (i) "Having read the story and considering the arguments presented, are you able to believe that Julie and Mark's having sex with each other will not negatively affect the quality of their relationship or how they feel about each other later on?"; (ii) "Having read the story and considering the arguments presented, are you able to

believe that Julie and Mark's having sex with each other will have no bad consequences for them personally and/or for those close to them?" (Royzman et al., 2015, p. 302–303). If participants responded "No" to either of these questions, their judgments were attributed to harm-based reasons, and therefore they could not be identified as dumbfounded.

The second principle identified by Royzman et al. (2015) is the *norm principle*. They argue that if people believe that committing a particular act is wrong, regardless of the circumstances, then, for these people, this belief may be sufficient to serve as a reason to condemn the behavior of the characters in the scenario. Royzman et al. (2015) presented participants with two statements: (a) "violating an established moral norm just for fun or personal enjoyment is wrong only in situations where someone is harmed as a result, but is acceptable otherwise"; (b) "violating an established moral norm just for fun or personal enjoyment is inherently wrong even in situations where no one is harmed as a result" (Royzman et al., 2015, p. 305). If participants endorsed (b) over (a) they reasoned that a judgment could be legitimately defended using a normative statement. They suggest that the "unsupported declarations" (Haidt et al., 2000, p. 12) identified by Haidt et al. (2000) are statements of a normative position, and that, rather than being a viewed as a dumbfounded response, they may be viewed as reasons for judgments.

Royzman et al. (2015) used the credulity check to assess if participants' judgments could be attributed to the harm principle, while attributing judgments to the norm principle was based on the norm statements. Royzman et al. (2015) use the phrase "fully convergent" to describe participants who, in their view, are eligible for analysis (Royzman et al., 2015, p. 306). According to Royzman et al. (2015), a participant is fully convergent if their judgment cannot be attributed to either the harm principle or the norm principle. Using these stricter criteria for dumbfounding, Royzman et al. (2015) initially identified 4 participants, from a sample of 53, who presented as dumbfounded. Each of these participants was then interviewed and the inconsistencies in their responses pointed out to them. During these

interviews 2 participants changed their judgment of the behavior and 1 participant changed her position on the normative statements. This left just 1 fully convergent, dumbfounded participant. This participant did not resolve the inconsistency in his responses to the questions, and, following post-experiment interviews, Royzman and colleagues found dumbfounding to occur once in a sample of 53. This was found to be not significantly greater than 0 (Royzman et al., 2015, p. 309), supporting the claim that moral dumbfounding is "highly irregular" or even "non-existent" (Royzman et al., 2015, p. 300; see also Guglielmo, 2018).

1.4 | Reasons or Rationalisations

The studies conducted by Royzman et al. (2015) introduce an additional level of methodological rigor to the study of moral dumbfounding. They clearly demonstrate that people will endorse a reason for a judgment if it is available to them. This undermines the dumbfounding narrative, that people defend a judgment in the absence of reasons, and poses a strong challenge to the existence of moral dumbfounding.

We (McHugh et al., 2017) have previously outlined some limitations with the conclusions presented by Royzman et al. (2015). Firstly, Royzman et al. (2015) suggest that people who present as morally dumbfounded do so in an attempt to avoid appearing "stubborn" or "inattentive" (2015, p. 310). However, Royzman et al. (2015) also employ the original Haidt et al. (2000) definition of moral dumbfounding, which defines moral dumbfounding as "the stubborn and puzzled maintenance of a judgment without supporting reasons"" (Haidt et al., 2000, p. 2; see also Haidt & Björklund, 2008, p. 197; Haidt & Hersh, 2001, p. 194). This means that according to Royzman et al. (2015), people who present as dumbfounded, paradoxically present as stubborn in an attempt to avoid appearing stubborn.

Secondly, the means by which Royzman et al. (2015) arrive at their estimate of 1 instance of moral dumbfounding out of a sample of 53 is problematic for the claim that

moral dumbfounding occurs as a result of social pressure. They present their estimate of 1/53 as not significantly greater than 0/53 ($z = 1$, $p = .315$).[1] However their original estimate of instances of moral dumbfounding was 4/53, which is significantly greater than 0/53 ($z = 2.04$, $p = .041$). These participants were invited back into the lab and the "inconsistencies" in their "responses were pointed out directly" to them (Royzman et al., 2015, p. 308). Furthermore they were then "advised to carefully review and, if appropriate, revise" their responses (Royzman et al., 2015, p. 308). This procedure subjected participants to social pressure to appear consistent in their responding. This illustrates that dumbfounded responding can be influenced by social pressure, however it does not support the stronger claim (by Royzman et al., 2015) that dumbfounded responding can be attributed to social pressure (McHugh et al., 2017). The role of social pressure in eliminating instances of dumbfounded responding is not acknowledged by Royzman et al. (2015).

Finally, demonstrating that people endorse principles that are consistent with their judgments does not provide evidence that these principles are guiding their judgments. In relying on participants' endorsing of a given principle to attribute their judgment to that principle, Royzman et al. (2015) may have falsely excluded some participants from analysis. Consider the following scenario to illustrate this point:

> Two friends (John and Pat) are bored one afternoon and trying to think of something to do. John suggests they go for a swim. Pat declines stating that it's too much effort – to get changed, and then to get dried and then washed and dried again after; he says he'd rather do something that requires less effort. John agrees and adds "Oh yeah, and there's that surfing competition on today so the place will be mobbed". To which Pat replies "Yeah exactly!" (McHugh et al., 2017, p. 20)

[1] No explanation for the responding of this participant is offered. Neither can this participant's response be explained by the theoretical position adopted by Royzman et al. (2015).

It is clear from reading this scenario that even though he endorsed it to support or to rationalise his decision, the surfing competition was not the reason for John's decision not to go to the beach. It would be incorrect to attribute his decision to this reason. The studies conducted by Royzman et al. (2015) do not guard against the possibility of this type of false attribution, and it is likely that some participants were incorrectly excluded from analysis on this basis. This possibility of false exclusion presents a key limitation Royzman et al. (2015) that casts doubt on their findings.

We suggest that attributing people's judgments to principles requires stronger evidence than endorsing alone. We propose two measures that may be useful in establishing whether or not a given principle may truly be identified as a reason for the judgments made by participants. Firstly, participants should be given the opportunity to provide the reason(s) that they based their judgment on, and the reasons provided should inform decisions of inclusion or exclusion.[2] Attributing participants' judgments to particular reasons/principles should account for both the endorsing and the articulating of the reason/principle. Secondly, if a principle is guiding the judgments of participants, this principle should be applied consistently across different contexts. We predict that when these two measures are applied evidence for dumbfounding will be found.

1.5 | The Current Studies

The aim of the current studies was to investigate whether or not people's moral judgments can be attributed to moral principles based on their endorsing of these principles. Specifically, aim to address the concerns raised by McHugh et al. (2017) and test the claim by Royzman et al. (2015) that participants' judgments in the *Incest* scenario can be attributed to the harm principle or the norm principle. Firstly, the degree to which participants articulate either the harm principle or the norm principle as informing their

[2] Participants in Royzman et al. (2015) provided reasons however these reasons did not inform their exclusion criteria.

judgment is examined (Study 1). Secondly, the consistency with which participants apply the harm principle across differing contexts is additionally assessed (Studies 2 and 3). We hypothesise that by developing more rigorous exclusion criteria the rates of false exclusion of participants would be reduced and that evidence for moral dumbfounding would be found, posing a challenge to the type of rationalist perspective described by Haidt (2001). The failure to identify dumbfounded responding would serve as support for these alternative perspectives (e.g., Gray et al., 2014; Guglielmo, 2018; Jacobson, 2012; Royzman et al., 2015; Sneddon, 2007; Wielenberg, 2014) and pose a challenge to SIM as described by Haidt (2001). Given that the exclusion criteria used by Royzman et al. (2015) were developed for the *Incest* dilemma, the studies reported here similarly focus on the *Incest* dilemma specifically.

2 | Study 1: Articulating and Endorsing

In Study 1 we use an existing method for the evoking of dumbfounded responding (McHugh et al., 2017), however, we incorporate to additional materials taken from Royzman et al. (2015) as a more stringent set of criteria for inclusion in analysis. This serves two purposes. If effective, it reduces the likelihood of false inclusions for analysis to identify rates of dumbfounded responding, and also allows us to assess rates at which participants will explicitly articulate or endorse the principles when given the opportunity to do so. In addition to the stricter measure of inclusion proposed by Royzman et al. (2015), we introduce an additional change designed to reduce the possibility of false exclusions. Study 1 was an extension the work of Royzman et al. (2015), using largely the same materials. One moral judgment vignette (*Incest*) was taken from Haidt et al. (2000, Appendix A). Targeted questions, designed to assess participants endorsements of the harm principle or the norm principle, were taken directly from Royzman et al. (2015).

As noted above, if a participant endorses a principle this does not necessarily provide evidence that this principle was guiding their judgment. Relying on the endorsing of principles to determine participants' eligibility for analysis may result in some participants

being falsely excluded from analysis, and any resulting estimate of the prevalence of dumbfounded responding would be inaccurate. In an attempt to control for the possibility of falsely attributing participants' judgments to principles based on endorsing alone, we included an open-ended response option to assess whether or not participants could also articulate these principles. This was presented to participants immediately after the presenting of the vignette. The inclusion or exclusion of participants from analysis, depended on both endorsing and articulating either principle. Participants' judgments were only attributed to a given principle if they both articulated and endorsed that principle. It was hypothesised that participants' endorsing of a principle would not be predictive of their ability to articulate this principle, and that by accounting for this, rates of false attribution and false exclusion would be reduced. We hypothesised that in reducing rates of false exclusion, dumbfounded responding would be observed.

2.1 | Method

2.1.1 | Participants and design. Study 1 was a frequency based extension of Royzman et al. (2015). A combined sample of 110 (60 female, 49 male, 1 other; M_{age} = 32.44, min = 18, max = 69, SD = 11.28) took part. Fifty-eight (25 female, 32 male, 1 other; M_{age} = 38.47, min = 19, max = 69, SD = 12.34) were recruited through MTurk.[3] Participation was voluntary and participants were paid 0.50 US dollars for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden). Fifty-two (35 female, 17 male; M_{age} = 25.71, min = 18, max = 38, SD = 3.80) were recruited through direct electronic correspondence. Participants in this sample were

[3] In order to prevent repeat participation from MTurk workers, this study and all remaining studies conducted on MTurk, were included as part of the same MTurk project as Study 3b from McHugh et al. (2017). In addition, a probe question was included to check if participants had encountered the scenario before. This probe included a follow-up question to determine the nature of participants' previous experience with the scenario.

undergraduate students, postgraduate students, and alumni from Mary Immaculate College (MIC), and University of Limerick (UL). Participation was voluntary and participants did not receive a reward for their participation. Previous research on moral dumbfounding found responses from an MTurk sample and a College sample are largely comparable (see McHugh et al., 2017 Study 3a and 3b).

2.1.2 | Procedure and materials. Data were collected using an online questionnaire generated using Questback (Unipark, 2013). The questionnaire opened with the information sheet and consent form. The main questionnaire was only accessible once consent had been provided. Following the consent form, participants were presented with questions relating to basic demographics. Participants were then presented with two statements to assess if participants' judgments may be grounded in the norm principle. These were taken directly from Royzman et al. (2015): (i) "violating an established moral norm just for fun or personal enjoyment is wrong only in situations where someone is harmed as a result, but is acceptable otherwise."; (ii) "violating an established moral norm just for fun or personal enjoyment is inherently wrong even in situations where no one is harmed as a result.". Participants read both statements and were asked to select the statement they "identify with the most". The order of these statements was randomised. Participants who selected (ii) were then asked to elaborate on their position through an open-ended response question. The purpose of these statements was to assess participants' own *prior beliefs* regarding moral judgment and justifications (see Royzman et al., 2015, p. 331). In order to prevent the potentially confounding influence of a salient example moral scenario, these statements were presented before the moral judgment task.

Participants were then presented with the *Incest* vignette (Appendix A) from the original moral dumbfounding study (Haidt et al., 2000). They were asked to rate on a seven-point Likert scale how right or wrong they would rate the behavior of Julie and Mark (where, 1 = *Morally wrong*; 4 = *Neutral*; 7 = *Morally right*). They were asked to provide a reason for their judgment through open-ended response, and, rated their confidence in their

judgment. Participants were then presented with a series of prepared counter-arguments designed to refute commonly used justifications for rating the behavior as "wrong" (Appendix B).

Dumbfounding was measured using a "critical slide" (developed by McHugh et al., 2017). The critical slide is a page in an online or computer based questionnaire specifically designed to measure dumbfounded responding. It contains a statement defending the behavior and a question as to how the behavior could be wrong ("Julie and Mark's behavior did not harm anyone, how can there be anything wrong with what they did?"). There are three possible answer options: (a) "There is nothing wrong"; (b) an admission of not having reasons ("It's wrong but I can't think of a reason"); and finally a judgment with accompanying justification (c) "It's wrong and I can provide a valid reason". The order of these response options is randomised. Participants who select (c) are prompted on a following slide to type a reason. In line with McHugh et al. (2017), the selecting of option (b), the admission of not having reasons, was taken to be a dumbfounded response.

Following the critical slide, participants rated the behavior, and rated their confidence in their judgment again. They also indicated, on a 7-point Likert scale, how much they changed their mind. A post-discussion questionnaire containing self-report reaction to the scenario across various dimensions (confidence, confusion, irritation, etc.) taken from Haidt et al. (2000) was administered after these revised judgments had been made (Appendix C).

Two targeted questions were taken directly from Royzman et al. (2015) to assess whether or not participants' judgments may be grounded in the harm principle: (i) "Having read the story and considering the arguments presented, are you able to believe that Julie and Mark's having sex with each other will not negatively affect the quality of their relationship or how they feel about each other later on?"; (ii) "Having read the story and considering the arguments presented, are you able to believe that Julie and Mark's having sex with each other will have no bad consequences for them personally and/or for those close

to them?". Participants responded "Yes" or "No" to each of these statements. The order of these questions was randomised.

Two other measures were also taken for exploratory purposes: Meaning in Life questionnaire (MLQ; Steger, Kashdan, Sullivan, & Lorentz, 2008). This ten item scale is made up of two five item sub scales: presence (e.g., "I understand my life's meaning.") and search (e.g., "I am looking for something that makes my life feel meaningful."). Responses were recorded using a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*); and CRSi7 a seven item scale taken from The Centrality of Religiosity Scale (Huber & Huber, 2012). Participants responded to questions relating to the frequency with which they engage in religious or spiritual activity (e.g., "How often do you think about religious issues?"). Responses were recorded using a 5-point Likert scale ranging from 1 (*never*) to 5 (*very often*). The seven item inter-religious version of the scale was selected because some non-religious activities (such as meditation) may also have a bearing on a person's ability to reason about moral issues.

2.2 | Results and Discussion

Eighty-seven of the total sample (N = 110; 79.09%) initially rated the behavior of Julie and Mark as wrong; no difference in initial rating between the MTurk sample (M = 1.98, SD = 1.52), and the MIC sample, (M = 2.10, SD = 1.39), t(107.94) = -0.41, p = .683, d = 0.08. Eighty-six of the total sample, (N = 110; 78.18%) rated the behavior as wrong after viewing the counter-arguments and the critical slide; no difference in revised rating between the MTurk sample, (M = 2, SD = 1.53), and the MIC sample, (M = 2.33, SD = 1.54), t(106.55) = -1.11, p = .268, d = 0.21. A paired samples t-test revealed a significant difference in rating of behavior from time one, initial rating, (M = 2.04, SD = 1.45), to time two, revised rating, (M = 2.15, SD = 1.54), t(109) = -2.38, p = .019, d = 0.08. This result may be due to changes in the severity of the judgments as opposed to changing the judgment. Further analysis revealed that only eight (7.27%) participants changed their judgment: two

participants changed their judgment from "wrong" to "neutral"; one participant changed their judgment from "right" to "neutral"; four changed their judgment from "neutral" to "right"; and one participant changed their judgment from "neutral" to "wrong". A chi-square test for independence revealed no significant association between time of judgment and valence of judgment made, $\chi^2(2, N = 220) = 0.73$, $p = .694$, $V = 0.06$. This rate of changing judgments is lower than the 12% reported in Haidt et al. (2000), however, as noted above, social pressure appears to influence responses in the dumbfounding paradigm. It is likely that the lower rates of changing judgments can be attributed to the reduced social pressure in a computerized task.

Ten participants (9%) indicated that they had encountered the scenario before. When asked to elaborate, participants provided anecdotes, or referred to previous readings (either fiction or philosophy). Two participants (2%) indicated that they had encountered it in a previous survey. The low numbers mean that any potential influence of previous experience on the results is negligible and these participants were not excluded from the analyses.

2.2.1 | Measuring dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Rates of of each response to the critical slide are for the entire sample ($N = 110$) are displayed in Figure 1. Twenty participants (18.18%) were initially identified as dumbfounded.[4] The exclusion criteria developed by Royzman et al. (2015) were applied, all participants who endorsed either the harm principle or the norm principle were excluded from analysis. This left a sample of 14 participants who were eligible for analysis. None of these 14 selected the dumbfounded response.

[4] Unsupported declarations and tautological responses provided in the open-ended responses resulted in an additional six participants presenting as potentially dumbfounded; given that Royzman et al. (2015) argue that these responses are an articulation of a norm/principle, these participants are not identified as dumbfounded here.

The purpose of the Study 1 was to assess if participants could articulate the principles identified by Royzman et al. (2015), independently of the targeted statements/questions, as these may serve as a prompt. A revised measure of convergence is developed here. A participant's endorsement of either principle should lead to their exclusion from analysis, only if the participant also articulated this principle when given the opportunity. The open-ended responses were analysed and coded for any mention of either the harm principle or the norm principle. Participants were only excluded from analysis if they both endorsed and articulated either principle. For the purposes of consistency with Royzman et al. (2015), unsupported declarations and tautological responses (identified as dumbfounded responses by McHugh et al., 2017) were coded as an articulation of the norm principle here.[5] As predicted, the number of participants who both articulated and endorsed either principle was much lower than the number of participants who only endorsed either principle. Fifty two participants were eligible for analysis according to the revised exclusion criteria. Eight of these participants (15.38%) selected the dumbfounded response, providing some evidence for moral dumbfounding. Figure 1 shows the responses to the critical slide for the entire sample and for participants eligible for analysis according to each measure of convergence.

2.2.2 | Consistency between endorsed principles and expressed judgments. The exclusion criteria developed by Royzman et al. (2015) (endorsing only), led to a large proportion of participants who selected "There is nothing wrong" to be excluded from analysis (12 participants; 54.55% of the 22 participants who selected this option). Both the harm principle and the norm principle provide legitimate reasons for participants to judge the behavior as wrong (Royzman et al., 2015). It follows that if a participant endorsed either principle, they would also judge the behavior as wrong. It is surprising then that, 12 of the

[5] By only identifying participants who explicitly admittied to not having a reason as dumbfounded we also reduced the potential risk of "false inclusions", where people provide a dumbfounded response through laziness or inattentiveness. While the motivations for selecting various responses cannot be known, previous research has identified the selecting of an admission of not having reasons as a conservative indicator of moral dumbfounding (McHugh et al., 2017, p. 16).

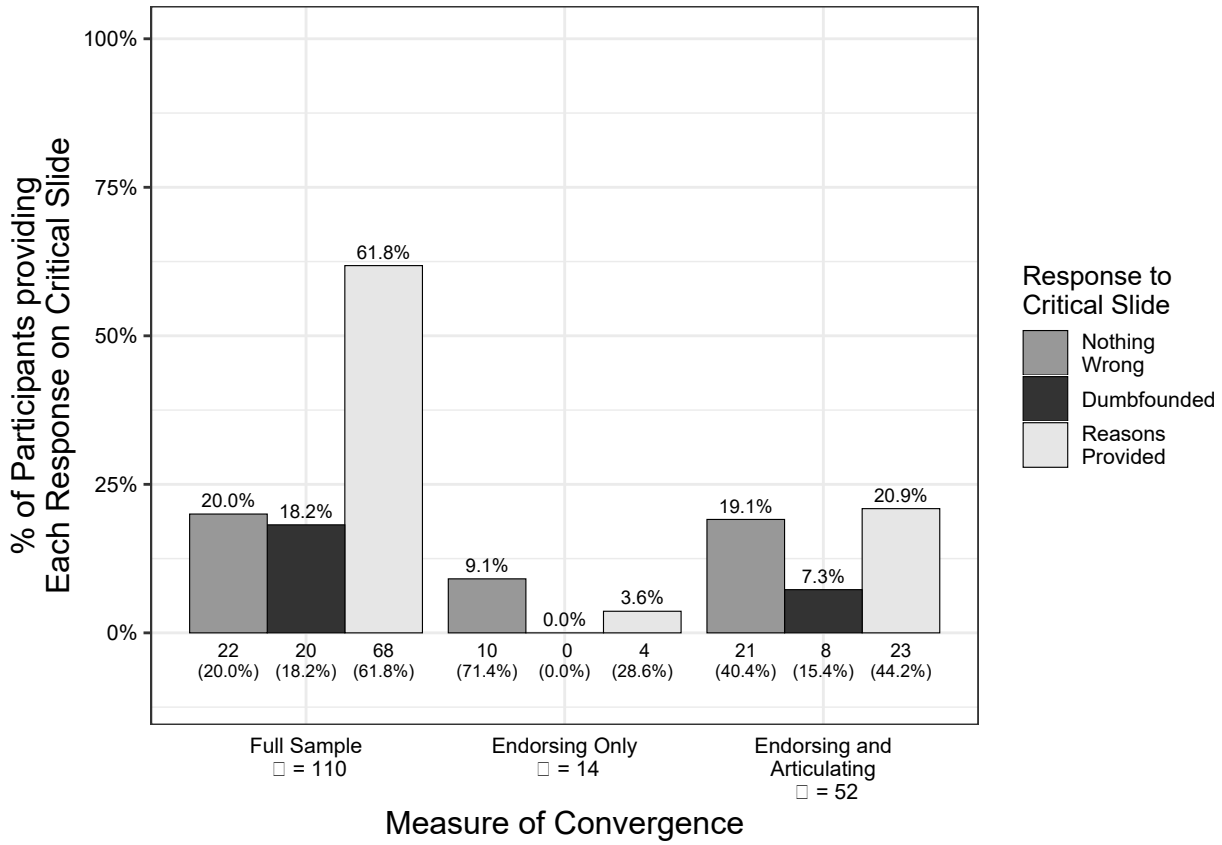

Figure 1. Responses to critical slide for the entire sample, and for each measure of convergence: (i) endorsing only, and (ii), endorsing and articulating; percentages of full sample displayed within plot, percentages of relevant sample displayed in parenthesis below the count.

22 participants who selected "There is nothing wrong" on the critical slide, also endorsed either the harm principle or the norm principle. The endorsing of these principles meant that these participants were excluded from analysis on the grounds they had a legitimate reason to rate the behavior as wrong. However, these participants did not rate the behavior as wrong. This demonstrates an inconsistency between the endorsing of the principles through targeted questions and statements and the apparent use of these principles as reasons guiding the participants' judgments. The endorsing only measure of convergence, using the targeted questions and statements developed by Royzman et al. (2015) led to participants being falsely excluded from analysis.

According to the revised criteria for exclusion, in which participants are only excluded from analysis if they were also able to articulate the principle that they endorsed, only one of the 22 participants (4.55%) who selected "There is nothing wrong" was excluded from analysis. The revised measure of convergence developed in Study 1 shows a reduced incidence of false exclusion of participants who selected "There is nothing wrong". This suggests that accounting for both the articulating and the endorsing of principles provides more accurate (though still not quite perfect) exclusion criteria.

The aim of Study 1 was to extend previous research by Royzman et al. (2015). They excluded participants from analysis based on their endorsing of either the harm principle or the norm principle through targeted questions/statements. Using these criteria for exclusion, they found minimal dumbfounded responding (1 participant from a sample of 53 (Royzman et al., 2015, p. 309)). It was hypothesised that their exclusion criteria were too broad, and that participants' endorsing of either principle does imply that participants can articulate the given principle. Revised criteria for exclusion were developed which accounted for both the endorsing and the articulation of either the harm principle or the norm principle. Our initial analysis replicated the findings of Royzman et al. (2015).

Further analysis, using the revised measure of convergence demonstrated considerably more consistency in the exclusion/inclusion of participants who selected "There is nothing wrong". These revised criteria identified eight (7.27% of the total sample of $N = 110$) participants as dumbfounded. Study 1 demonstrated inconsistency in the endorsing and articulation of the harm principle and the norm principle, and provided evidence for moral dumbfounding, however rates of dumbfounded responding were low, with the majority of participants (68; 61.82%) providing reasons for their judgments. A second study was devised to assess the consistency in the application of the harm principle across differing contexts, along with the endorsing, and articulation of the each principle.

3 | Study 2: Applying Moral Principles Across Contexts

In Study 1, we tested if participants could articulate the harm principle and the norm principle as identified by Royzman et al. (2015). In Study 2, we investigated the role of the harm principle in the making of judgments. Specifically, we examined if the harm principle can legitimately be said to be guiding the judgments of participants. This was done by assessing whether or not the harm principle is applied consistently across different contexts

Drawing on the research by Royzman et al. (2015), the harm principle may summarised as follows "it is wrong for two people to engage in an activity whereby harm may occur". Royzman et al. (2015) do not offer clarification on specific types of harm that may fall under this principle, it is therefore assumed that this is a generalised principle concerning any form of harm. According to the argument proposed by Royzman et al. (2015), participants' moral judgments are grounded in this principle, such that applying this principle to the *Incest* dilemma gives people a good reason to judge the behavior of Julie and Mark as wrong. If this general harm principle is to be considered as guiding participants' judgments, it should be consistently applied across differing contexts.

Study 2 tested if this was the case by including a set of targeted questions relating to the generalisation and application of the harm principle across different contexts (the rest of the materials were largely the same as those used in Study 1). We hypothesised that participants' responses to these targeted questions would reveal inconsistency in the application of the harm principle across differing contexts. Any exclusion criteria based on the harm principle should account for the endorsing of the principle (Royzman et al., 2015), articulating the principle (Study 1), and the application of the principle (Study 2).

3.1 | Method

3.1.1 | Participants and design. Study 2 was a frequency-based extension of Study 1. The aim was to investigate the prevalence of moral dumbfounding when controlling

for (a) the consistency with which people articulate and endorse the norm principle and the harm principle, and (b) the consistency with which people apply the norm principle principle. A combined sample of 111 (67 female, 44 male; $M_{\text{age}} = 34.23$, min = 19, max = 74, $SD = 11.42$) took part.

Sixty-one (36 female, 25 male; $M_{\text{age}} = 39.08$, min = 20, max = 74, $SD = 12.25$) were recruited through MTurk. Participation was voluntary and participants were paid 0.50 US dollars for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden). Fifty (31 female, 19 male; $M_{\text{age}} = 28.32$, min = 19, max = 48, $SD = 6.65$) were recruited through direct electronic correspondence. Participants in this sample were undergraduate students, postgraduate students, and alumni from Mary Immaculate College (MIC), and University of Limerick (UL). Participation was voluntary and participants were not reimbursed for their participation.

3.1.2 | Procedure and materials. Data were collected using an online questionnaire generated using Questback (Unipark, 2013). The questionnaire in Study 2 was the same as that presented in Study 1, with the inclusion of three additional targeted questions which aimed to assess the consistency with which participants generalise and apply the harm principle. The questions were: (a) "How would you rate the behavior of two people who engage in an activity that could potentially result in harmful consequences for either of them?"; (b) "Do you think boxing is wrong?"; (c) "Do you think playing contact team sports (e.g. rugby; ice-hockey; American football) is wrong?". Responses to (a) were recorded on a 7-point Likert scale (where, 1 = *Morally wrong*; 4 = *Neutral*; 7 = *Morally right*). Responses to (b) and (c) were recorded using a binary "Yes/No" option. These questions were presented sequentially, in randomised order. The randomised sequence was grouped as Block A. Similarly all slides and questions directly relating the moral scenario were grouped as Block B. Block B also included the targeted questions relating to the endorsing of the harm principle. The order of presentation of these blocks was randomised.

As with Study 1, the questionnaire opened with the information sheet, and the main body of the questionnaire could not be accessed until participants consented to continue. Once consent was given participants were asked a number of questions relating to basic demographics. They were then presented with the two targeted statements relating to the norm principle (in randomised order) and asked to select the statement they "identify with the most". Participants were then presented with either Block A (containing the targeted questions relating to the application of the harm principle) or Block B (containing the moral scenario, related questions, and targeted questions relating to the endorsing of the harm principle). Following this participants were presented with the second block. As in Study 1, the questionnaire ended with the MLQ (Steger et al., 2008); and CRSi7 (Huber & Huber, 2012).

3.2 | Results and Discussion

Seventy-nine of the total sample (N = 111; 71.17%) initially rated the behavior of Julie and Mark as wrong. An independent samples t-test revealed no difference in initial rating between the MTurk sample (M = 2.08, SD = 1.48), and the MIC sample, (M = 2.68, SD = 1.83), $t(93.31) = 1.86$, $p = .066$, $d = 0.36$. Sixty seven of the total sample, (N = 111; 60.36%) rated the behavior as wrong after viewing the counter-arguments and the critical slide. An independent samples t-test revealed a significant difference in revised rating between the MTurk sample, (M = 2.31, SD = 1.53), and the MIC sample, (M = 3, SD = 1.84), $t(95.40) = 2.11$, $p = .037$, $d = 0.41$. A paired samples t-test revealed a significant difference in rating of behavior from time one, initial rating, (M = 2.35, SD = 1.67), to time two, revised rating, (M = 2.62, SD = 1.54), $t(110)$ = -3.47, $p < .001$, $d = 0.16$. Further analysis revealed that although 15 participants changed their judgment, only two participants changed fully the valence of their judgment, changing their judgment from "wrong" to "right". Of the other changes in judgment, ten participants changed their judgment from "wrong" to "neutral"; two participants changed their judgment from "right"

to "neutral"; and one changed their judgment from "neutral" to "right". A chi-square test for independence revealed no significant association between time of judgment and valence of judgment made, $\chi^2(2, N = 222) = 3.40$, $p = .183$, $V = 0.12$.

Eighteen participants (16%) indicated that they had encountered the scenario before. As in Study 1, when asked to elaborate, participants provided anecdotes, or referred to previous readings/TV (either fiction or philosophy), 8 participants (7%) indicated that they had encountered it in a previous survey. The number of participants indicating previous experience with the scenario was higher than in Study 1 and as such the possibility that it may have confounded the results was investigated. An independent samples t-test revealed no difference in judgment between participants who had previously seen the scenario, ($M = 2.83$, $SD = 1.86$), and participants who had not previously seen the scenario, ($M = 2.26$, $SD = 1.62$), $t(22.31) = 1.23$, $p = .232$, $d = 0.35$. Furthermore, a chi-squared test for independence revealed no significant association between previous experience with the scenario and response to the critical slide, $\chi^2(2, N = 111) = 3.16$, $p = .206$, $V = 0.17$. These participants were not excluded from the analyses.

3.2.1 | Testing for order effects. The order of the blocks had no influence on the any of the responses of interest (see supplementary materials for details of analysis). Of the questions relating to the application of the harm principle, there were differences in responding to general question only ("How would you rate the behavior of two people who engage in an activity that could potentially result in harmful consequences for either of them?"). This question was more abstract than the two questions it appeared with, in which participants were asked to judge a named behavior (boxing or contact team sports). The description in the general question could apply to either of the named behaviors. Participants who responded to this question first rated the behavior as more wrong than participants who responded to it after reading one or both of the named behaviors. It seems likely that the named behaviors provided an example of a situation in which the behavior described in the general question may be acceptable, leading participants to respond more

favorably to the general question.

3.2.2 | Measuring dumbfounding. As in Study 1, participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Rates of each response to the critical slide are for the entire sample ($N = 111$) are displayed in Figure 1. Twenty one participants (18.92%) were initially identified as dumbfounded.[6] The exclusion criteria developed by Royzman et al. (2015; the endorsing of either principle) were applied, and this left a sample of 20 who were eligible for analysis. Two of these fully convergent participants selected the dumbfounded response. We then applied the revised criteria for exclusion (both articulating and endorsing either principle) developed in Study 1, and the number of participants eligible for analysis increased to 61. Of these, nine (14.75%) selected the dumbfounded response. Again this also led to a reduction in false exclusions, three 3 of the 36 (8.33) participants who selected "There is nothing wrong" were excluded by this measure.

The responses to the three targeted questions relating the application of the harm principle were analysed together. Only one participant was consistent in their application of the harm principle across all three targeted questions and this meant that only one participant was consistent in the application, articulation, and, endorsing of the harm principle (as measured by the open-ended responses and the targeted questions taken from Royzman et al. (2015)). This was combined with the exclusion criteria developed in Study 1 leaving a sample of 73 participants who were eligible for analysis. Ten (9.01% of the total sample) of these participants selected the dumbfounded response. The responses to the critical slide across all measures of convergence used are displayed in Figure 2.

3.2.3 | Consistency between endorsed principles and expressed judgments. As in Study 1, the initial criteria for exclusion (endorsing only) excluded a large proportion

[6] Unsupported declarations and tautological responses provided in the open-ended responses resulted in an additional six participants presenting as potentially dumbfounded; again, these participants are not identified as dumbfounded here.

of the participants who selected "There is nothing wrong"; 20 of the 36 participants (55.56%) who selected "There is nothing wrong" were excluded. When articulation of the principles was accounted for, only three (8.33%) of these 36 participants were excluded. This is higher than in Study 1 (one participant, 4.55% of those who selected "There is nothing wrong"), however in reducing the obvious false exclusion of participants who selected "There is nothing wrong" it remains an improvement on the original criteria. This suggests that accounting for participants' ability to articulate the principles endorsed provides a more accurate criteria for exclusion than accounting only for the endorsing of a given principle. Furthermore, when the applying of the harm principle was also accounted for, only one of the 36 participants who selected "There is nothing wrong" was excluded. The criteria for convergence developed here lead to greater consistency between a participant's eligibility for analysis and their judgment made than the original criteria described by Royzman et al. (2015).

Study 2 investigated the consistency with which people apply, articulate, and endorse the harm principle. Only one participant consistently applied, articulated, and endorsed the harm principle. As such, the harm principle as a basis for exclusion from analysis becomes practically redundant, and it seems unlikely that there is a generalised harm principle that underlies moral judgments (though does not rule out the possibility of more focused, content specific harm principles). The endorsing and articulation of the norm principle resulted in the exclusion of 37 participants. The degree to which the articulation or the endorsing of the norm principle may render participants ineligible for consideration as dumbfounded is unclear, this is discussed in more detail below. However, even if participants are excluded from analysis based on the norm principle, dumbfounded responding is still observed, with ten participants (13.70% of sample eligible for analysis; 9.01% of the total sample) selecting the admission of having no reason on the critical slide. As in Study 1, rates of observed dumbfounding are low, and providing reasons appears to be the preferred response, with more participants (54; 48.65%) providing reasons than selecting either of the other responses to the critical slide.

4 | Study 3: Replication and Extension

Studies 1 and 2 demonstrated that people do not consistently articulate and endorse the norm principle, or consistently articulate, endorse and apply the harm principle. Both studies found evidence of dumbfounding, however the exclusion of participants resulted in relatively small numbers of participants being eligible for analysis. As such we conducted a third study, an attempt to replicate Study 2, with a larger sample.

4.1 | Method

4.1.1 | Participants and design. Study 3 was a frequency-based replication of Study 2. The aim was to investigate the prevalence of moral dumbfounding when controlling for (a) the consistency with which people articulate and endorse the norm principle and the harm principle, and (b) the consistency with which people apply the norm principle principle. A total sample of 502 (287 female, 212 male; $M_{\text{age}} = 39.05$, min = 18, max = 81, $SD = 12.46$) took part. All participants were recruited through MTurk. Participation was voluntary and participants were paid 0.50 US dollars for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden).

4.1.2 | Procedure and materials. The materials and procedure were identical to Study 2.

4.2 | Results and Discussion

Three-hundred-and-seventy-nine of the total sample ($N = 502$; 75.50%) rated the behavior of Julie and Mark as wrong initially; and 357 participants, ($N = 502$; 71.12%) rated the behavior as wrong after viewing the counter-arguments and the critical slide. A paired samples t-test revealed a significant difference in rating of behavior from time one, initial rating, ($M = 2.21$, $SD = 1.72$), to time two, revised rating, ($M = 2.38$, $SD = 1.79$), $t(501) =$ -4.74, $p < .001$, $d = 0.10$. However a chi-square test for independence revealed no significant

association between time of judgment and valence of judgment made, $\chi^2(2, N = 1004) = 3.59$, $p = .166$, $V = 0.08$.[7]

4.2.1 | Testing for order effects. As in Study 2, the order of the blocks did influence on the any of the responses of interest, and the general harm question was the only question relating to the application of the harm principle that varied significantly with order (see supplementary materials for details of analysis). Again, it is likely that encountering a behaviour where harm may be acceptable (through the content of the other two questions), led participants to respond to the general question more favourably.

4.2.2 | Measuring dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. This option was selected by 88 participants (17.53% of the entire sample $N = 502$).[8]

The exclusion criteria developed by Royzman et al. (2015; the endorsing of either principle) were applied, and this left a sample of 84 who were eligible for analysis. Of these, 9 participants selected the dumbfounded response.

We then applied the exclusion criteria developed in Study 1 (both articulating and endorsing either principle), and the number of participants eligible for analysis increased to 294. Of these, 52 (17.69%) selected the dumbfounded response.

Finally, the exclusion criteria developed in Study 2 were applied, leaving a sample of 345 participants who were eligible for analysis; Sixty nine of whom (13.75% of the total

[7] Further analysis revealed that 42 participants changed their judgment, only seven participants changed fully the valence of their judgment, with five changing their judgment from "wrong" to "right", and two changing their judgement from "right" to "wrong". Of the other changes in judgment, twenty two participants changed their judgment from "wrong" to "neutral"; six participants changed their judgment from "right" to "neutral"; and four changed their judgment from "neutral" to "right".

[8] Unsupported declarations and tautological responses provided in the open-ended responses resulted in an additional 50 participants presenting as potentially dumbfounded; again, these participants are not identified as dumbfounded here.

sample) selected the dumbfounded response. The responses to the critical slide for the entire sample, and for each measure of convergence used are displayed in Figure 2.

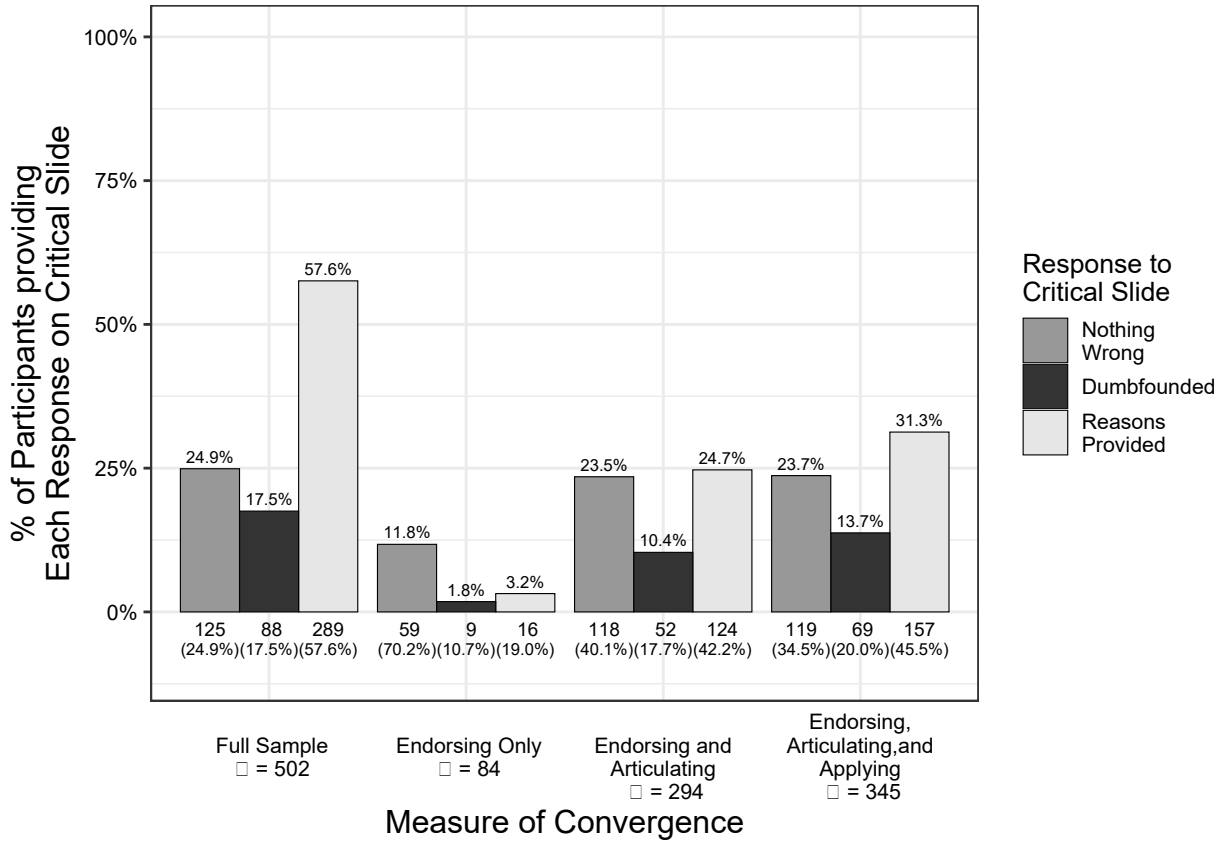

Figure 2. Responses to critical slide for the entire sample, and for each measure of convergence: (i) endorsing only, (ii) endorsing and articulating, and (iii), endorsing, articulating, and applying; percentages of full sample displayed within plot, percentages of relevant sample displayed in parenthesis below the count.

4.2.3 | Consistency between endorsed principles and expressed judgments. As in Studies 1 and 2, the exclusion criteria developed here resulted in fewer false exclusions. In the current study, the exclusion criteria developed by Royzman et al. (2015, endorsing only), led to 66 of the 125 participants who selected "There is nothing wrong" being excluded from analysis (52.80%). Conversely, applying the exclusion criteria developed in Study 1 resulted in seven of these 125 participants being excluded (5.60%); and the exclusion criteria from Study 2 resulted in six of these 125 participants being excluded (4.80%).

Further analysis, using the revised measure of convergence demonstrated considerably more consistency in the exclusion/inclusion of participants who selected "There is nothing wrong". These revised criteria identified sixty-nine (20% of the total eligible sample of $N = 345$) participants as dumbfounded. Study 1 provided evidence for moral dumbfounding and demonstrated inconsistency in the endorsing and articulation of the harm principle and the norm principle, a second study was devised to assess the consistency in the application of the harm principle across differing contexts, along with the endorsing, and articulation of the each principle. Study 3 replicated the findings of both Studies 1 and 2 with a larger sample. By applying our revised exclusion criteria, we found clear evidence for the existence of moral dumbfounding, though observed rates of dumbfounding were low, with the majority of participants (157; 45.51%) providing reasons.

The analyses of the individual difference variables are reported in the Supplementary Materials (Appendix D).

5 | General Discussion

The overarching goal of Studies 1, 2, and 3 was to re-assess the occurrence of moral dumbfounding. That is, we examined whether the judgments of dumbfounded participants can be attributed to moral principles based on their endorsing of these principles. This was done by assessing the consistency with which participants articulate and apply these moral principles. Royzman et al. (2015) argue that, if participants endorse a principle, their judgment can be attributed to that principle. They claimed that by attributing participants' judgments to particular principles in this way, moral dumbfounding can be eliminated. However, attributing judgments to reasons based on the endorsing of a related principle is problematic. Stronger evidence that a participant's judgment may be attributed to a given principle should account for (a) the participant's ability to articulate this principle, independent of a prompt; or (b) the consistency with with the participant applies the principle across differing contexts. Three studies were conducted to address these issues.

All three studies showed that participants do not consistently articulate principles that they may endorse. This inconsistency between the endorsing and articulation of principles that are purported to be governing moral judgments suggests that endorsing alone provides a poor measure of whether these principles directly underpin a given judgment. In these cases participants' judgments were not attributed to these principles, and evidence for dumbfounding was found, though rates of dumbfounding were quite low. Studies 2 and 3 demonstrated that people do not consistently apply the harm principle across different contexts. This poses a challenge to the argument that the judgments of dumbfounded participants can be attributed to the harm principle (e.g., Royzman et al., 2015; see also Gray et al., 2014; Jacobson, 2012). Our studies showed evidence for dumbfounding. Despite the low rates of dumbfounding observed, the consistency across all three studies provides some evidence that dumbfounded responding may indeed be indicative of a state of dumbfoundedness, rather than being entirely attributed to features of the experimental design.

5.1 | The Norm Principle and Unsupported Declarations

In all three studies, unsupported declarations were coded as an articulation of the norm principle, and therefore not taken as dumbfounded responses. However, in previous work, we identified parallels between the providing of unsupported declarations and the providing of admissions of not having reasons (similar proportion of time spent (a) smiling/laughing, (b) in silence; see McHugh et al., 2017). There is also a strong theoretical case for the inclusion of unsupported declarations as dumbfounded responses. Propositional beliefs/deontological judgments may be viewed as habitual/model-free intuitions (e.g., Crockett, 2013; Cushman, 2013b). The reasons for these judgments are independent of the intuition. Stating the content of the intuition, is not the same as providing a reason for the intuition. Royzman et al. (2015) argue that endorsing the propositional belief is sufficient evidence of that belief playing an influential role in relevant judgments, however, this is

holding participants to a different standard. There is a difference between having a reason for an intuition/propositional belief and claiming the direct basis for a judgment is an associated propositional belief. In view of this, it is possible that by not including unsupported declarations or tautological reasons as dumbfounded responses, the rates of dumbfounding reported here are not representative of the phenomenon, providing instead an overly conservative estimate. However, even according to this stricter measure adopted here, evidence for dumbfounding was found.

5.2 | Consistency Between Endorsed Principles and Expressed Judgments

The most convincing evidence that the exclusion criteria developed in these studies are more accurate than the criteria proposed by Royzman et al. (2015) is the greater consistency between valence of judgment and eligibility for analysis. Participants' eligibility for analysis is determined by whether or not their judgment can be attributed to either the harm principle or the norm principle. If a participant's judgment can be attributed to a given principle, this participant is deemed to have a reason for their judgment and they cannot be identified as dumbfounded (rendering them ineligible for analysis). In order for a judgment to legitimately be attributed to a particular principle, it is necessary that the valence of the judgment is consistent with what is predicted by the application of that principle. In the case of both principles, applying either the harm principle or the norm principle (as described by Royzman et al., 2015) results in the behavior being judged as wrong. This means that the judgments of participants who selected "There is nothing wrong" cannot be attributed to either principle. Any participants who are excluded from analysis but selected "There is nothing wrong", are clearly identifiable as being falsely excluded from analysis such that this may be used as a measure of the relative accuracy of the different exclusion criteria employed.

According to Royzman et al. (2015), a participant's judgment can be attributed to a given principle if they endorse this principle. However, in each of the studies reported here,

excluding participants based on the endorsing of a principle resulted in over half of the participants who selected "There is nothing wrong" to be falsely excluded from analysis; participants' judgments were incorrectly attributed to either the harm principle or the norm principle (12 of the 22 participants who selected "There is nothing wrong" in Study 1 were falsely excluded 54.55%; 20 of the 36 participants who selected "There is nothing wrong" in Study 2 were falsely excluded 55.56%; and 66 of the 125 participants who selected "There is nothing wrong" in Study 3 were falsely excluded 52.80%). This suggests that the endorsing of a principle is a flawed indicator of the degree to which the principle is guiding participants' judgments.

We made two changes to the exclusion criteria that aimed to reduce the numbers of participants being falsely excluded from analysis. We hypothesised that providing participants with an opportunity to articulate the reasons for their judgment would more accurately identify the principles that guided participants' judgments than their endorsing of particular principles. This was found to be the case; in Study 1, only one of the 22 participants who selected "There is nothing wrong" was falsely excluded from analysis; in Study 2 only three of the 36 participants who selected "There is nothing wrong" were falsely excluded from analysis; and in Study 3 seven of the 125 participants who selected "There is nothing wrong" were falsely excluded from analysis. Taking participants' articulating of the reasons for their judgments into account reduced measurable rate of false exclusion from 54.55% to 4.55% in Study 1; 55.56% to 8.33% in Study 2; and 52.80% to 5.60% in Study 3. Furthermore, in Studies 2 and 3, with specific reference to the harm principle, we hypothesised that assessing the degree to which people's judgments could be attributed to the harm principle would be related to whether or not they apply the harm principle across different contexts. Again this was found to be the case, as evidenced by a further reduction in the measurable rate of false exclusion from 8.33% (3/36) to 2.78% (1/36) in Study 2, and from 5.60% (7/125) to 4.80% (6/125) in Study 3.

5.3 | Implications

The existence of moral dumbfounding and the associated support for intuitionist theories of moral judgment (e.g. Cushman et al., 2010; Haidt, 2001; Hauser, Young, & Cushman, 2008; Prinz, 2005; see also Crockett, 2013; Cushman, 2013b; Greene, 2008, 2013) has been questioned in recent years. The majority of these challenges are theoretical (e.g., Jacobson, 2012; Sneddon, 2007; Wielenberg, 2014). The work of Gray et al. (2014), appeared to give some empirical weight to these challenges, while Royzman et al. (2015) extended these challenges to the dumbfounding paradigm specifically. We conducted three studies addressing specific methodological limitations associated with the work by Royzman et al. (2015). Their criteria for exclusion were found to be overly liberal, as evidenced by the high rates of false exclusion of participants who selected "There is nothing wrong". and evidence for dumbfounding was found. Adopting the more rigorous exclusion criteria developed here led to a reduction in the false exclusion of participants. In using these criteria, evidence for dumbfounding was found, and the explanation of dumbfounded responding proposed by Royzman et al. (2015) was not supported.

Our findings provide further evidence that the distinction between implicit and explicit cognition (e.g., Bonner & Newell, 2010; Evans, 2003, 2006, 2008; Evans & Over, 2013; Reber, 1989) extends to the moral domain. It has long been known that people have poor introspective awareness of how judgments are made (e.g., Nisbett & Wilson, 1977) and it appears that in some cases this may also be true for moral judgments.

5.4 | Limitations and Future Directions

The research we present here consists of three studies with a combined sample of $N = 723$, from MTurk ($N = 621$) and third level institutions ($N = 102$). Follow-up studies should investigate the phenomenon with larger and more diverse samples. Such follow-up work may inform investigations into the influence of cultural and societal norms on the prevalence of moral dumbfounding. Previous work by Haidt and Hersh (2001) provides suggestive evidence

that political orientation may influence a person's susceptibility to moral dumbfounding; furthermore, there is some evidence to indicate that cultural and socio-economic factors may also play a role (Haidt et al., 1993). Future research should draw on the methods developed here and by both McHugh et al. (2017) and Royzman et al. (2015) to investigate these influences further.

The procedures we used were very similar across both studies. They were also very similar to those used by McHugh et al. (2017) and by Royzman et al. (2015). A more rigourous test of moral dumbfounding should employ a variety of methods. We recommend that future research develops a broader selection of "dumbfounding scenarios", and investigate the feasibility of alternative procedures that may elicit dumbfounding.

The role of social pressure and conversational norms in the emergence of moral dumbfounding is not well understood. The studies described here were conducted using online surveys and therefore there was no immediate social pressure on participants to either appear consistent or to conform to conversational norms. Furthermore, the argument proposed by Royzman et al. (2015), that participants' judgment are grounded in reasons (harm-based/norm-based) and that they drop these reasons in response to social pressure is not supported by the evidence presented here; harm-based/norm based reasons were not consistently articulated or applied by participants in these studies. It is apparent then that dumbfounded responding cannot be attributed to social pressure alone. The processes by which we make moral judgments also give rise to moral dumbfounding. This means that isolating the underlying mechanisms that give rise to moral dumbfounding may contribute to our overall understanding of the making of moral judgments.

6 | Conclusion

Based on three studies we conclude: moral dumbfounding seems to be real, if not as widespread as initial reports might suggest (Haidt, 2001; Haidt et al., 2000; Haidt & Hersh,

886 2001). By reconsidering approaches of earlier research, our procedures found clear evidence
887 for this phenomenon. People are not always able to justify their moral judgments. Indeed, in
888 our studies, between 13% and 18% of people showed dumbfounding. Gaining insights into
889 the occurrence and underlying processes equips society with the tools to confront and reduce
890 dumbfounding. Further research in the area may inform improvements in the conduct of
891 public debate, particularly in relation to polarizing issues. Perhaps in the future, the
892 influence dumbfounding in public discourse and public policy (e.g., MacNab, 2016; Sim,
893 2016) will be reduced or even eliminated.

894 **7 | Data Accessibility Statement**

895 All participant data, and analysis scripts can be found on this paper's project page on
896 the Open Science Framework at https://osf.io/m4ce7/.

897 All statistical analysis was conducted using R (Version 3.6.0; R Core Team, 2017) and
898 the R-packages *afex* (Version 0.23.0; Singmann, Bolker, & Westfall, 2015), *boot* (Version
899 1.3.23; Davison & Hinkley, 1997), *Cairo* (Version 1.5.10; Urbanek & Horner, 2019), *car*
900 (Version 3.0.3; Fox & Weisberg, 2011; Fox, Weisberg, & Price, 2018), *carData* (Version 3.0.2;
901 Fox et al., 2018), *citr* (Version 0.3.2; Aust, 2016), *DescTools* (Version 0.99.28; et mult. al.,
902 2019), *desnum* (Version 0.1.1; McHugh, 2017), *devtools* (Version 2.1.0; Wickham & Chang,
903 2017), *emmeans* (Version 1.4; Lenth, 2019), *extrafont* (Version 0.17; Chang, 2014), *foreign*
904 (Version 0.8.72; R Core Team, 2018), *Formula* (Version 1.2.3; Zeileis & Croissant, 2010),
905 *ggplot2* (Version 3.2.1; Wickham, 2009), *koRpus* (Version 0.11.5; Michalke, 2018a, 2019),
906 *koRpus.lang.en* (Version 0.1.3; Michalke, 2019), *lme4* (Version 1.1.21; Bates, Mächler, Bolker,
907 & Walker, 2015), *lmtest* (Version 0.9.37; Zeileis & Hothorn, 2002), *lsmeans* (Lenth, 2016), *lsr*
908 (Version 0.5; Navarro, 2015), *MASS* (Version 7.3.51.4; Venables & Ripley, 2002a), *Matrix*
909 (Version 1.2.17; Bates & Maechler, 2017), *metap* (Version 1.1; Dewey, 2017), *mlogit* (Version
910 1.0.1; Croissant, 2013), *nnet* (Version 7.3.12; Venables & Ripley, 2002b), *papaja* (Version
911 0.1.0.9842; Aust & Barth, 2018), *plyr* (Version 1.8.4; Wickham, 2011), *powerMediation*

(Version 0.2.9; Qiu, 2018), *pwr* (Version 1.2.2; Champely, 2018), *QuantPsyc* (Version 1.5; Fletcher, 2012), *reshape2* (Version 1.4.3; Wickham, 2007), *scales* (Version 1.0.0; Wickham, 2016), *sjstats* (Version 0.17.5; Lüdecke, 2018), *sylly* (Version 0.1.5; Michalke, 2018b), *tibble* (Version 2.1.3; Müller & Wickham, 2017), *usethis* (Version 1.5.1; Wickham & Bryan, 2019), *VGAM* (Version 1.1.1; Yee & Wild, 1996; Yee, 2010, 2013; Yee & Hadi, 2014; Yee, Stoklosa, & Huggins, 2015), *wordcountaddin* (Version 0.3.0.9000; Marwick, 2019), and *zoo* (Version 1.8.6; Zeileis & Grothendieck, 2005).

References

Aust, F. (2016). *Citr: 'RStudio' Add-in to Insert Markdown Citations*. Retrieved from https://CRAN.R-project.org/package=citr

Aust, F., & Barth, M. (2018). *Papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bates, D., & Maechler, M. (2017). *Matrix: Sparse and Dense Matrix Classes and Methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Bauman, C. W., McGraw, A. P., 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. doi:10.1111/spc3.12131

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition*, *38*(2), 186–196. doi:10.3758/MC.38.2.186

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas. *Psychological Science*, *29*(7), 1084–1093. doi:10.1177/0956797617752640

Brand, C. (2016). *Dual-Process Theories in Moral Psychology: Interdisciplinary Approaches to Theoretical, Empirical and Practical Considerations*. Springer.

Cameron, C. D., Payne, B. K., & Doris, J. M. (2013). Morality in high definition: Emotion differentiation calibrates the influence of incidental disgust on moral judgments.

Journal of Experimental Social Psychology, *49*(4), 719–725. doi:10.1016/j.jesp.2013.02.014

Champely, S. (2018). *Pwr: Basic Functions for Power Analysis*. Retrieved from https://CRAN.R-project.org/package=pwr

Chang, W. (2014). *Extrafont: Tools for using fonts*. Retrieved from https://CRAN.R-project.org/package=extrafont

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: A moral dilemma validation study. *Emotion Science*, *5*, 607. doi:10.3389/fpsyg.2014.00607

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264. doi:10.1016/j.neubiorev.2012.02.008

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366. doi:10.1016/j.tics.2013.06.005

Croissant, Y. (2013). *Mlogit: Multinomial logit model*. Retrieved from https://CRAN.R-project.org/package=mlogit

Cushman, F. A. (2013a). Action, Outcome, and Value A Dual-System Framework for Morality. *Personality and Social Psychology Review*, *17*(3), 273–292. doi:10.1177/1088868313495594

Cushman, F. A. (2013b). The role of learning in punishment, prosociality, and human uniqueness. In K. Sterelny, B. Calcott, & B. Fraser (Eds.), *Signaling, Commitment and Emotion, Vol. 2: Psychological and Environmental Foundations of Cooperation*. MIT Press.

965 Cushman, F. A., Young, L., & Greene, J. D. (2010). Multi-system Moral Psychology. In J.
966 M. Doris (Ed.), *The Moral Psychology Handbook* (pp. 47–71). Oxford; New York:
967 Oxford University Press.

968 Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*.
969 Cambridge: Cambridge University Press. Retrieved from
970 http://statwww.epfl.ch/davison/BMA/

971 Dewey, M. (2017). *Metap: Meta-analysis of significance values*.

972 Dickinson, D. L., & Masclet, D. (2018). *Using Ethical Dilemmas to Predict Antisocial*
973 *Choices with Real Payoff Consequences: An Experimental Study* (SSRN Scholarly
974 Paper No. ID 3205879). Rochester, NY: Social Science Research Network. Retrieved
975 from https://papers.ssrn.com/abstract=3205879

976 et mult. al., A. S. (2019). *DescTools: Tools for Descriptive Statistics*. Retrieved from
977 https://cran.r-project.org/package=DescTools

978 Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in*
979 *Cognitive Sciences*, *7*(10), 454–459. doi:10.1016/j.tics.2003.08.012

980 Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and
981 evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378–395.
982 doi:10.3758/BF03193858

983 Evans, J. S. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social
984 Cognition. *Annual Review of Psychology*, *59*(1), 255–278.
985 doi:10.1146/annurev.psych.59.103006.093629

986 Evans, J. S. B. T., & Over, D. E. (2013). *Rationality and Reasoning*. Psychology Press.

987 Fine, C. (2006). Is the emotional dog wagging its rational tail, or chasing it? *Philosophical*

Explorations, *9*(1), 83–98. doi:10.1080/13869790500492680

Flanagan, O., Sarkissian, H., & Wong, D. (2008). Naturalizing Ethics. In W. Sinnott-Armstrong (Ed.), *Moral Psychology Volume 1: The evolution of morality adaptations and innateness* (pp. 1–26). Cambridge, Mass.; London, England: The MIT press.

Fletcher, T. D. (2012). *QuantPsyc: Quantitative Psychology Tools*. Retrieved from https://CRAN.R-project.org/package=QuantPsyc

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (Second.). Thousand Oaks CA: Sage. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Fox, J., Weisberg, S., & Price, B. (2018). *carData: Companion to Applied Regression Data Sets*. Retrieved from https://CRAN.R-project.org/package=carData

Gray, K. J., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*(4), 1600–1615. doi:10.1037/a0036149

Greene, J. D. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong, *Moral Psychology Volume 3: The neurosciences of morality: Emotion, brain disorders, and development* (pp. 35–79). Cambridge (Mass.): the MIT press.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*.

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 (New York, N.Y.)*, *293*(5537), 2105–2108. doi:10.1126/science.1062872

Guglielmo, S. (2018). Unfounded dumbfounding: How harm and purity undermine evidence

for moral dumbfounding. *Cognition*, *170*, 334–337.
doi:10.1016/j.cognition.2017.08.002

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.
doi:10.1037/0033-295X.108.4.814

Haidt, J., & Björklund, F. (2008). Social Intuitionists Answer Six Questions about Moral Psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology Volume 2, The cognitive science of morality: Intuition and diversity* (pp. 181–217). London: MIT.

Haidt, J., Björklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished Manuscript, University of Virginia.*

Haidt, J., & Hersh, M. A. (2001). Sexual Morality: The Cultures and Emotions of Conservatives and Liberals. *Journal of Applied Social Psychology*, *31*(1), 191–221.
doi:10.1111/j.1559-1816.2001.tb02489.x

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*(4), 613–628.
doi:10.1037/0022-3514.65.4.613

Hauser, M. D., Young, L., & Cushman, F. A. (2008). Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions. In W. Sinnott-Armstrong (Ed.), *Moral psychology Volume 2, The cognitive science of morality: Intuition and diversity* (pp. 107–155). London: MIT.

Huber, S., & Huber, O. W. (2012). The Centrality of Religiosity Scale (CRS). *Religions*, *3*(3), 710–724. doi:10.3390/rel3030710

Jacobson, D. (2012). Moral Dumbfounding and Moral Stupefaction. In *Oxford studies in*

normative ethics (Vol. 2, p. 289).

Johnson-Laird, P. N. (2006). *How we reason*. Oxford ; New York: Oxford University Press.

Kennett, J., & Fine, C. (2009). Will the Real Moral Judgment Please Stand Up? *Ethical Theory and Moral Practice*, *12*(1), 77–96. doi:10.1007/s10677-008-9136-4

Kohlberg, L. (1969). *Stages in the development of moral thought and action*. New York: Holt, Rinehart & Winston.

Kohlberg, L. (1971). *From is to Ought: How to Commit the Naturalistic Fallacy and Get Away with it in the Study of Moral Development*.

Lenth, R. (2019). *Emmeans: Estimated Marginal Means, aka Least-Squares Means*. Retrieved from https://CRAN.R-project.org/package=emmeans

Lenth, R. V. (2016). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. doi:10.18637/jss.v069.i01

Lüdecke, D. (2018). *Sjstats: Statistical Functions for Regression Models*. Retrieved from https://CRAN.R-project.org/package=sjstats

MacNab, S. (2016). MSPs to consider "abhorrent" call to legalise incest. *The Scotsman*. Retrieved from http://www.scotsman.com/news/politics/msps-to-consider-abhorrent-call-to-legalise-incest-1-4009185

Marwick, B. (2019). *Wordcountaddin: Word counts and readability statistics in R markdown documents*.

McHugh, C. (2017). *Desnum: Creates some useful functions*. Retrieved from https://github.com/cillianmiltown/R_desnum

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2017). Searching for Moral

Dumbfounding: Identifying Measurable Indicators of Moral Dumbfounding. *Collabra: Psychology*, *3*(1). doi:10.1525/collabra.79

Mercier, H. (2016). The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700. doi:10.1016/j.tics.2016.07.001

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. doi:10.1017/S0140525X10000968

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.

Michalke, M. (2018a). *koRpus: An R Package for Text Analysis*. Retrieved from https://reaktanz.de/?c=hacking&s=koRpus

Michalke, M. (2018b). *Sylly: Hyphenation and Syllable Counting for Text Analysis*. Retrieved from https://reaktanz.de/?c=hacking&s=sylly

Michalke, M. (2019). *koRpus.Lang.En: Language Support for 'koRpus' Package: English*. Retrieved from https://reaktanz.de/?c=hacking&s=koRpus

Mustonen, A.-M., Paakkonen, T., Ryökäs, E., & Nieminen, P. (2017). Abortion debates in Finland and the Republic of Ireland: Textual analysis of experiential thinking and argumentation in parliamentary and layperson discussions. *Reproductive Health*, *14*(1), 163. doi:10.1186/s12978-017-0418-y

Müller, K., & Wickham, H. (2017). *Tibble: Simple Data Frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Narvaez, D. (2005). The neo-Kohlbergian tradition and beyond: Schemas, expertise, and character. In G. Carlo & C. Pope-Edwards (Eds.), *Nebraska symposium on motivation* (Vol. 51, p. 119).

Navarro, D. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5)*. Adelaide, Australia: University of Adelaide. Retrieved from http://ua.edu.au/ccs/teaching/lsr

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231. Retrieved from http://psycnet.apa.org/journals/rev/84/3/231/

Plunkett, D., & Greene, J. D. (2019). Overlooked Evidence and a Misunderstanding of What Trolley Dilemmas Do Best: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, 0956797619827914. doi:10.1177/0956797619827914

Prinz, J. J. (2005). Passionate Thoughts: The Emotional Embodiment of Moral Concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking* (pp. 93–114). Cambridge University Press.

Qiu, W. (2018). *powerMediation: Power/Sample Size Calculation for Mediation Analysis*. Retrieved from https://CRAN.R-project.org/package=powerMediation

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

R Core Team. (2018). *Foreign: Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...* Retrieved from https://CRAN.R-project.org/package=foreign

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. doi:10.1037/0096-3445.118.3.219

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision Making*, *10*(4), 296–313.

Rozin, P., Haidt, J., MacCauley, C., McKay, D., & Olatunji, B. O. (2008). Disgust: The body and soul emotion in the 21st century. In *Disgust and its disorders* (pp. 9–29). American Psychological Association.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574–586. doi:10.1037/0022-3514.76.4.574

Sim, P. (2016, January 26). MSPs throw out incest petition. *BBC News: Scotland Politics*. Retrieved from http://www.bbc.com/news/uk-scotland-scotland-politics-35401195

Singmann, H., Bolker, B., & Westfall, J. (2015). *Afex: Analysis of Factorial Experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Sneddon, A. (2007). A Social Model of Moral Dumbfounding: Implications for Studying Moral Reasoning and Moral Judgment. *Philosophical Psychology*, *20*(6), 731–748. doi:10.1080/09515080701694110

Steger, M. F., Kashdan, T. B., Sullivan, B. A., & Lorentz, D. (2008). Understanding the Search for Meaning in Life: Personality, Cognitive Style, and the Dynamic Between Seeking and Experiencing Meaning. *Journal of Personality*, *76*(2), 199–228. doi:10.1111/j.1467-6494.2007.00484.x

Stepniak, D. (1995). Televising Court Proceedings Forum: Televising Court Proceedings. *University of New South Wales Law Journal*, (2), 488–492. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/swales18&i=501

Todd, P. M., & Gigerenzer, G. (Eds.). (2012). *Ecological rationality: Intelligence in the world*. Oxford ; New York: Oxford University Press.

Topolski, R., Weaver, J. N., Martin, Z., & McCoy, J. (2013). Choosing between the emotional dog and the rational pal: A moral dilemma with a tail. *Anthrozoös*, *26*(2), 253–263. doi:10.2752/175303713X13636846944321

Triskiel, J. (2016). Psychology Instead of Ethics? Why Psychological Research Is Important but Cannot Replace Ethics. In C. Brand (Ed.), *Dual-Process Theories in Moral Psychology: Interdisciplinary Approaches to Theoretical, Empirical and Practical Considerations* (pp. 77–98). Springer.

Unipark, Q. (2013). *QuestBack Unipark.(2013)*.

Urbanek, S., & Horner, J. (2019). *Cairo: R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (x11 and win32) output*. Retrieved from https://CRAN.R-project.org/package=Cairo

Venables, W. N., & Ripley, B. D. (2002a). *Modern Applied Statistics with S* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Venables, W. N., & Ripley, B. D. (2002b). *Modern Applied Statistics with S* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of*

Statistical Software, *40*(1), 1–29. Retrieved from http://www.jstatsoft.org/v40/i01/

Wickham, H. (2016). *Scales: Scale Functions for Visualization*. Retrieved from https://CRAN.R-project.org/package=scales

Wickham, H., & Bryan, J. (2019). *Usethis: Automate Package and Project Setup*. Retrieved from https://CRAN.R-project.org/package=usethis

Wickham, H., & Chang, W. (2017). *Devtools: Tools to Make Developing R Packages Easier*. Retrieved from https://CRAN.R-project.org/package=devtools

Wielenberg, E. J. (2014). *Robust Ethics: The Metaphysics and Epistemology of Godless Normative Realism*. OUP Oxford.

Yee, T. W. (2010). The VGAM Package for Categorical Data Analysis. *Journal of Statistical Software*, *32*(10), 1–34. Retrieved from http://www.jstatsoft.org/v32/i10/

Yee, T. W. (2013). Two-parameter reduced-rank vector generalized linear models. *Computational Statistics and Data Analysis*. Retrieved from http://ees.elsevier.com/csda

Yee, T. W., & Hadi, A. F. (2014). Row-column interaction models, with an R implementation. *Computational Statistics*, *29*(6), 1427–1445.

Yee, T. W., Stoklosa, J., & Huggins, R. M. (2015). The VGAM Package for Capture-Recapture Data Using the Conditional Likelihood. *Journal of Statistical Software*, *65*(5), 1–33. Retrieved from http://www.jstatsoft.org/v65/i05/

Yee, T. W., & Wild, C. J. (1996). Vector Generalized Additive Models. *Journal of Royal Statistical Society, Series B*, *58*(3), 481–493.

Zeileis, A., & Croissant, Y. (2010). Extended Model Formulas in R: Multiple Parts and

Multiple Responses. *Journal of Statistical Software*, *34*(1), 1–13. doi:10.18637/jss.v034.i01

Zeileis, A., & Grothendieck, G. (2005). Zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, *14*(6), 1–27. doi:10.18637/jss.v014.i06

Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News*, *2*(3), 7–10. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Appendices

Appendix A: Moral Scenario

Julie and Mark, who are brother and sister, are travelling together in France. They are both on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy it, but they decide not to do it again. They keep that night as a special secret between them, which makes them feel even closer to each other (Haidt et al., 2000).

Appendix B: Sample Statements to Challenge Judgement

- Do you not agree that any concerns regarding reproductive complications are eased by their using of two forms of contraception?
- And do you accept that they are both consenting adults, and that they both consented and enjoyed it?
- And do you concede that nobody else was affected by their actions?

Appendix C: Post Discussion Questionnaire

How sure were you about your judgement?

1 2 3 4 5 6 7

Not at all Extremely sure

_____

How much did you change your mind?

1 2 3 4 5 6 7

Not at all Extremely

_____

How confused were you?

1 2 3 4 5 6 7

Not at all Extremely confused

_____

How irritated were you?

1 2 3 4 5 6 7

Not at all Extremely irritated

_____

How much was your judgement based on reason?

1 2 3 4 5 6 7

Not at all Extremely

How much was your judgement based on “gut” feeling?

1 2 3 4 5 6 7

Not at all Extremely

Appendix D: Supplementary Materials: Additional Analyses

Study 2: Test for Order Effects

Recall that the questions were blocked for randomisation. Tests for effects of the order of the blocks revealed no difference in initial rating, $t(106.87) = -1.64$, $p = .104$, $d = 0.29$; no difference in responding to the critical slide, $\chi^2(2, N = 111) = 4.76$, $p = .093$, $V = 0.21$; and no difference in response to the generic potential harm question ("How would you rate the behavior of two people who engage in an activity that could potentially result in harmful consequences for either of them?"), $t(85.40) = -1.02$, $p = .312$, $d = 0.20$. A chi-squared test for independence revealed no significant association between order of blocks and judgments of boxing, $\chi^2(1, N = 111) = 2.86$, $p = .091$, $V = 0.16$, or the question regarding contact team sports, $\chi^2(1, N = 111) = 0.19$, $p = .660$, $V = 0.04$.

The order of the questions regarding the application of the harm principle was also randomised. A one-way ANOVA revealed a significant difference in responses to the question "How would you rate the behavior of two people who engage in an activity that could potentially result in harmful consequences for either of them?" (1 = *Extremely wrong*; 4 = *Neutral*; 7 = *Extremely right*) depending on when it was presented $F(2, 109) = 4.757$ $p = .010$, partial $\eta^2 = .080$. Tukey's post-hoc pairwise revealed that, when this question was responded to first, participants ratings were significantly lower ($M = 2.80$, $SD = 1.43$) than when it was responded to second ($M = 3.57$, $SD = 1.21$), $p = .040$, or third ($M = 3.67$, $SD = 1.31$) , $p = .014$; and there was no difference in responding to this question second ($M = 3.57$, $SD = 1.21$) or third ($M = 3.67$, $SD = 1.31$), $p = .932$.

A chi-squared test for independence revealed no significant association between order these questions and responses to the question "Do you think boxing is wrong?", $\chi^2(2, N = 111) = 4.88$, $p = .087$, $V = 0.21$. Similarly, a chi-squared test for independence revealed a significant association between order these questions and responses to the question "Do you think playing contact team sports (e.g. rugby; ice-hockey; American football) is wrong?",

$\chi^2(2,\ N = 111) = 1.79,\ p = .409,\ V = 0.13.$

Study 3: Test for Order Effects

As in Study 2, the questions were blocked for randomisation. Tests for effects of the order of the blocks revealed no difference in initial rating, $t(465.55) = 1.76,\ p = .079,\ d = 0.16$; no difference in responding to the critical slide, $\chi^2(2,\ N = 502) = 1.12,\ p = .570,\ V = 0.05$; no difference in responses to the generic potential harm question, $t(443.45) = 0.99,\ p = .322,\ d = 0.09$. no association with judgments of boxing, $\chi^2(1,\ N = 502) = 1.03,\ p = .310,\ V = 0.05$, or the question regarding contact team sports, $\chi^2(1,\ N = 502) = 1.15,\ p = .283,\ V = 0.10$, depending on order of blocks.

Regarding the three questions assessing the application of the harm principle, a one-way ANOVA revealed a significant difference in responses to the generic potential harm question depending on when it was presented $F(2,\ 499) = 23.512\ p < .001$, partial $\eta^2 = .086$. Tukey's post-hoc pairwise revealed that, when this question was responded to first, participants ratings were significantly lower ($M = 2.60,\ SD = 1.46$) than when it was responded to second ($M = 3.50,\ SD = 1.44$), $p < .001$, or third ($M = 3.47,\ SD = 1.20$) , $p < .001$; and there was no difference in responding to this question second ($M = 3.50,\ SD = 1.44$) or third ($M = 3.47,\ SD = 1.20$), $p = .983$. As in Study 2, it seems likely that the named behaviours in the other questions provide an example of potential harm that is acceptable, leading to a more favourable response to this more abstract question. There was no significant association between question order and responses to the question "Do you think boxing is wrong?", $\chi^2(2,\ N = 502) = 1.12,\ p = .570,\ V = 0.05$; or "Do you think playing contact team sports (e.g. rugby; ice-hockey; American football) is wrong?", $\chi^2(1,\ N = 502) = 1.03,\ p = .310,\ V = 0.05$.

Study 3: Individual Differences

A series of logistic regressions were conducted to investigate if dumbfounded responding was related to any of the individual difference variables Religiosity (as measured by CRSi7 Huber and Huber 2012), or Meaning in Life (Presence and Search, measured using MLQ Steger et al. 2008). We first report the results for each variable individually, followed by the combined model.

Religiosity. The overall mean Religiosity score was $M = 2.57$, $SD = 1.17$. The mean religiosity scores for participants depending on response to the critical slide were as follows: $M = 2.84$, $SD = 1.17$ for participants who provided reasons, $M = 2.42$, $SD = 1.11$ for participants who were dumbfounded, and $M = 2.28$, $SD = 1.12$ for participants who selected "There is nothing wrong".

A multinomial logistic regression revealed a statistically significant association between Religiosity and response to the critical slide, $\chi^2(2,\ N = 502) = 17.38$, $p < .001$, The observed power was 0.97. Religiosity explained approximately 2.40% (McFadden R square) of the variance in responses to the critical slide. Participants with higher religiosity scores were significantly more likely to provide reasons than to present as dumbfounded, Wald = 6.14, $p = .013$, odds ratio = 0.73, 95% CI [0.57, 0.94], or select "There is nothing wrong" Wald = 15.24, $p < .001$, odds ratio = 0.65, 95% CI [0.53, 0.81]. See Figure 3.

Meaning in Life (Presence). The overall mean Meaning in Life (Presence) score was $M = 4.74$, $SD = 1.66$. The mean Meaning in Life (Presence) scores for participants depending on response to the critical slide were as follows: $M = 5.01$, $SD = 1.67$ for participants who provided reasons, $M = 4.35$, $SD = 1.42$ for participants who were dumbfounded, and $M = 4.62$, $SD = 1.73$ for participants who selected "There is nothing wrong".

A multinomial logistic regression revealed a statistically significant association between Meaning in Life (Presence) and response to the critical slide, $\chi^2(2,\ N = 345) = 8.46$,

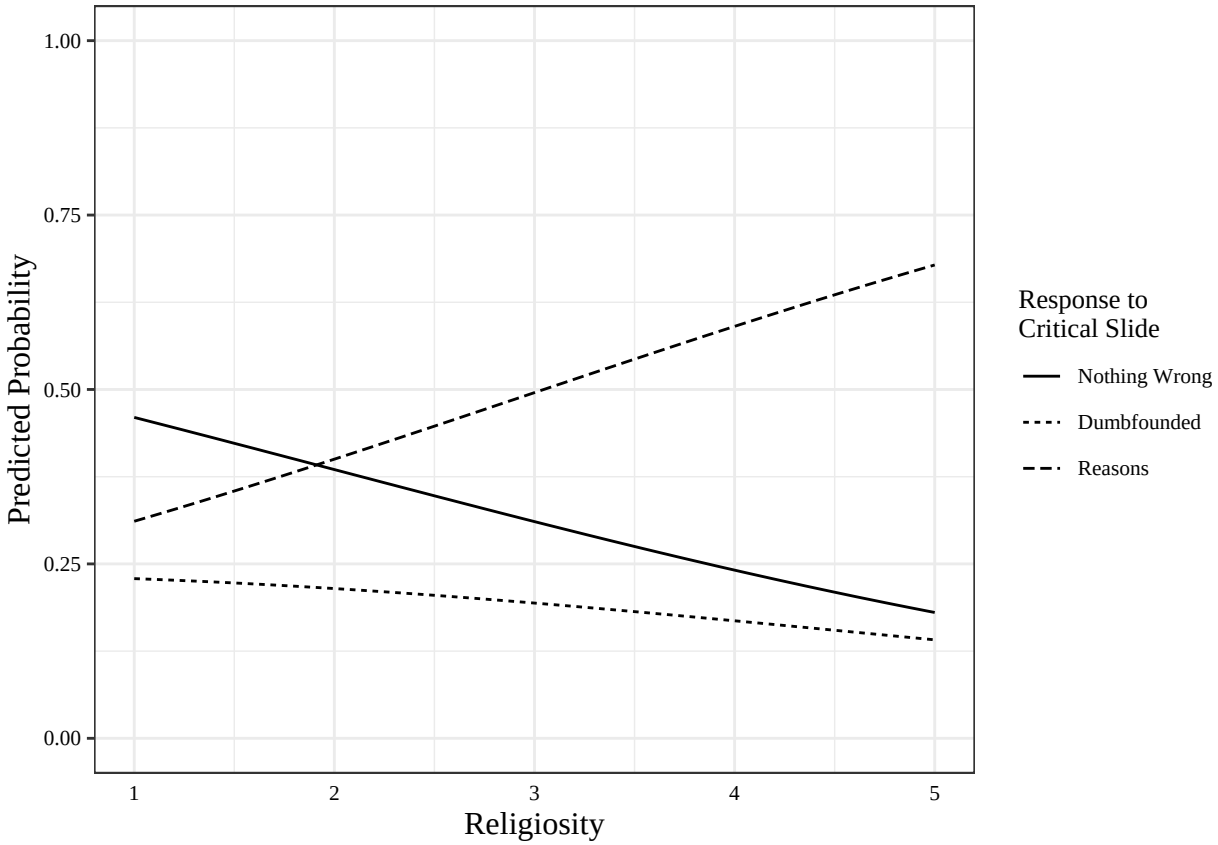

Figure 3. Probability of selecting each response to the critical slide depending on Religiosity

$p = .015$, The observed power was 0.74. Meaning in Life explained approximately 1.17% (McFadden R square) of the variance in responses to the critical slide. Participants with higher MLQ: presence scores were significantly more likely to provide reasons than to present as dumbfounded, Wald = 7.46, $p = .006$, odds ratio = 0.79, 95% CI [0.66, 0.93]. (Participants with higher MLQ: presence scores were marginally more likely to provide reasons than to select "There is nothing wrong" Wald = 3.77, $p = .052$, odds ratio = 0.86, 95% CI [0.74, 1.00].) See Figure 4.

Meaning in Life (Search). The overall mean Meaning in Life (Search) score was $M = 4.47$, $SD = 1.73$. The mean Meaning in Life (Search) scores for participants depending on response to the critical slide were as follows: $M = 4.42$, $SD = 1.75$ for participants who provided reasons, $M = 4.55$, $SD = 1.68$ for participants who were dumbfounded, and $M = 4.49$, $SD = 1.73$ for participants who selected "There is nothing wrong".

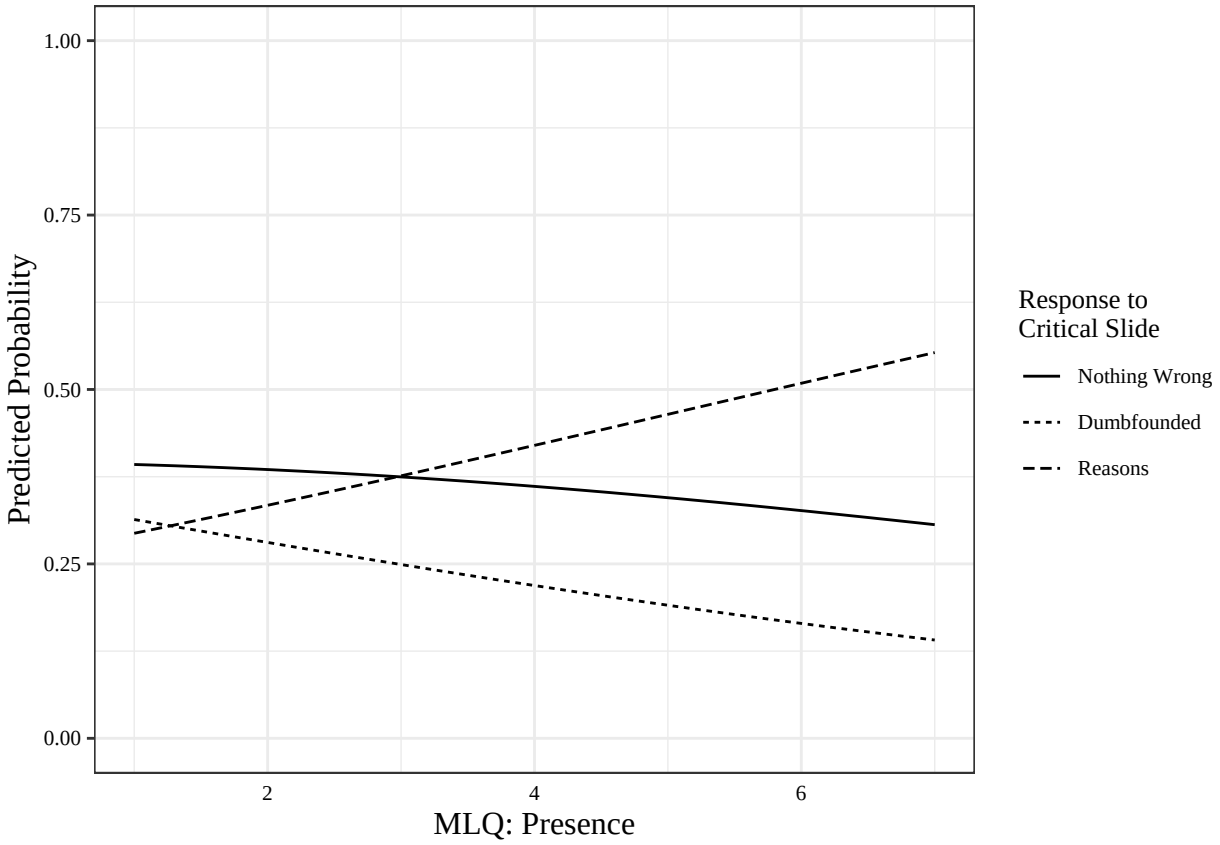

Figure 4. Probability of selecting each response to the critical slide depending on MLQ: Presence

A multinomial logistic regression revealed no statistically significant association between Search for Meaning in Life and response to the critical slide, $\chi^2(2, N = 345) = 0.3$, $p = .859$, The observed power was 0.07. See Figure 5.

Individual Differences. When analysed together, a multinomial logistic regression revealed a statistically significant association between the three individual difference variables and response to the critical slide, $\chi^2(6, N = 345) = 22.15$, $p = .001$, The observed power was 0.99. The model explained approximately 3.07% (McFadden R square) of the variance in responses to the critical slide. Religiosity was the only significant predictor (see Table 1). Participants who scored higher in Religiosity were significantly more likely to provide reasons than to select "There is nothing wrong", Wald = 12.90, p , odds ratio = 0.65, 95% CI [0.52, 0.52]. It seems religiosity was more related to valence of judgement than

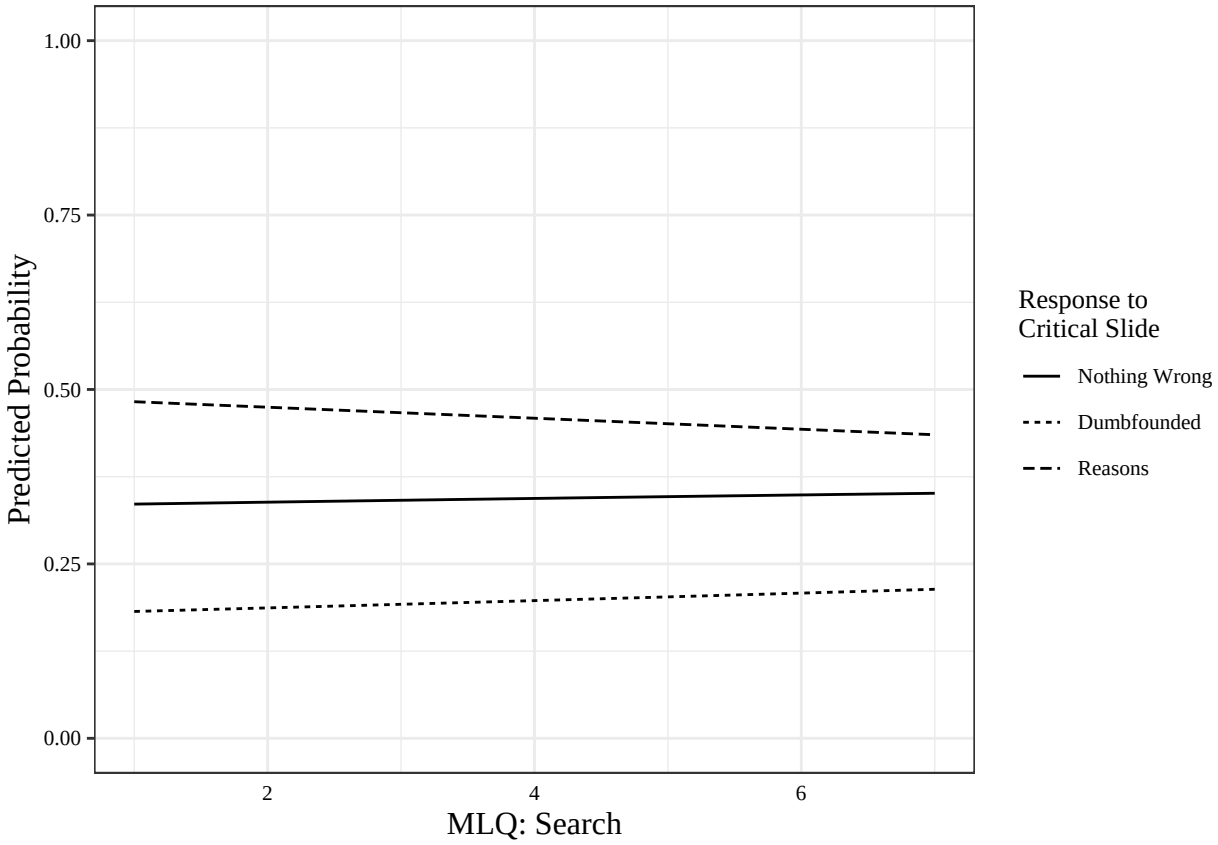

Figure 5. Probability of selecting each response to the critical slide depending on MLQ: Search

to ability to provide reasons Wald = 3.04, p , odds ratio = 0.78, 95% CI [0.59, 0.59].

A linear regression was conducted to assess the relationship between the individual difference variables (Religiosity, Meaning and Life Presence, Meaning in Life Search) and initial judgement. The model significantly predicted valence of judgement, $R^2 = .04$, $F(3, 497) = 6.22$, $p < .001$. Religiosity the only significant predictor, $b = -0.21$, 95% CI $[-0.35, -0.07]$, $t(497) = -3.01$, $p = .003$ (MLQ: presence $b = -0.08$, 95% CI $[-0.18, 0.03]$, $t(497) = -1.44$, $p = .152$; MLQ: search $b = 0.06$, 95% CI $[-0.03, 0.15]$, $t(497) = 1.25$, $p = .212$). Participants who scored higher in Religiosity were more likely to condemn the actions of Julie and Mark.

A final multinomial logistic regression was conducted that included Initial Judgement

Table 1

Multinomial logistic regression predicting responses to the critical slide where providing reasons is the referent in each case.

Variable	Response	*B*	S.E.	Wald	*df*	*p*	O.R.	Lower	Upper
Religiosity	Dumbfounded	-0.247	0.141	3.045	6	.081	0.781	0.592	1.031
	Nothing wrong	-0.426	0.119	12.899	6	<.001**	0.653	0.518	0.824
MLQ: Presence	Dumbfounded	-0.173	0.095	3.29	6	.070	0.841	0.698	1.014
	Nothing wrong	-0.043	0.082	0.275	6	.600	0.958	0.815	1.125
MLQ: Search	Dumbfounded	0.054	0.09	0.358	6	.522	1.055	0.885	1.259
	Nothing wrong	0.074	0.076	0.95	6	.207	1.077	0.928	1.25

Note. * = sig. at $p < .05$; ** = sig. at $p < .001$

Table 2

Multinomial logistic regression predicting responses to the critical slide where providing reasons is the referent in each case.

Variable	Response	*B*	S.E.	Wald	*df*	*p*	O.R.	Lower	Upper
Religiosity	Dumbfounded	-0.236	0.147	2.588	8	.108	0.79	0.593	1.053
	Nothing wrong	-0.875	0.23	14.524	8	<.001**	0.417	0.266	0.654
MLQ: Presence	Dumbfounded	-0.155	0.099	2.453	8	.117	0.856	0.705	1.04
	Nothing wrong	0.031	0.139	0.05	8	.823	1.031	0.786	1.353
MLQ: Search	Dumbfounded	0.039	0.092	0.179	8	.672	1.04	0.868	1.245
	Nothing wrong	0.031	0.136	0.05	8	.823	1.031	0.789	1.347
Initial Judgement	Dumbfounded	0.39	0.139	7.821	8	.005*	1.477	1.124	1.942
	Nothing wrong	1.98	0.219	81.533	8	<.001**	7.241	4.712	11.128

Note. * = sig. at $p < .05$; ** = sig. at $p < .001$

as a predictor variable. The results are shown in Table 2. Overall the model was a significant predictor of response to the critical slide, $\chi^2(8, N = 345) = 292.33$, $p < .001$, The observed power was 1. The model explained approximately 40.54% (McFadden R square) of the variance in responses to the critical slide. As shown in Table 2, Religiosity appeared to be related only to valence of judgement on the critical slide, initial judgement appeared to predict valence of judgement and ability to provide reasons, with more extreme judgements of "wrong" most strongly predicting the providing of reasons. The relative probabilities of selecting each response to the critical slide depending on initial judgement are displayed in Figure 6.

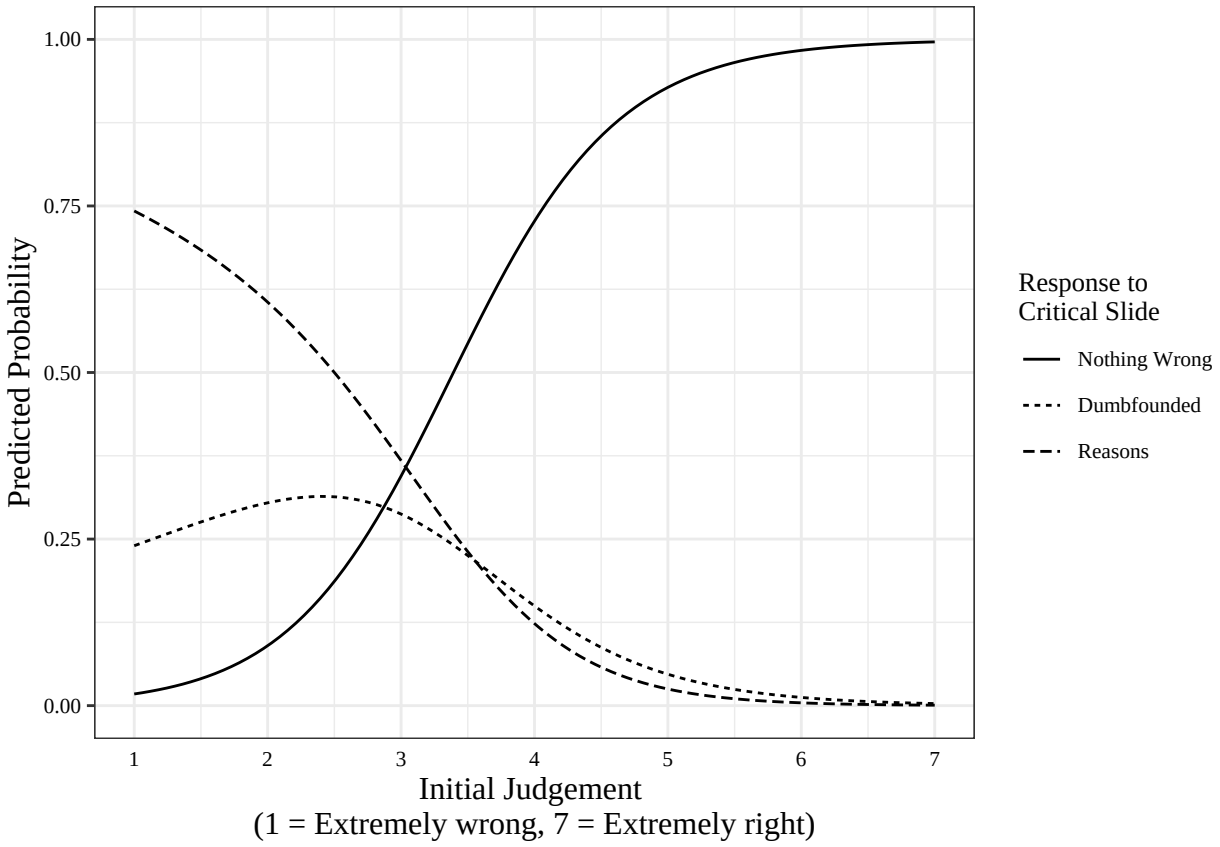

Figure 6. Probability of selecting each response to the critical slide depending on Initial Judgement.