

The Means/Side-Effect Distinction in Moral Cognition: A Meta-Analysis

Adam Feltz & Joshua May¹

Forthcoming in *Cognition*

Abstract: Experimental research suggests that people draw a moral distinction between bad outcomes brought about as a means versus a side effect (or byproduct). Such findings have informed multiple psychological and philosophical debates about moral cognition, including its computational structure, its sensitivity to the famous Doctrine of Double Effect, its reliability, and its status as a universal and innate mental module akin to universal grammar. But some studies have failed to replicate the means/byproduct effect especially in the absence of other factors, such as personal contact. So we aimed to determine how robust the means/byproduct effect is by conducting a meta-analysis of both published and unpublished studies ($k = 101$; 24,058 participants). We found that while there is an overall small difference between moral judgments of means and byproducts (*standardized mean difference* = 0.87, 95% CI 0.67 – 1.06; *standardized mean change* = 0.57, 95% CI 0.44 – 0.69; *log odds ratio* = 1.59, 95% CI 1.15 – 2.02), the mean effect size is primarily moderated by whether the outcome is brought about by personal contact, which typically involves the use of personal force.

Total Word Count: 9,758

Keywords: trolley problem; moral dilemmas; moral cognition; linguistic analogy; Double Effect; Meta-analysis; instrumental dilemmas; incidental dilemmas

1. Introduction

Many people find it morally questionable for physicians to kill their terminally ill patients as a *means* to ending suffering, even when patients competently request it. That's active euthanasia, which is illegal in many countries. Yet in the same jurisdictions it is typically legal for physicians to commence palliative care that merely has the known *side effect* of hastening a terminal patient's death. The distinction between harming as a means and harming as a byproduct can also be observed when the stakes are much lower. Some experimental studies suggest that people regard destroying one piece of property as a *means* to saving five other pieces of property as morally worse than sacrificing one as a mere side effect (or byproduct) of saving the greater goods (e.g. Millar et al 2014). Suppose, for example, that you can save someone's five rare books by diverting some spilled bleach that's fast approaching them. It may

¹ Authorship is equal; author names are ordered alphabetically by surname.

seem morally acceptable to save these books, even if you know that as a side effect the caustic liquid will then flow toward just one rare book and destroy it. But it strikes many as less morally appropriate to save the five if doing so involves using someone else's beloved book as a means to diverting the bleach.

The distinction fits with a venerable theory in moral philosophy that is associated with the Doctrine of Double Effect. The Doctrine is complicated and variously formulated. John Mikhail (2011: 149), for example, articulates it as follows:

[A]n otherwise prohibited action, such as battery or homicide, which has both good and bad effects may be permissible if the prohibited act itself is not directly intended, the good but not the bad effects are directly intended, the good effects outweigh the bad effects, and no morally preferable alternative is available.

A core element of any formulation of the Doctrine is something like the means/byproduct distinction, embodied in what we can dub the *Means Principle*: all else being equal, bringing about a bad outcome as a means to a noble goal is morally worse, or more difficult to justify, than bringing about the same outcome as a side effect (McIntyre 2001; Mikhail 2011; Wedgwood 2011).² Some, especially those in the Catholic tradition, have used the Doctrine to reconcile the ideas that a human fetus is a person, it's always wrong to intentionally kill an innocent person, but it's sometimes permissible to save the life of a mother via hysterectomy even knowing that this will have the side effect of killing the fetus (Foot 1967). Killing someone as a mere byproduct of a worthy goal may be justified, according to Double Effect, given that only the good effect of one's action is intended. Many proponents of the Doctrine, however, are secular. Indeed, something like the Means Principle is arguably presupposed in many aspects of American criminal law (Sarch in press) and is codified in the American Medical Association (AMA Opinion 2.21).

Whether the Means Principle plays a role in ordinary moral thinking has implications for longstanding debates in philosophy, psychology, and public policy. First, arguments in favor of Double Effect have commonly rested on it explaining firm commonsense intuitions about cases (e.g. Foot 1967; McIntyre 2001; Scanlon 2008; Wedgwood 2011). Such arguments suffer if our basic mode of moral thinking is not committed to the significance of the means/byproduct distinction. Second, the Means Principle may serve to underwrite conceptions of moral cognition as involving tacit computation (Cushman et al 2006) that is perhaps universal and innate (Hauser et al 2007; Mikhail 2011). Third, many who have attempted to empirically debunk deontological or non-utilitarian ways of thinking have regarded the Means Principle as a core element of the targeted approach to ethics, since it treats more than outcomes as morally significant (Greene 2013; Sinhababu 2013). Thus, the status of the Means Principle in ordinary moral cognition informs a wide range of debates, from empirical questions about human nature to moral questions about the plausibility of certain ethical traditions.

Participants in such debates have understandably focused attention on numerous experiments that have reported a means/byproduct effect in moral judgment. Some researchers have reported the effect among adults and often when using the famous trolley dilemmas or other similar sacrificial dilemmas involving life and death (e.g. Mikhail 2002; Cushman et al 2006; Hauser et al 2007; Sinnott-Armstrong et al 2008; Moore et al. 2008; Cushman & Young 2011).

² Side effects are notoriously difficult to define, but for our purposes a *side effect* is (roughly) an effect of an individual's action that is not a goal of hers or a means to one of her goals (cf. Cushman & Mele 2008: 179).

But other studies have apparently generated the effect using non-trolley dilemmas or in other areas of moral judgment, such as situations involving bodily harm, cheating, financial loss, pollution, and property damage (e.g. Nichols & Mallon 2006; DeScioli et al 2012; Gold et al 2013; Millar et al 2014; Kelman & Kreps 2014; May ms-b). Moreover, some studies have found the means/byproduct effect in children (e.g. Pellizzoni et al 2010; Mikhail 2011) and non-Western populations (e.g. Hauser et al 2007; Abarbanell & Hauser 2010; Mikhail 2011; Moore et al 2011a; Ahlenius & Tännsjö 2012; Kawai et al 2014).

However, there are reasons to worry that the purported effect is not robust or perhaps even non-existent. First, some of these studies appear to have confounds, conjoining harming as a means with other factors relevant to moral cognition, including contact, commission, battery, and personal force (Greene et al 2009; May 2014; Mikhail 2014). Consider, for example, what is arguably the most famous pair of cases in this literature: Switch and Footbridge. In *Switch*, the protagonist can either do nothing and let an empty runaway trolley kill five innocent people stuck on the tracks or flip a switch that will divert the trolley to a side-track with only one innocent person on it. Here, sacrificing one for the greater good involves only causing a death as a side effect of a noble goal. In *Footbridge*, the protagonist is on a bridge and can save the five only by pushing a man onto the tracks who is large enough to stop the trolley with his body. Here, the actor can promote the greater good by killing not only as a means but in a violent way that requires up-close and personal contact with the victim. And some experiments suggest an important interaction effect between harming as a means and using personal force (Greene et al 2009).

A second worry is that the differences in people's judgments are rather small when we focus on vignettes that don't involve confounds such as personal contact. Consider, for example, one pair of trolley cases that remove the contact confound. In *Loop*, the side track circles back around and the trolley will return toward the five if it continues around the loop, but there is one innocent man stuck on the looping track who is large enough to stop the trolley from continuing on to kill the five. Like Footbridge, killing the man on the loop track is likely to be represented as harming as a means, but it doesn't involve up-close and personal contact. Now contrast Loop with *Man-in-Front*, in which the only change is that behind the one man on the loop track is a large boulder sufficient on its own to stop the trolley. Killing the one now looks to be a mere side effect of smashing the empty trolley into the boulder. Some early studies report that moral judgments diverge about this minimal pair of cases: more people regard Man-in-Front as morally permissible than Loop (Hauser et al 2007; Mikhail 2011).

Such studies may provide hope that the Means Principle does guide ordinary moral cognition when problematic confounds are removed. However, some have noted (e.g. Enoch 2013: 10) that, while the differences in permissibility judgments between the classic Switch/Footbridge pair are consistently quite large (e.g. 85% and 12%), the differences in the Man-in-Front/Loop pair are much smaller (e.g. 72% and 56% in Hauser et al 2007) suggesting that personal contact moderates the effect. Similarly, studies using continuous measures of moral judgments often find exceedingly small differences on a fine-grained scale (see e.g. Cushman et al 2006; Cushman & Young 2011). So, even if the differences in intuitions are statistically significant, it's unclear whether the Means Principle has a powerful impact on moral cognition (cf. May 2014; Cushman 2016).

A final worry about the means/byproduct effect involves replication. At least for vignettes that don't appear to involve confounds, there have been some failures to replicate the means/byproduct effect (e.g. Waldmann & Dieterich 2007; Greene et al 2009; Zimmerman 2013;

May ms-a). For example, while the proportion of permissibility judgments tends to go down when death is brought about as a means (Loop) rather than a mere side-effect (Man-in-Front), two subsequent studies using this same categorical measure have found that participants are inclined to think it's permissible to sacrifice the one for the greater good in both cases (see Table 1).

Table 1: Examples of Inconsistent Categorical Data

	Mikhail (2002/2011)	Hauser et al (2007)	Zimmerman (2013)	May (ms)
Loop	48%	56%	89%	77%
Man-in-Front	62%	72%	89%	77%

Note. Percentages indicate the proportion of participants reporting that the choice of killing one to save five is morally permissible.

One might try to explain the small effect size or the inconsistent data by pointing to another body of research on the *side-effect effect*. Participants are consistently more inclined to say that an individual brought about a side effect intentionally if the side effect is bad as opposed to good (Knobe 2010). When researchers try to study the means/byproduct effect, they generate cases in which the protagonist is generating bad outcomes, such as the death of an innocent person. But, given such negative consequences of the well-intentioned action, it might be difficult for people to see such outcomes as being mere side effects (as in Switch). So such “side effects” might be represented as more intentional, perhaps even intended, much like cases in which the protagonist is supposed to have generated an outcome as a means (as in Footbridge). Perhaps for some participants the side-effect effect masks the difference between generating a bad outcome as a means versus a byproduct, which leads to similar moral judgments about the relevant pairs of cases.

There are at least two reasons to doubt that this masking account explains the different results across studies. First, if people treat the bad side effects as more like harming as a means, then permissibility judgments should be *equally low* in both cases, not equally high. Second, some studies indicate that, for whatever reason, the side-effect effect doesn't play a role in the trolley-type cases, since people do treat harming as a byproduct as less intentional than harming as a means (e.g. Sinnott-Armstrong et al 2008, Table 1; Cushman & Young 2011; Watkins & Laham 2016).

There is thus a need for a meta-analysis to estimate the overall effect size in an effort to determine the robustness of the apparent means/byproduct effect. Such an endeavor could also reveal whether the effect is moderated by other factors, particularly personal contact. The results of our meta-analysis are complex but suggest that, while the means/byproduct distinction is reflected in everyday moral cognition, the effect is almost completely moderated by personal contact. Our results are consistent with various studies contained within the meta-analysis itself, including those that report a pure means/byproduct effect (e.g. Cushman et al. 2006) and those that report an interaction with personal contact (Greene et al. 2009). However, as we'll later discuss, the exact theoretical upshot of this state of affairs is important and as yet unresolved.

2. Methods

2.1 Search Criteria

The main goal of the meta-analysis was to determine the extent to which the Means Principle is present in everyday moral judgments about the rightness or wrongness of various acts. To that end, we established the following criteria for including empirical studies in the meta-analysis.

First, the studies had to compare moral judgments about both outcomes caused as a means *and* as a byproduct. So, for example, studies were excluded if they simply measured moral judgments about a harm caused as a byproduct (e.g. in the Switch case) but not as a means.

Second, the judgments had to be about what one *should or shouldn't do* in a situation. For our purposes, this category includes judgments about an action's *wrongness, moral acceptability or appropriateness*, and which choice one *would* make (given that it's clear a moral dilemma is at issue). Previous research suggests there are important differences between judgements about acts being *wrong/inappropriate* judgments about individuals as *blameworthy/deserving of punishment* (e.g. Cushman 2008; O'Hara et al. 2010). So, while other judgments may concern ethics or morality (e.g. attributions of responsibility, praise, blame, or virtue), we excluded studies that did not measure participant's beliefs about which action one ought to perform. Provided the measured judgments were about what one should do, they could be about any type of action, not just those that cause harm (e.g., in the classic Switch and Footbridge cases). While harm is a paradigmatic instance of a moral violation, the Means Principle should surface for other actions that are morally relevant, such as violations of property rights.

2.2 Literature Search

Our search for articles occurred from July 7, 2015 to July 12, 2015 and then again from December 1, 2016 to February 10, 2017. We began by searching Google Scholar, Web of Science, and ProQuest Psychology Journals for articles using the following keywords (no restrictions on publication date): *common sense morality, action versus omission and moral judgments, trolley problem, trolley problems and studies, trolley problem and empirical data, trolley dilemma and studies, trolley dilemma and empirical data, moral dilemmas and studies, moral dilemma and empirical data, doctrine of double effect and studies, instrumental vs incidental studies*. With the resulting papers, we did forward and backward citation searches. We ultimately identified 69 unique papers whose abstracts we reviewed to see if the paper likely contained studies relevant to the means/byproduct distinction. If the abstract indicated the paper likely contained relevant studies, we read the full papers to determine if the inclusion criteria were met. This method left 50 articles with data to be extracted.

We also attempted to minimize publication bias to avoid inflating mean effect sizes. Compared to unpublished studies, published studies more often are "conventionally significant" and are likely to have higher estimated effect sizes. So including only published studies would tend to over-estimate the mean effect size in the meta-analysis. Unpublished studies were solicited via personal communication and by posting calls for unpublished data on well-trafficked, highly visible blogs and listservs (i.e., Experimental Philosophy Blog, Brains Blog, The Moral Psychology Group on Facebook, the Judgment and Decision Making Listserv, the Society for Philosophy and Psychology Listserv). These calls yielded an additional 11 studies that met the inclusion criteria for the meta-analysis.

2.3 Variables in Studies

The studies included in the meta-analysis ($k = 101$, Table 2) varied in a number of ways. First, some measured moral judgments using a categorical variable (e.g. permissible/impermissible) while others used a continuous variable (e.g. a 9-point scale of moral acceptability).

Second, some studies had a within-subjects (rather than between-subjects) design, where each participant responded to both kinds of cases under investigation (e.g. Switch and Footbridge). Some of the studies that used a within-subjects design tested for order effects by presenting each participant with multiple vignettes in different orders. In such cases, we calculated the mean response across all orders of presentation for use in the meta-analysis.

Third, some studies conducted multiple experiments on the same participants but not in a between-subjects design (e.g., participants were given five means scenarios and five byproduct scenarios). In those instances, we treated those multiple studies as one unit in the meta-analysis and took the mean of responses.

Fourth, most of the studies used slightly different materials that are too idiosyncratic to describe meaningfully. However, one major dimension that clearly differentiated some studies from others was the use of *personal or bodily contact* (e.g. pushing or bumping) in the means condition but not in the byproduct condition. We decided to focus only on the broad and intuitive notion of contact rather than related categories that are more fine-grained and theory-driven, such as personal force (Greene et al 2009), prototypically violent acts (2013), and battery (Mikhail 2014).

Table 2: Studies and brief description of studies in the meta-analysis

Study #	Authors	Year	Design	Data	Personal Contact
1	Moore et al	2008	Between	Categorical	No
2	May	unpublished	Between	Categorical	No
3	Mikhail	2002	Between	Categorical	Yes
4	Nichols & Mallon	2006	Between	Categorical	No
5	Zimmerman	2013	Between	Categorical	No
6	Mikhail	2011	Between	Categorical	Yes
7	Mikhail	2011	Between	Categorical	Yes
8	Mikhail	2011	Between	Categorical	No
9	May	unpublished	Between	Categorical	No
10	May	unpublished	Between	Categorical	No
11	May	unpublished	Between	Categorical	No
12	Hauser et al	2007	Between	Categorical	Yes
13	Hauser et al	2007	Between	Categorical	No
14	Nakamura	unpublished	Between	Categorical	Yes
15	Pellizzoni et al	2010	Within	Categorical	Yes
16	Nichols & Mallon	2006	Within	Categorical	No
17	Ahlenius & Taennsjoe	2012	Within	Categorical	Yes
18	Cote et al	2013	Within	Categorical	Yes
19	Lotto et al	2014	Within	Categorical	Yes

20	Koengis et al	2007	Within	Categorical	yes
21	Rusch	2015	Within	Categorical	Yes
22	Sarlo et al	2012	Within	Categorical	Yes
23	Costa et al	2014	Within	Categorical	Yes
24	Fumagalli et al	2010	Within	Categorical	Yes
25	Koenigs et al	2012	Within	Categorical	Yes
26	Manfrinait et al	2013	Within	Categorical	Yes
27	Moore et al	2011a	Within	Categorical	Yes
28	Moore et al	2011a	Within	Categorical	Yes
29	More et al	2011b	Within	Categorical	Yes
30	Millar et al	2014	Between	Continuous	Yes
31	Millar et al	2014	Between	Continuous	Yes
32	Millar et al	2014	Between	Continuous	Yes
33	Millar et al	2014	Between	Continuous	No
34	Greene et al	2009	Between	Continuous	No
35	Geren et al	2009	Between	Continuous	Yes
36	Shallow et al	2011	Between	Continuous	Yes
37	Gold et al	2013	Between	Continuous	Yes
38	Wiegmann et al	2013	Between	Continuous	Yes
39	Wiegmann et al	2013	Between	Continuous	Yes
40	Wiegmann et al	2013	Between	Continuous	Yes
41	Wiegmann et al	2013	Between	Continuous	Yes
42	Lombrozo	2008	Between	Continuous	Yes
43	Sinnott-Armstrong et al	2008	Between	Continuous	No
44	Waldmann & Dieterich	2007	Between	Continuous	No
45	Zimmerman	2013	Between	Continuous	No
46	Waldmann & Wiegmann	2010	Between	Continuous	No
47	Shepard	unpublished	Between	Continuous	No
48	Amit & Greene	2012	Between	Continuous	Yes
49	Broeders et al	2011	Between	Continuous	Yes
50	Broeders et al	2011	Between	Continuous	Yes
51	Broeders et al	2011	Between	Continuous	Yes
52	Kelman & Kreps	2014	Between	Continuous	Yes
53	Kelman & Kreps	2014	Between	Continuous	Yes
54	Kelman & Kreps	2014	Between	Continuous	Yes
55	Horne & Powell	2016	Between	Continuous	Yes
56	Horne & Powell	2016	Between	Continuous	Yes
57	Watkins & Laham	2016	Between	Continuous	Yes
58	Watkins	2016	Between	Continuous	Yes
59	Watkins	2016	Between	Continuous	Yes
60	Watkins	2016	Between	Continuous	Yes
61	Watkins & Laham	unpublished	Between	Continuous	No
62	Cao et al	2017	Between	Continuous	Yes

63	Cao et al	2017	Between	Continuous	Yes
64	Cao et al	2017	Between	Continuous	Yes
65	Sytsma & Livengood	unpublished	Between	Continuous	Yes
66	Sytsma & Livengood	unpublished	Between	Continuous	Yes
67	Sytsma & Livengood	unpublished	Between	Continuous	Yes
68	Wiegmann et al	2012	Between	Continuous	Yes
69	Wiegmann et al	2012	Between	Continuous	No
70	Cushman & Young	2011	Within	Continuous	No
71	Cushman & Young	2011	Within	Continuous	No
72	Cushman & Young	2011	Within	Continuous	No
73	Cushman & Young	2011	Within	Continuous	No
74	Cushman et al	2006	Within	Continuous	No
75	Schwitzgebel & Cushman	2012	Within	Continuous	Yes
76	Liao et al	2012	Within	Continuous	Yes
77	Kawai et al	2014	Within	Continuous	Yes
78	Kawai et al	2014	Within	Continuous	Yes
79	Kawai et al	2014	Within	Continuous	Yes
80	Kawai et al	2014	Within	Continuous	Yes
81	Abarbanell & Hauser	2010	Within	Continuous	Yes
82	DeScioli et al	2012	Within	Continuous	No
83	DeScioli et al	2012	Within	Continuous	No
84	Wiegmann et al	2013	Within	Continuous	Yes
85	Lombrozo	2008	Within	Continuous	Yes
86	Lotto et al	2014	Within	Continuous	Yes
87	Tempesta et al	2011	Within	Continuous	Yes
88	Christensen et al	2014	Within	Continuous	Yes
89	Christensen et al	2014	Within	Continuous	No
90	Ugazio et al	2012	Within	Continuous	Yes
91	Ugazio et al	2012	Within	Continuous	Yes
92	Watkins	2016	Within	Continuous	Yes
93	Watkins	2016	Within	Continuous	Yes
94	Watkins	2016	Within	Continuous	Yes
95	Laham & Watkins	unpublished	Within	Continuous	No
96	Wiegmann et al	2012	Within	Continuous	No
97	Wiegmann & Waldman	2014	Within	Continuous	Yes
98	Wiegmann & Waldman	2015	Within	Continuous	Yes
99	Wiegmann & Waldman	2016	Within	Continuous	Yes
100	Wiegmann & Waldman	2017	Within	Continuous	Yes
101	Wiegmann & Waldman	2018	Within	Continuous	Yes

3. Results

Meaningful comparisons between studies with different designs are difficult to make in a meta-analysis (Lipsey & Wilson, 2001). So three separate meta-analyses were performed on studies collecting: within-subjects continuous data (*within-continuous*), between-subjects continuous data (*between-continuous*), and categorical data. We did not conduct separate meta-analyses on within and between subjects categorical data. Rather, the study design was a moderator variable for categorical data to determine if differences in designs matter to overall effect size estimates (Morris & DeShon, 2002). Each meta-analysis used different effect size estimates: log odds ratios for the categorical data;³ standardized mean differences for the between-continuous data; and standardized mean change for the within-continuous data.

For within-continuous studies, one necessary piece of information is the correlation between the two key moral judgments under investigation (i.e., judgments about cases of generating an outcome either as a byproduct or as a means). No study that used a within-continuous design reported the correlation, so we estimated that the correlation is .5 and used that value in calculations. (The correlation is used when calculating the standard errors and, consequently, estimates of variability. If a larger r value is used, SEs should be smaller than when a smaller r value is used.)

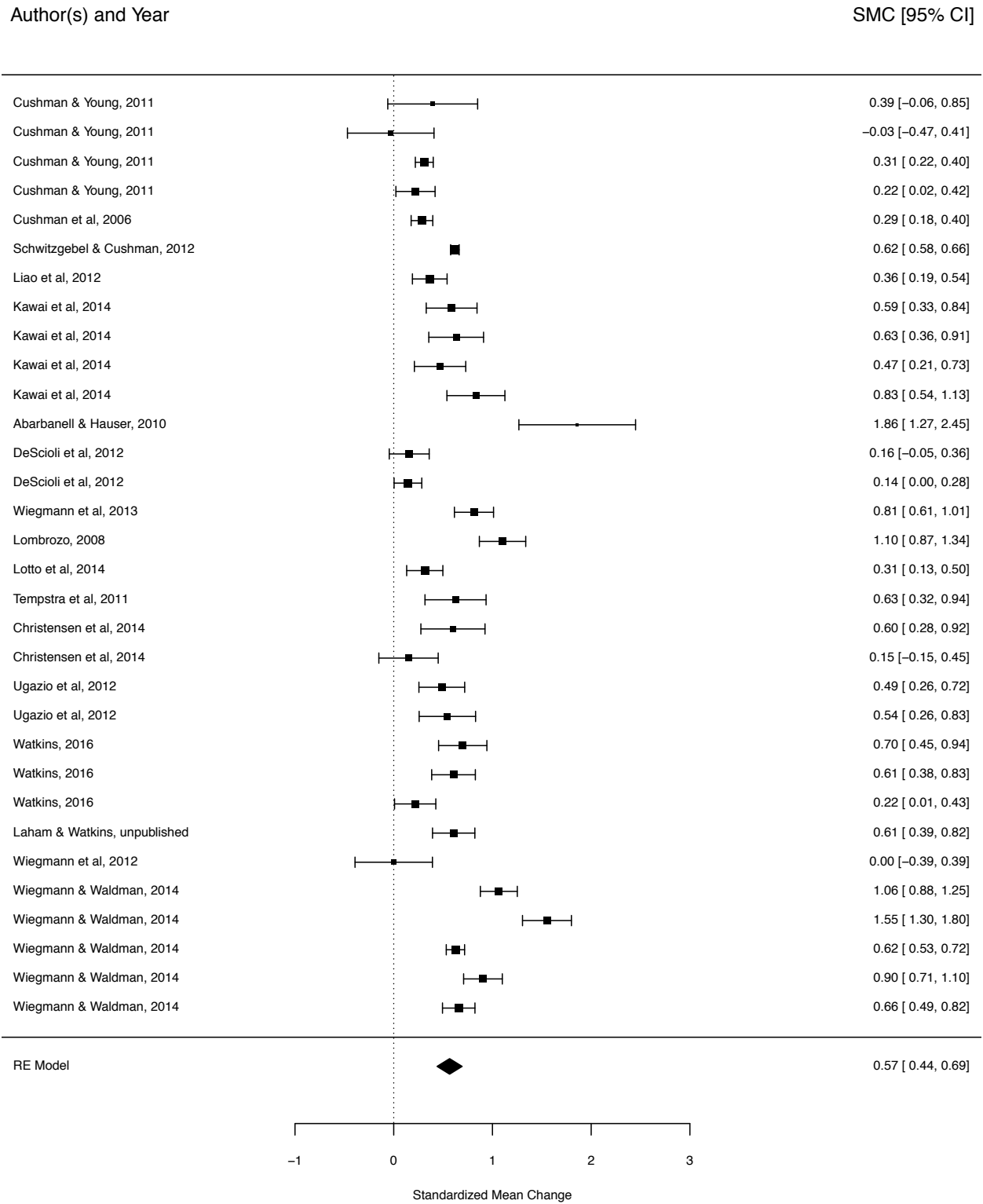
Despite the differences in designs, the three meta-analyses proceeded in the same way. The Metafor (Viechtbauer, 2010) package for R (R Core Team, 2014) was the tool of choice for all analyses and graphics. Since variability in estimates typically decreases with larger sample sizes, we weighted studies by the inverse of the variance ($1/\text{variance}$) (Lipsey & Wilson, 2001; Viechtbauer 2010), giving those with larger samples more importance in the meta-analysis than studies with smaller samples. For each meta-analysis, we fit random-effects models with restricted maximum likelihood estimation (Viechtbauer, 2010), which estimate the precision of the effect size taking into account sampling level error and random error. We also performed a moderator analysis for each meta-analysis to help identify a source of differences in the distribution of effect sizes. One prominent confound in the studies was contact versus no contact. For this reason, the two authors independently coded the contact/no contact moderator with near perfect agreement (95%, $kappa = .88$, $z = 8.78$, $p < .001$), with the few instances of disagreement settled by discussion.

3.1 Within-Continuous Data

We first analyzed studies that used a within-subjects design and measured moral judgments with a continuous variable ($k = 31$). There was an overall, moderately sized mean effect: *standardized mean change* = 0.57, 95% CI 0.44 - 0.69, $SE = .06$, $z = 8.73$, $p < .001$ (see Figure 1).

³ A standard effect size for categorical variables is the odds ratio, but the odds ratio is not easily incorporated into a meta-analysis. The odds ratio is centered on 1 (indicating no effect), with values less than 1 indicating a negative relation, values greater than 1 as positive relations. And the odds ratio can never be less than 0 which makes it peculiar. To illustrate, an odds ratio of 0.5 is of the same magnitude as an odds ratio of 2. Consequently, most meta-analyses take the natural logarithm of the odds ratio as the unit to be analyzed in a meta-analysis (Lipsey & Wilson, 2001). To convert back to the more easily interpretable odds ratio one can use the following formula: $e^{\log OR}$, where e is the base of the natural logarithm (about 2.718) (Lipsey & Wilson, 2001, p. 54).

Figure 1: Forest Plot for Continuous, within-subjects studies

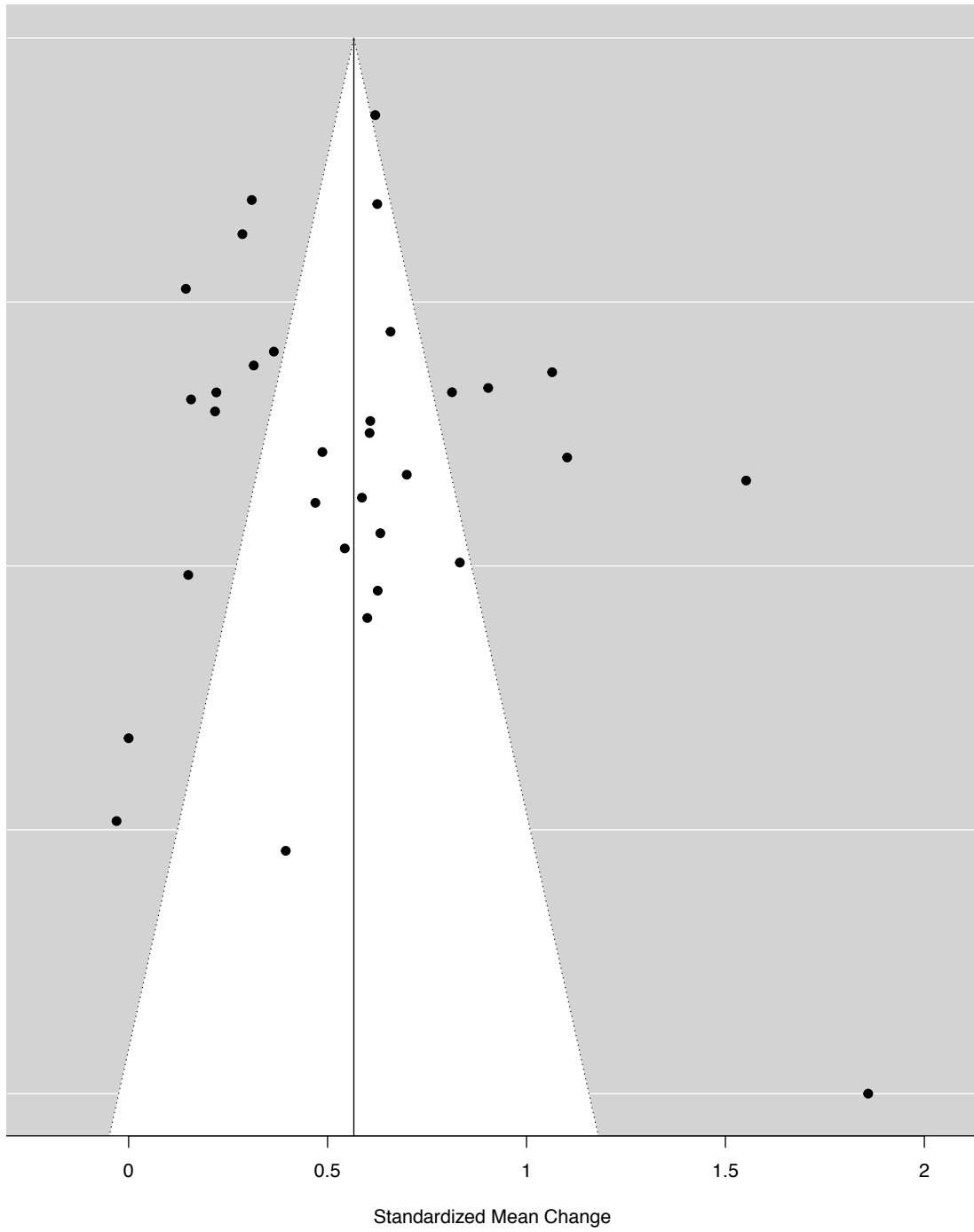


A visual inspection of the funnel plot did not indicate publication bias (see Figure 2). This was confirmed with a regression test using the standard error as the predictor, $t(30) = 0.06$, $p = .95$ (Egger et al, 1997; Viechtbauer, 2010). While the regression suggested a lack of publication bias, we performed two tests for Failsafe N s.⁴ The Rosenthal method estimates the number of studies with an average effect size of zero would be needed to be included in the meta-analysis for the mean effect size to be non-significant (e.g., $p > .05$) (Rosenthal, 1979). The Rosenthal method suggested that 12,567 studies would have to be added for the mean effect size to be non-significant. We also conducted a different Failsafe N based on Orwin's (1983) method. Orwin's method provides the number of studies that would need to be added to the meta-analysis with an average effect size of zero to reach some target mean effect size. We selected 0.13 as the target mean effect size because that value is the difference between the estimated standardized mean change and the lower bound of the 95% confidence interval. This method suggested that 110 studies would have to be added to the meta-analysis to reach that target overall mean effect size. Consequently, this meta-analysis should be robust to studies that are not included.

As the funnel indicates, there was a great deal of heterogeneity in effect sizes, $Q(df = 31) = 320.87$, $p < .001$, $I^2 = .94$. The substantial heterogeneity suggested that moderator variables might be able to account for some of the variance in the distribution of effect sizes. A mixed-model analysis using contact/no contact as the moderator variable indicated that this moderator accounted for a significant amount of the variation in effect sizes: $Q_m(df = 1) = 18.04$, $p < .001$. More specifically, the mean effect size was significant when contact was involved (*standardized mean change* = 0.71, 95% CI 0.59 – 0.83 $z = 11.5$, $p < .001$), but was substantially less (about a third the size) when contact was absent (*standardized mean change* = 0.24, 95% CI 0.05 – 0.42, $z = 2.53$, $p = 0.01$).

⁴ Failsafe N s are sometimes difficult to interpret and the regression on the funnel plot is a better method to detect publication bias (Becker, 2005). However, the failsafe N s can be helpful to illustrate what the magnitude of unreported studies would have to be to change mean effect sizes to non-significant.

Figure 2: Funnel plot for continuous, within studies



3.2 Between-Continuous Data

Next we analyzed studies that used continuous measures of moral judgments in a between-subjects design ($k = 40$). There was an overall, moderately sized mean effect: *standardized mean difference* = 0.87, 95% CI 0.67 – 1.06, $SE = .1$, $z = 8.69$, $p < .001$ (see Figure 3).

A visual inspection of the funnel plot did not reveal evidence for publication bias (see Figure 4). A regression test using the standard error as the predictor did not reveal evidence of publication bias, $t(38) = 1.28$, $p = .21$ (see Figures 3 and 4). To provide additional evidence against publication bias in this meta-analysis, we tested file-drawer effects with two different methods. The Rosenthal method suggested that 12,616 studies with no effects would have to be added for the mean effect size to be non-significant. We also used the Orwin Method with the target mean effect size of 0.2 (the difference between the estimated mean effect size and the lower bound of the 95% confidence interval). The Orwin method suggested that 98 studies with a mean effect of zero are required to reduce the overall mean effect size to that target level. Consequently, this meta-analysis should be robust to studies that are not included.

Again there was a great deal of heterogeneity in effect sizes, $Q(df = 39) = 428.6$, $p < .001$, $I^2 = .92$. So we performed a mixed-model analysis using contact/no contact as the moderator variable. The analysis suggested that the contact moderator accounted for a significant amount of the variation in effect sizes: $Q_m(df = 1) = 6.84$, $p = .009$. The mean effect size, again, was significant when contact was involved (*standardized mean difference* = 0.99, 95% CI 0.79-1.2, $z = 9.61$, $p < .001$), but was much smaller (by more than a half) when contact was absent (*standardized mean difference* = 0.41, 95% CI 0.02 – 0.8, $z = 2.04$, $p = 0.02$).

Figure 3: Forest plot for between, continuous studies

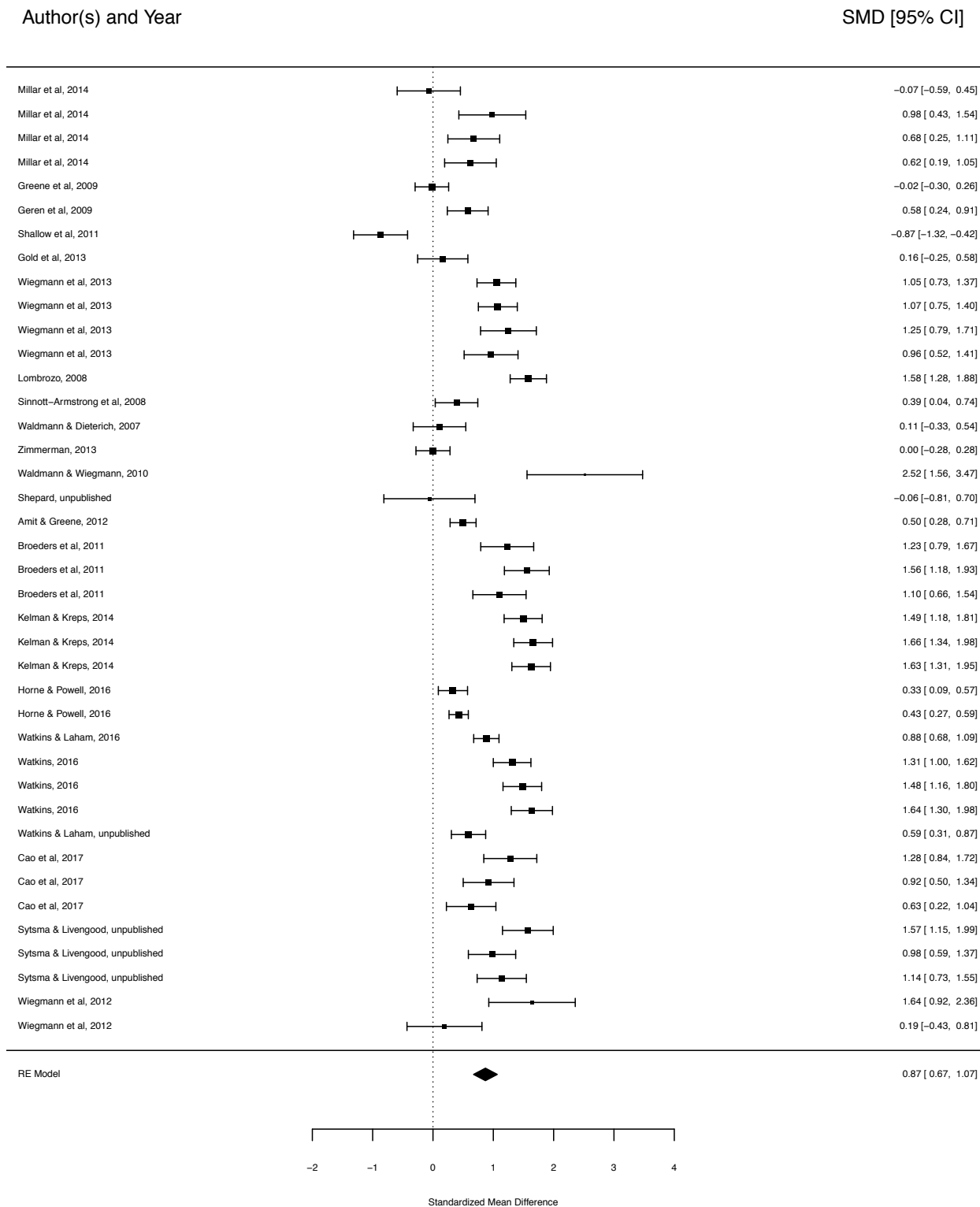
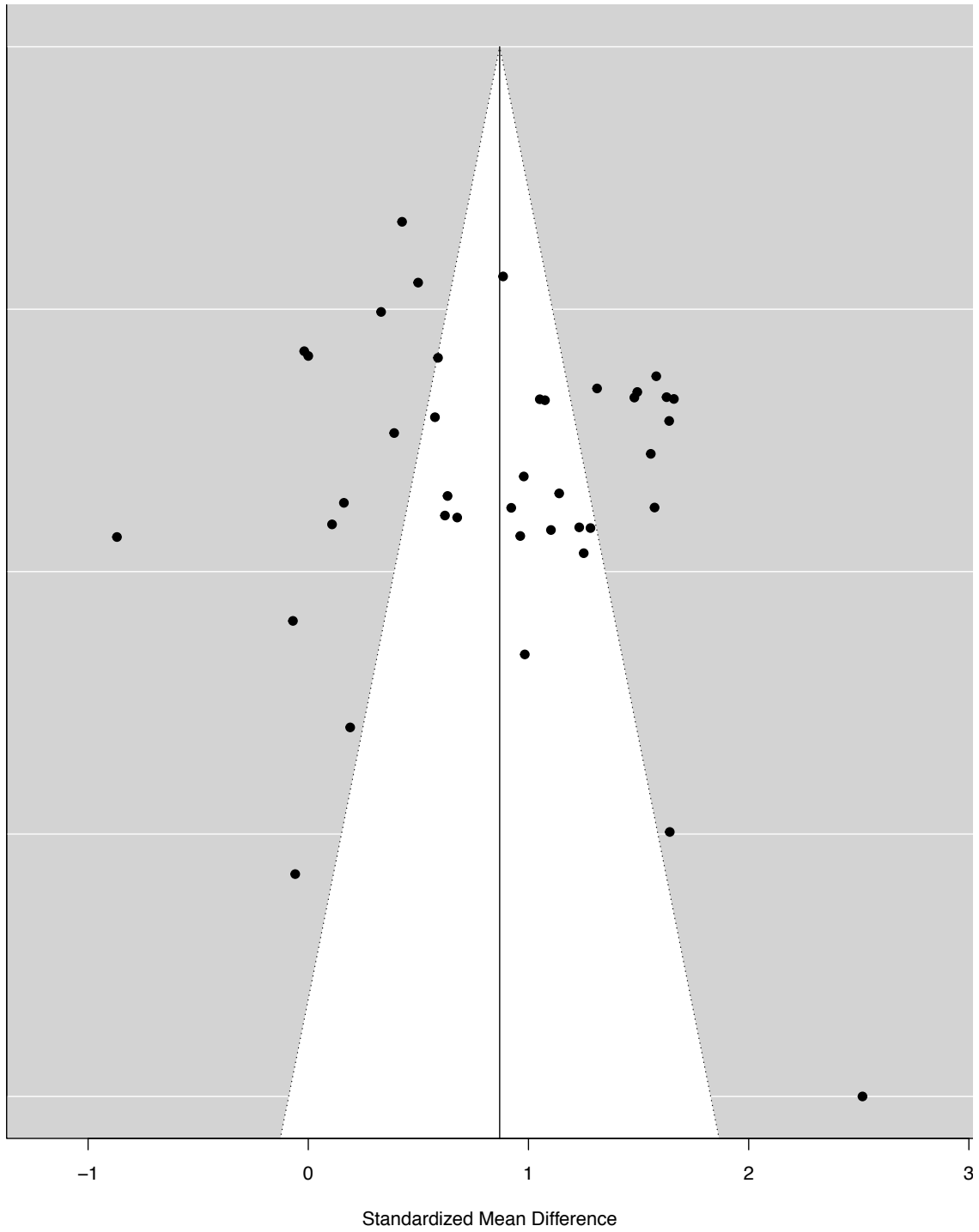


Figure 4: Funnel plot for between, continuous studies



3.3 Categorical Data

Finally, we analyzed studies that used categorical measures of moral judgment, including both between- or within-subjects designs ($k = 30$). There was an overall, large mean effect: *log odds ratio* = 1.59, 95% CI 1.15 – 2.02, SE = .22, $z = 7.09$, $p < .001$ (see Figure 5).

A visual inspection of the funnel plot did not reveal evidence for publication bias (see Figure 6). A regression test using the standard error as the predictor did not reveal evidence of publication bias, $t(28) = 0.33$, $p = .74$ (see Figures 5 and 6). To provide additional evidence against publication bias in this meta-analysis, we tested file-drawer effects with two different methods. The Rosenthal method suggested that 9,062 studies with no effects would have to be added for the mean effect size to be non-significant. We also used the Orwin method with target mean effect size of 0.44 (the difference between the overall mean effect size and the lower bound of the 95% confidence interval). The Orwin method suggested that 88 studies would have to be added to reach that target effect size. Consequently, this meta-analysis should be robust to studies that are not included.

As with the previous analyses, there was a great deal of heterogeneity in effect sizes $Q(df = 29) = 758.24$, $p < .001$, $I^2 = .95$. Studies using categorical measures called for two different moderator analyses. First, we tested whether between- or within-subjects designs significantly influenced mean effect sizes (Morris & DeShon, 2002). A mixed-model analysis using the study design as the moderator did not account for a significant amount of the variation in effect sizes: $Q_m(df = 1) = 0.17$, $p = 0.69$. When a within-subjects design was used (*log odds ratio* = 1.67, 95% CI 1.07 – 2.27, $z = 5.47$, $p < .001$) the overall mean effect size and confidence intervals were similar to when a between-subjects design was used (*log odds ratio* = 1.49, 95% CI 0.83 – 2.15, $z = 4.41$, $p < .001$). Second, a mixed-model analysis using contact/no contact as the moderator variable suggested again that the personal contact moderator accounted for a significant amount of the variation in effect sizes: $Q_m(df = 1) = 8.25$, $p = .004$. When personal contact was involved, the pattern of results typical of the Means Principle was large and significant (*log odds ratio* = 2.97, 95% CI 1.50 – 2.45, $z = 8.14$, $p < .001$) but was much smaller in the absence of personal contact (*log odds ratio* = 0.78, 95% CI 0.13 – 1.44, $z = 2.33$, $p = .02$). To convert these values back into a more interpretable effect size, the odds ratio involving contact is very large, $OR = 19.48$, whereas the odds ratio involving non-contact is small, $OR = 2.18$.

Figure 5: Forest plot for categorical studies

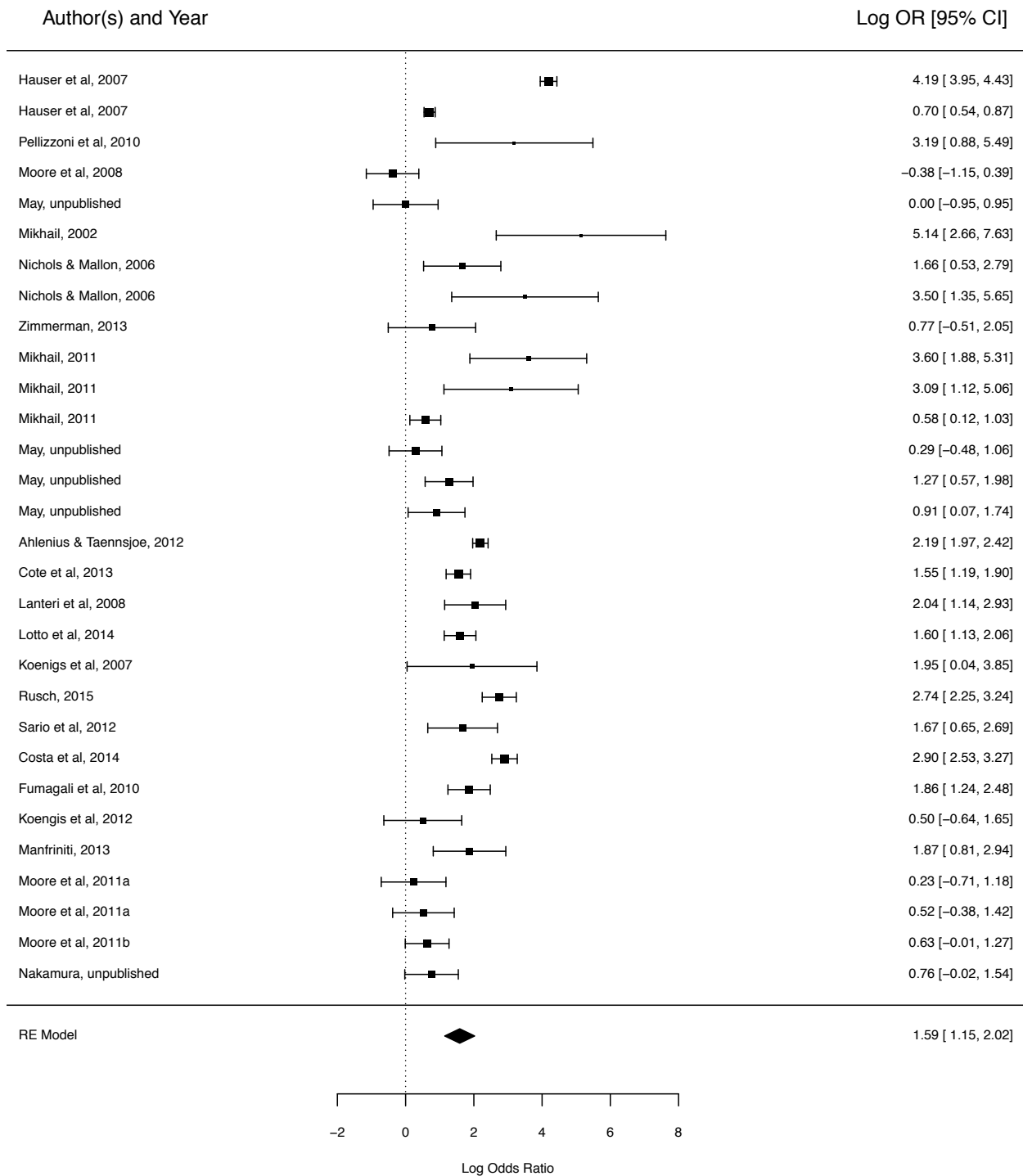
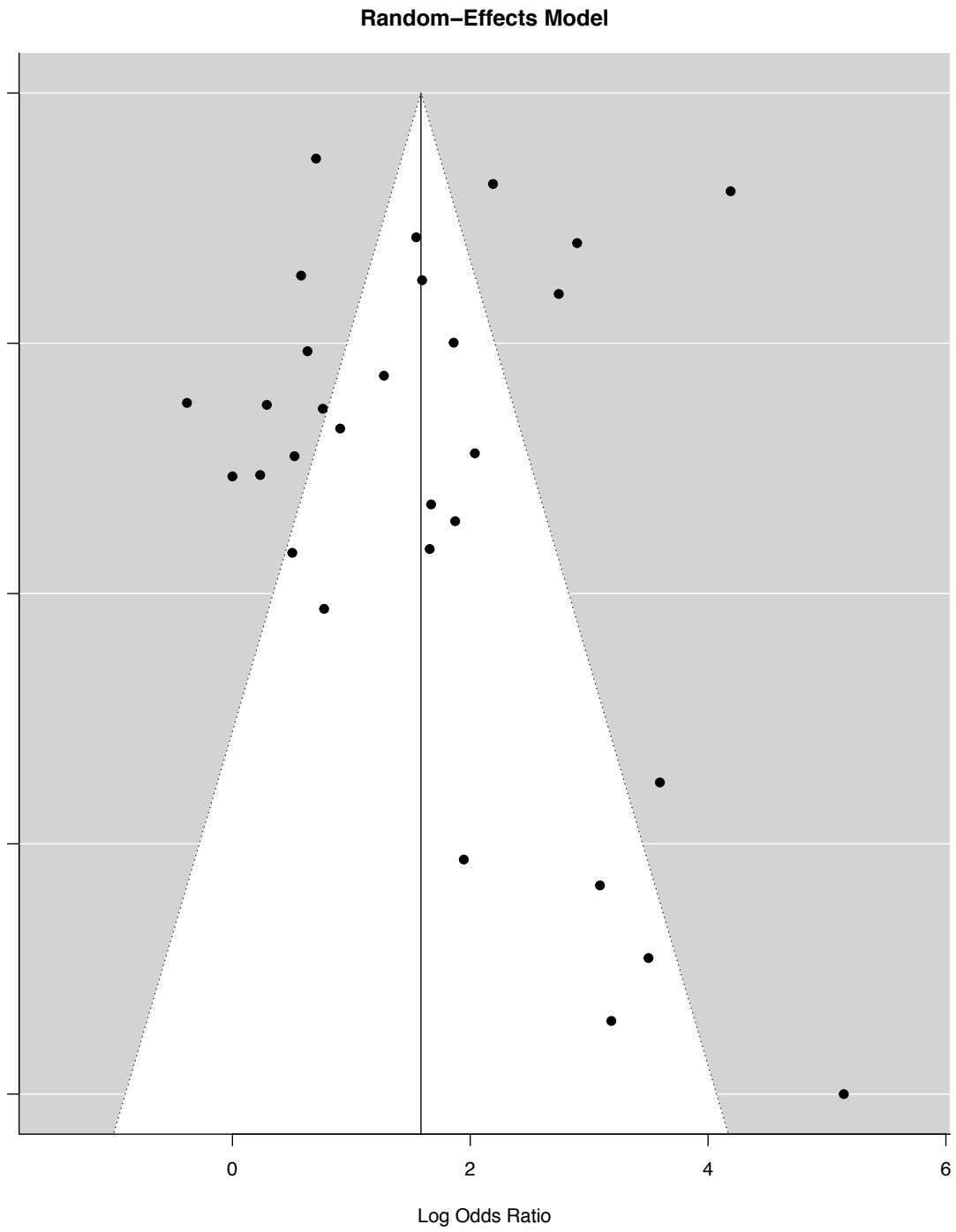


Figure 6: Funnel plot for categorical studies



3. Discussion

Our meta-analysis is clear. First, the means/byproduct effect does appear in everyday moral judgments. Across one hundred and one studies using different designs, samples, and types of measures of moral judgment, the asymmetry in responses typical of the Means Principle was medium to large. However, consistent with two experiments reported by Greene and his collaborators (2009), these overall effects were qualified by an important moderator: personal contact. When personal contact was not present, the pattern of response predicted by the Means Principle was dramatically reduced. The estimates of the mean effect size are also likely to be accurate and robust given the absence of evidence of publication bias.

So is the Doctrine of Double Effect, or something quite like it, tacitly assumed in ordinary moral cognition? Does such a principle have a legitimate claim to being an innate element of universal moral grammar? To address the theoretical debates about the means/byproduct distinction, we must determine the significance of its being moderated by personal contact. There are three kinds of theoretical conclusions one might draw concerning the Means Principle.

4.1 *Skeptical Conclusion*

One might argue that the Means Principle alone doesn't shape ordinary moral cognition. The effect is sufficiently robust only when combined with personal contact or a related factor. Perhaps the small effect one can detect should be construed as a performance error or otherwise not a phenomenon on which we should build our theories of moral psychology. Greene (2013: 246), for example, argues that the means/side-effect distinction, along with act/omission and personal force, "are not three separate criteria, employed in checklist fashion." Instead, these factors are "intertwined... forming an organic whole" that triggers a psychological "alarm gizmo" which is sensitive to *prototypically violent* acts.

Given that the Doctrine of Double Effect relies on the Means Principle alone and not on its interaction with anything like personal contact or prototypical violence, the skeptical conclusion has important implications. Contrary to many theorists, Double Effect wouldn't have "considerable intuitive appeal" (McIntyre 2001: 219); wouldn't be able to "explain some otherwise puzzling cases" (Scanlon 2008: 1); wouldn't be presupposed by a "range of moral beliefs that have been comparatively stable over time and across cultures" (Wedgwood 2011: 392); and wouldn't be an element of an innate "moral grammar" (Mikhail 2011). Moreover, if the Means Principle isn't implicit in our basic modes of moral thinking, then we lose a presumptive reason for believing that Double Effect and similar principles are ethically and legally sound. The skeptical conclusion may thus support attempts to debunk such principles as morally suspect or unreliable (e.g. Greene 2013; Sinhababu 2013). Even if debunkers have nothing to undermine in ordinary moral thinking, its absence may still help to undermine the principle's plausibility.

However, the skeptical conclusion is doubtful given that the means/byproduct effect is not entirely moderated by personal contact. While violent acts may typically involve harming as a means in an active and personal way, the means/byproduct effect appears to arise when the relevant outcome isn't even harmful, active, or up-close and personal. The pure unmoderated effect may be small, but our meta-analysis suggests it is real, even when using a variety of

measures of moral cognition. Some studies failed to detect the pure effect but this is most likely because those studies are not sufficiently powered.

4.2 *Non-Skeptical Conclusion*

A second possible conclusion is decidedly *non-skeptical*: the means/byproduct distinction does shape ordinary moral cognition and personal contact helps to facilitate it. Personal contact could, for example, merely draw one's attention to the fact that an outcome is brought about as a means to an end, which one does treat as independently significant (cf. Locke ms). On this account, personal contact and related factors merely *enable* one to detect or notice that the outcome is generated as a means rather than a byproduct.

Compare how this might go in a non-moral example. Your judgment that an object is red is influenced by how it visually appears, which is heavily moderated by whether your eyes are open. However, while your perceptual experience is treated as relevant to judging the color of the object, having your eyes open isn't. Indeed, your visual experience arguably provides a reason for believing that the object is red while the fact that your eyes are open doesn't. One fact is a reason while the other is a mere "enabling condition" (cf. Burge 1988: 654-5). Similarly, personal contact may merely help many people focus on the factor they treat as morally relevant: causing as a means.

Our meta-analysis is compatible with this non-skeptical account. However, proponents of Double Effect and related principles don't typically take the means/byproduct distinction to be difficult to appreciate, let alone noticeable only when personal contact is also contrasted. So our meta-analysis suggests that a non-skeptical account requires some theoretical revision.

4.3 *Moderately Skeptical Conclusion*

A third kind of conclusion is somewhere in between the two extremes and is thus *moderately skeptical*: moral judgment is influenced by whether a harm is brought about by an action that involves *both* personal contact and harming as a means. Why might we care primarily about the combination of these two factors? We can think of four possible theories.

First, we may have evolved to be sensitive to prototypically violent acts, which involve actively and personally causing harm as a means (cf. Greene 2013: ch. 9). While the Means Principle does not necessarily have independent significance in moral cognition, it interacts with something like personal contact, making harming as a means an essential element of a broader category. The account is moderately skeptical because it does undermine the idea that the Means Principle itself is an element of moral cognition, which may fund attempts to undermine its normative significance. However, it's important to consider that some of the studies reporting a means/byproduct effect involve non-violent actions, such as property damage (e.g. Millar et al 2014), financial loss (e.g. Gold et al 2013), and minor impersonal harm (e.g. Kelman & Kreps 2014; May ms-b). Unfortunately, since too few studies use such stimuli, we couldn't meaningfully test for violence as a moderator. Further research could help adjudicate these issues.

Second, instead of prototypical violence, the more general factor that influences moral cognition could be something like *agential involvement*—the degree to which the agent is involved in generating the consequences (Wedgwood 2011; May forthcoming). Even if harming as a means doesn't by itself influence moral thinking, there is some evidence that it contributes to our perception that the individual intentionally generated the relevant outcome (cf. Cushman

& Young 2011). Like impersonal contact and omissions, generating an outcome as a side effect may contribute to the perception that the agent was less involved (compare Foot 1984), constituting one way to “get one’s hands dirty.” Like the appeal to prototypical violence, this view is moderately skeptical because the means/byproduct distinction is only significant insofar as it’s bound up with the broader category of agential involvement.

A third version of the moderately skeptical account involves individual differences. Perhaps for *some people* the means/byproduct distinction is prominent whereas for others it is not. There is growing evidence that there are important individual differences in people’s philosophically relevant judgments (Feltz & Cokely in press; 2013). To illustrate, imagine that for roughly half of people the means/byproduct distinction is strong (e.g., *standardized mean difference* = 1). However, for the other half of people the distinction is completely absent (e.g., *standardized mean difference* = 0). If we average across those groups of people in the experiment, we would get a standardized mean difference of about 0.5—roughly what we actually observe. Consequently, the meta-analysis is consistent with a moderately skeptical account where the distinction does matter but only for *some people*.

A final route to moderate skepticism about the Means Principle restricts its scope. Empirical and theoretical investigations of the means/byproduct effect have treated such principles as being unrestricted or as applying to all moral norms. Perhaps, though, the principle is relevant to only some types of norms (compare Young & Saxe 2011; Walen 2016). While many of the experimental vignettes have involved dilemmas of life and death, perhaps these are precisely the situations in which it makes least sense to draw the distinction. In general, we don’t seem to think people are entitled to risk causing the death of others, either as a means or as a byproduct. However, it seems more important to distinguish whether someone violates other norms purposefully or merely knowingly, particularly when we think people are generally entitled to risk the negative outcomes involved in violating the norm, such as pollution or property damage. Yet these are some of the studies in which we find a means/byproduct effect in the absence of personal contact (e.g. DeScioli et al 2012; Millar et al 2014; May ms-b). Again, unfortunately few studies differ in this respect, so we could not test for different norm-violations as moderators. Further research could help to settle the matter.

This last account also fits with some findings that have implications for hotly debated topics in applied ethics. For example, in the euthanasia debate, it is often thought that bringing about a death as a goal is worse than bringing about a death as a merely foreseen consequence. Some studies suggest that this general pattern is not reflected in everyday judgments about the moral permissibility of some kinds of euthanasia. One series of experiments systematically manipulated the description of euthanasia (e.g., physician assisted suicide, aid in dying, euthanasia) but found no difference in moral responses when the death was brought about as a means as opposed to a foreseen side-effect (Feltz, in press; but compare DeScioli et al 2012). Rather, the major factor that distinguishes moral judgments about euthanasia is the voluntariness of the death. When the patient specifically requests death, it is deemed to be much more permissible than when it is not brought about voluntarily. This general pattern of results has been replicated using recent scales for measuring attitudes about euthanasia (Feltz & Cokely, ms; Ho, 1988). Thus, the means/byproduct distinction might not be as important as many have thought for explaining certain contemporary moral disputes. If the distinction has any importance in commonsense moral reasoning, it may be restricted to only certain norms. Further studies could address this issue by extending existing research (e.g. DeScioli 2012; Millar et al 2014) that

probes intuitions about real moral disputes (e.g. abortion, torture, casualties of war, affirmative action), rather than just trolley dilemmas.

4. Conclusion

On the basis of experimental findings, many scientists, policy-makers, and philosophers have concluded that people's moral intuitions consistently treat bringing about a bad outcome as a means to an end as much worse than doing so as a mere side effect. Killing one innocent person to save five, for example, is apparently conceived as more morally acceptable if the death of the one is a mere side effect of saving the five rather than an integral part of one's plan to promote the greater good. Proponents of the relevant principle, such as the Doctrine of Double Effect, have praised the empirical results as demonstrating its importance in moral and legal reasoning. Opponents who have sought to debunk a principle like Double Effect have likewise assumed its presence in moral cognition.

The present meta-analysis, however, was motivated by failures to replicate the means/byproduct effect and by its frequently small size when reported. We found that the effect is robust but particularly when combined with personal contact. For example, people generally think it's morally impermissible to kill one innocent person as a means to saving five when the killing involves pushing the one to his death. But we're less inclined to regard this as morally dubious if the man can be used as a means by merely flipping a switch.

The significance of the moderator result is difficult to determine at present. Some may take it as a death knell to Double Effect and its attendant Means Principle, but there is room to argue for less skeptical conclusions. Personal contact might merely draw one's attention to the means/byproduct distinction, for example. Our meta-analysis alone cannot discern which conclusion is correct, but it does constrain the kinds of options available and guide future research. What's clear is that the means/byproduct distinction does shape ordinary moral thinking, although its precise role has yet to be fully uncovered.

Acknowledgements: We would like to thank Amy Joy Patterson for helping collect the data for the studies. For discussion and feedback, we thank Nick Byrd, Jason Shephard, Hanne Watkins, and the anonymous referees for this journal. Funding for some of this research was provided by a Dean's Humanities Grant from the University of Alabama at Birmingham.

Ethics Approval: The individual unpublished studies conducted by May were approved by the IRB at the University of Alabama at Birmingham (Protocol X140528011).

References

References with an asterisk mark studies included in the meta-analysis.

- * Abarbanell, L., & Hauser, M. D. (2010). Mayan morality: An exploration of permissible harms. *Cognition* 115(2): 207–224.
- * Ahlenius, H., & Tännjö, T. (2012). Chinese and Westerners Respond Differently to the Trolley Dilemmas. *Journal of Cognition and Culture*, 12(3-4), 195-201.
- * Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means visual imagery and moral judgment. *Psychological science*, 23(8), 861-868.
- * Broeders, R., Van Den Bos, K., Müller, P. A., & Ham, J. (2011). Should I save or should I not kill? How people solve moral dilemmas depends on which rule is most accessible. *Journal of Experimental Social Psychology*, 47(5), 923-934.
- Burge, T. (1988). Individualism and self-knowledge. *The Journal of Philosophy*, 649-663.
- * Cao, F., Zhang, J., Song, L., Wang, S., Miao, D., & Peng, J. (2017). Framing Effect in the Trolley Problem and Footbridge Dilemma: Number of Saved Lives Matters. *Psychological Reports* 120(1), 88–101.
- * Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apestequia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PloS one*, 9(4), e94842.
- * Côté, S., Piff, P. K., & Willer, R. (2013). For whom do the ends justify the means? Social class and utilitarian moral judgment. *Journal of Personality and Social Psychology*, 104(3), 490-503.
- * Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: a moral dilemma validation study. *Frontiers in psychology*, 5(607), 1-18.
- Cushman, F. A. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2), 353–380.
- Cushman, F. (2016). “The Psychological Origins of the Doctrine of Double Effect.” *Criminal Law and Philosophy* 10(4): 763-776.
- * Cushman, F., & Young, L. (2011). “Patterns of Moral Judgment Derive from Nonmoral Psychological Representations.” *Cognitive Science* 35(6): 1052-1075.
- * Cushman, F., Young, L., & Hauser, M. (2006). “The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm.” *Psychological Science*, 17(12), 1082–1089.
- * DeScioli, P., Asao, K., & Kurzban, R. (2012). “Omissions and Byproducts Across Moral Domains.” *PloS One* 7: e46963.
- Enoch, D. (2013). “On Analogies, Disanalogies, and Moral Philosophy: A Comment on John Mikhail's Elements of Moral Cognition.” *Jerusalem Review of Legal Studies* 8: 1-25.
- Feltz, A. (in press). Everyday attitudes about euthanasia and the slippery slope argument. In M. Cholbi & J. Varelius (Eds.), *New Directions in the Ethics of Assisted Suicide*. New York: Springer.
- Feltz, A., & Cokely, E.T. (2013). Predicting philosophical disagreement. *Philosophy Compass*, 8/10, 978-989.
- Feltz, A. & Cokely, E.T. (in press). Personality and philosophical bias. In J. Sytsma & W. Buckwalter (Eds.) *The Blackwell Companion to Experimental Philosophy*.
- Feltz, A., & Cokely, E.T. (ms). The General Euthanasia Scale.
- Fitzpatrick, S. (2014). “Distinguishing Between Three Versions of the Doctrine of Double Effect Hypothesis in Moral Psychology.” *Review of Philosophy & Psychology* 5(4):505-525.
- Foot, P. (1967). “The Problem of Abortion and the Doctrine of the Double Effect.” *Oxford Review*, 5, 5–15.
- Foot, P. (1984). “Killing and Letting Die.” *Abortion: Moral and Legal Perspectives*, J. L. Garfield and P. Hennessey (eds.), 177–85. Amherst, MA: University of Massachusetts Press.
- * Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Spota, S., Zago, S., et al. (2010). Gender-related differences in moral judgments. *Cognitive processing*, 11(3), 219-226.

- * Gold, N., Pulford, B. D., & Colman, A. M. (2013). Your Money or Your Life: Comparing Judgements in Trolley Problems Involving Economic and Emotional Harms, Injury and Death. *Economics and Philosophy* 29(02): 213–233.
- * Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2009). “Pushing moral buttons: The interaction between personal force and intention in moral judgment.” *Cognition*, 111 (3): 364 –371.
- Greene, J. (2013). *Moral Tribes*. Penguin Press.
- * Hauser, M., Cushman, F., Young, L., Jin, R., J. Mikhail (2007). “A Dissociation Between Moral Judgments and Justifications.” *Mind and Language* 22(1):1–21.
- Ho, R. (1998). Assessing attitudes toward euthanasia: an analysis of the subcategorical approach to right to die issues. *Personality and Individual Differences*, 25(4), 719-734.
- * Horne, Z., & Powell, D. (2016). How Large Is the Role of Emotion in Judgments of Moral Dilemmas? *PLoS ONE*, 11(7), e0154780.
- * Kawai, N., Kubo, K., & Kubo-Kawai, N. (2014). “Granny dumping”: Acceptability of sacrificing the elderly in a simulated moral dilemma. *Japanese Psychological Research* 56(3): 254–262.
- * Kelman, M., & Kreps, T. A. (2014). Playing with Trolleys: Intuitions About the Permissibility of Aggregation. *Journal of Empirical Legal Studies*, 11(2), 197-226.
- Knobe, J. (2010). “Person as Scientist, Person as Moralist.” *Behavioral and Brain Sciences*, 33, 315-329.
- * Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social cognitive and affective neuroscience*, 7(6), 708-714.
- * Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908-911.
- * Laham, S. & Watkins, H. M. (2017). Cognitive Load and Moral Principles, Study 4. Unpublished manuscript.
- * Lanteri, A., Chelini, C., & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics*, 83(4), 789-804.
- * Liao, S. M., A. Wiegmann, J. Alexander, and G. Vong. (2012). “Putting the Trolley in Order: Experimental Philosophy and the Loop Case.” *Philosophical Psychology* 25: 661–671.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, Calif.: Sage Publications.
- Locke, D. (ms). The Normative Significance of Neuroscience Reconsidered. Unpublished manuscript.
- * Lombrozo, T. (2009). The Role of Moral Commitments in Moral Judgment. *Cognitive Science* 33(2): 273–286.
- Lopez, T., Zamzow, J., Gill, M., & Nichols, S. (2009). “Side Constraints and the Structure of Commonsense Ethics.” *Philosophical Perspectives* 23(1): 305-319.
- * Lotto, L., Manfrinati, A., & Sarlo, M. (2014). A new set of moral dilemmas: Norms for moral acceptability, decision times, and emotional salience. *Journal of Behavioral Decision Making*, 27(1), 57-65.
- * Manfrinati, A., Lotto, L., Sarlo, M., Palomba, D., & Rumiati, R. (2013). Moral dilemmas and moral principles: When emotion and cognition unite. *Cognition & emotion*, 27(7), 1276-1291.
- May, J. (2014). “Moral Judgment and Deontology: Empirical Developments.” *Philosophy Compass* 9(11): 745-755.
- May, J. (forthcoming). *Regard for Reason in the Moral Mind*. Oxford University Press.
- * May, J. (ms-a). “The Death of Double Effect?” Unpublished manuscript.
- * May, J. (ms-b). “Intuitive Moral Judgments and the Restricted Means Principle.” Unpublished data.
- McIntyre, A. (2001). “Doing Away with Double Effect.” *Ethics* 111(2): 219-255.
- Mikhail, J. (2002). Aspects of the Theory of Moral Cognition. *Public Law & Legal Theory Working Paper Series*. <http://ssrn.com/abstract=762385>
- * Mikhail, J. (2011). *Elements of Moral Cognition*. Cambridge University Press.

- Mikhail, J. (2014). "Any Animal Whatever? Harmful Battery and its Elements as Building Blocks of Moral Cognition." *Ethics* 124(4): 750-786.
- * Millar, J. C., Turri, J., & Friedman, O. (2014). "For the Greater Goods? Ownership Rights and Utilitarian Moral Judgment." *Cognition* 133: 79-84.
- * Moore, A. B., Clark, B. A., & Kane, M. J. (2008). "Who Shalt not Kill? Individual differences in working memory capacity, executive control, and moral judgment." *Psychological Science*, 19(6): 549-557.
- * Moore, A. B., Lee, N. L., Clark, B. A., & Conway, A. R. (2011a). In defense of the personal/impersonal distinction in moral psychology research: Cross-cultural validation of the dual process model of moral judgment. *Judgment and Decision Making*, 6(1), 186-195.
- * Moore, A. B., Stevens, J., & Conway, A. R. (2011b). Individual differences in sensitivity to reward and punishment predict moral judgment. *Personality and Individual Differences*, 50(5), 621-625.
- * Nakamura, K. (2012). The Footbridge Dilemma Reflects More Utilitarian Thinking Than The Trolley Dilemma: Effect Of Number Of Victims In Moral Dilemmas. *Proceedings of the Thirty-fourth Annual Conference of the Cognitive Science Society*.
- Nelkin, D. K. & Rickless, S. C. (2014). "Three Cheers for Double Effect." *Philosophy and Phenomenological Research* 89(1): 125-158.
- Nelkin, D. K., & Rickless, S. C. (2015). "So Close, yet So Far: Why Solutions to the Closeness Problem for the Doctrine of Double Effect Fall Short." *Noûs* 49(2): 376-409.
- * Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition* 100(3): 530-542.
- O'Hara, R. E., Sinnott-Armstrong, W., & Sinnott-Armstrong, N. A. (2010). Wording effects in moral judgments. *Judgment and Decision Making*, 5(7), 547-554.
- Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8, 157-159.
- * Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental Science* 13(2): 265-270.
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- * Rusch, H. (2015). Do Bankers Have Deviant Moral Attitudes. Negative Results from a Tentative Survey. *Rationality, Markets and Morals*, 6, 6-20.
- Sarch, A. (in press). Double Effect and the Criminal Law. *Criminal Law and Philosophy*.
- * Sarlo, M., Lotto, L., Manfrinati, A., Rumiati, R., Gallicchio, G., & Palomba, D. (2012). Temporal dynamics of cognitive-emotional interplay in moral decision-making. *Journal of Cognitive Neuroscience*, 24(4), 1018-1029.
- Scanlon, T. M. (2008). *Moral Dimensions*. Harvard University Press.
- * Schwitzgebel, E., & Cushman, F. A. (2012). Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers. *Mind & Language* 27(2): 135-153.
- * Shallow, C., Iliev, R., & Medin, D. (2011). Trolley problems in context. *Judgment and Decision Making* 6(7): 593-601.
- * Shepard, J. (2016). Unpublished raw data, untitled work. Department of Psychology, Emory University, Atlanta, GA.
- Sinhababu, N. (2013). "Unequal Vividness and Double Effect" *Utilitas* 25(3): 291-315.
- * Sinnott-Armstrong, W. Mallon, R., McCoy, T., & Hull, J. G. (2008). "Intention, Temporal Order, and Moral Judgments." *Mind & Language* 23(1): 90-106.
- * Spranca, M., Minsk, E., & Baron, J. (1991). "Omission and Commission in Judgment and Choice." *Journal of Experimental Social Psychology* 27: 76-105.
- * Sytsma, J. & Livengood, J. (ms) Intervention, Bias, Responsibility... and the Trolley Problem. PhilSci Archive. <http://philsci-archive.pitt.edu/12283/>

- * Tempesta, D., Couyoumdjian, A., Moroni, F., Marzano, C., De Gennaro, L., & Ferrara, M. (2012). The impact of one night of sleep deprivation on moral judgments. *Social neuroscience*, 7(3), 292-300.
- Thomson, J. J. (1992). *The Realm of Rights*. Harvard University Press.
- * Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion*, 12(3), 579-590.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36 (3), 1–48.
- * Waldmann, M., & Dieterich, J. (2007). “Throwing a Bomb on a Person versus Throwing a Person on a Bomb: Intervention Myopia in Moral Intuitions.” *Psychological Science*, 18: 247–253.
- * Waldmann, M. R., & Wiegmann, A. (2010). A Double Causal Contrast Theory of Moral Intuitions in Trolley Dilemmas. *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- * Wiegmann, A., Lippold, M., & Grigull, R. (2013). On the Robustness of Intuitions in the two best-known Trolley Dilemmas. *Proceedings of the Annual Conference of the Cognitive Science Society*, 3759–3764.
- * Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, 25(6), 813-836.
- * Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, 131(1), 28-43.
- Walen, A. (2016). “The Restricting Claims Principle Revisited: Grounding the Means Principle on the Agent–Patient Divide.” *Law and Philosophy* 35(2): 211–247.
- * Watkins, H. M. & Laham, S. (2017). Cognitive Load and Moral Principles, Study 5. Unpublished data.
- * Watkins, H. M. (2016). The Moral Psychology of Killing in War (Doctoral dissertation, The University of Melbourne, Australia).
- * Watkins, H. M., & Laham, S. M. (2016). An investigation of the use of linguistic probes "by" and “in order to” in assessing moral grammar. *Thinking & Reasoning* 22(1), 16–30.
- Wedgwood, R. (2011). “Defending Double Effect.” *Ratio* 24(4):384–401.
- * Young, L., & Saxe, R. (2011). “When Ignorance is No Excuse: Different Roles for Intent Across Moral Domains.” *Cognition* 120(2): 202-214.
- * Young, L. & Tsoi, L. (2013). “When Mental States Matter, When They Don’t, and What That Means for Morality.” *Social and Personality Psychology Compass* 7(8): 585–604.
- * Zimmerman, A. (2013). “Mikhail’s Naturalized Moral Rationalism.” *Jerusalem Review of Legal Studies* 8: 44-65.