The Moral Dilution Effect: Irrelevant Information Influences Judgments of Moral Character

Cillian McHugh[1] & Eric R. Igou[1]

[1] University of Limerick

Author Note

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

Abstract

Across five studies we investigated the moral dilution effect

Keywords: keywords

Word count: TBC

The Moral Dilution Effect: Irrelevant Information Influences Judgments of Moral Character

Recent developments in the psychology of moral judgments propose that making a moral judgment is an act of categorization (McHugh, McGann, Igou, & Kinsella, 2022). A corollary of this view is that moral judgments vary depending on how well a target maps onto a template of a prototypical morally relevant act/actor (Schein & Gray, 2018)

Pilot Study 1

The aim of this pilot study was to develop and test materials that could be used to study the dilution effect for moral characters. We developed diagnostic and non-diagnostic character descriptions. We hypothesized that moral evaluations of the diagnostic descriptions would be more severe (more immoral) than for the non-diagnostic descriptions.

Pilot Study 1: Method

Pilot 1: Participants and design

The pilot study was a within-subjects design. The independent variable was description type with two levels, *diagnostic* and *non-diagnostic*. We used two dependent variables. The first dependent variable was the four item moral perception scale (MPS-4), participants rated the characters on four dimensions using 7-point bipolar scales. The dimensions and scale endpoints were: Bad-Good, Immoral-Moral, Violent-Peaceful, Merciless-Empathetic, this showed excellent reliability, $\alpha = 0.93$. The second dependent variable was a single item moral perception measure (MM-1) which consisted of a 100-point slider ranging from 0 = *Very Immoral* to 100 = *Very Moral*. Both dependent variables were taken from Walker et al. (2021).

A total sample of 235 (89 female, 142 male, 1 non-binary, 1 prefer not to say; M_{age} = 36.45, min = 20, max = 72, SD = 10.23) started the survey. Participants were recruited from MTurk.

We removed participants who failed both manipulation checks ($n = 23$), leaving a total sample of 212 participants (80 female, 128 male, 1 non-binary, 1 prefer not to say; $M_{\text{age}} = 36.63$, min = 20, max = 72, $SD = 10.34$).

Pilot 1: Procedure and materials

Data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with descriptions of six characters.

Moral character descriptions were developed by combining descriptions relating to three different moral foundations. These descriptions were adapted from the items of the extended character morality questionnaire (Grizzard et al., 2020). A sample description reads: *Imagine a person named Sam. Throughout their life they have been known to be cruel, act unfairly, and to betray their own group.* Full text of these descriptions can be found in the supplementary materials.

We developed neutral descriptions that included information relating to physical appearance/attributes, hobbies/activities, and family information, e.g., *Imagine a person named Jackie. They have red hair, play tennis four times a month, and have one older sibling and one younger sibling.*

Character descriptions did not specify the gender of the charcters, and all characters had names that could be either male or female (Sam, Robin, Francis, Alex, Jackie, Charlie). All participants read six descriptions, four moral descriptions and two neutral. Pilot Study 1 was pre-registered at https://aspredicted.org/3VK_8FD.

Pilot 1: Results

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (diagnostic), $M_{\text{MPS-4}} = 4.35$, $SD_{\text{MPS-4}} = 1.90$, *Francis* (diagnostic), $M_{\text{MPS-4}} = 4.46$, $SD_{\text{MPS-4}} = 1.73$, *Alex* (diagnostic), $M_{\text{MPS-4}} = 4.44$, $SD_{\text{MPS-4}} = 1.79$, *Robin* (diagnostic), $M_{\text{MPS-4}} = 4.35$, $SD_{\text{MPS-4}} = 1.96$, *Jackie* (non-diagnostic), $M_{\text{MPS-4}} = 5.40$, $SD_{\text{MPS-4}} = 1.01$,

Charlie (non-diagnostic), $M_{\text{MPS-4}} = 5.38$, $SD_{\text{MPS-4}} = 1.01$. For the diagnostic descriptions, there was no significant variation depending on the description, $F(3,600) = 1.58$, $p = .194$, partial $\eta^2 = 0.00$. For the non-diagnostic descriptions there was no significant difference in ratings depending on description, $t(211) = -0.67$, $p = .506$, $d = 0.05$.

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (diagnostic), $M_{\text{MM-1}} = 55.67$, $SD_{\text{MM-1}} = 30.47$; *Francis* (diagnostic), $M_{\text{MM-1}} = 58.22$, $SD_{\text{MM-1}} = 28.61$; *Alex* (diagnostic), $M_{\text{MM-1}} = 56.80$, $SD_{\text{MM-1}} = 29.45$; *Robin* (diagnostic), $M_{\text{MM-1}} = 55.49$, $SD_{\text{MM-1}} = 31.38$; *Jackie* (non-diagnostic), $M_{\text{MM-1}} = 73.00$, $SD_{\text{MM-1}} = 14.72$; *Charlie* (non-diagnostic), $M_{\text{MM-1}} = 72.94$, $SD_{\text{MM-1}} = 14.79$. For the diagnostic descriptions, we observed significant variation depending on the description, $F(3,608) = 3.01$, $p = .032$, partial $\eta^2 = 0.001$. When correcting for multiple comparisons, pairwise comparisons did not reveal significant differences between descriptions. We note that without correction, *Francis* appeared to be rated as more moral than both *Robin* ($p = .012$), and *Sam* ($p = .009$). For the non-diagnostic descriptions there was no significant difference in ratings depending on description, $t(211) = -0.09$, $p = .929$, $d = 0.01$.

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(2) = 860.16$, $p < .001$. Condition was a significant predictor in the model $b = -0.49$, $t(211.05) = -8.54$, $p < .001$, with the non-diagnostic (neutral) descriptions being rated as more moral than the diagnostic (morally relevant) descriptions of immoral characters Figure 1.

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model

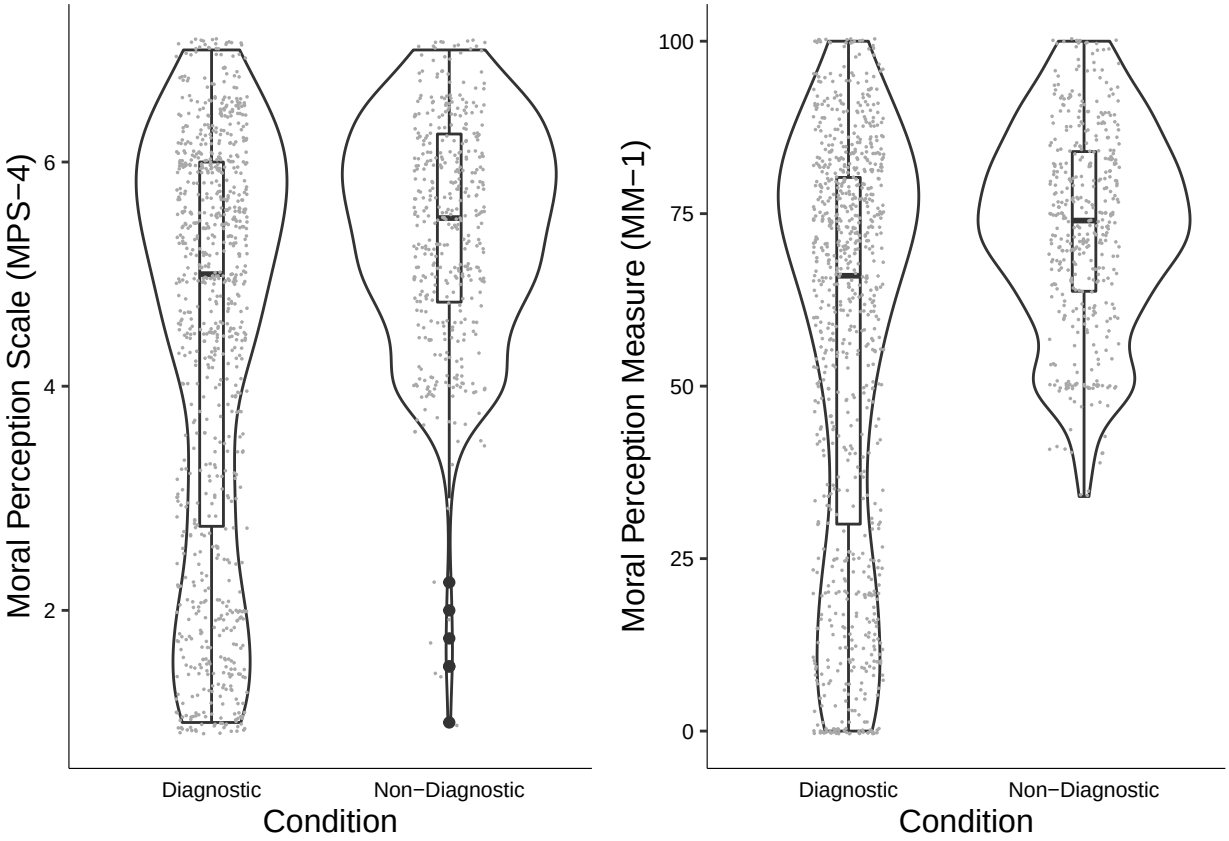

Figure 1

Pilot Study 1: Differences in moral perception depending on condition

significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(2) = 924.82$, $p < .001$. Condition was a significant predictor in the model b = -8.22, $t(210.98)$ = -8.60, $p < .001$, with the non-diagnostic (neutral) descriptions being rated as more moral than the diagnostic (morally relevant) descriptions, see Figure 1.

Study 1 - Bad Characters

Study 1: Method

The aim of Study 1 is to test if the dilution effect exists in the moral domain. Participants were presented with descriptions of four characters, two descriptions will only contain diagnostic information (morally relevant information) and two will additionally contain non-diagnostic information (non morally relevant information) along with the

diagnostic information. We hypothesize that moral perceptions of the diagnostic only descriptions will be more severe than for the descriptions that also contain non-diagnostic information.

Study 1: Participants and design

Study 1 was a within-subjects design. The independent variable was condition with two levels, diagnostic information only (diagnostic), and non-diagnostic information additionally included (non-diagnostic). We used the same two dependent variables as in Pilot Study 1, the four item moral perception scale (MPS-4) which showed good reliability, $\alpha = 0.83$, and the single item moral perception measure MM-1.

A total sample of 901 (302 female, 523 male, 0 non-binary, 5 other; 2 prefer not to say, $M_{\text{age}} = 26.16$, min = 18, max = 76, $SD = 10.14$) started the survey. Participants were recruited from the student population at University of [BLINDED].

Participants who failed both manipulation checks were removed ($n = 100$), leaving a total sample of 801 participants (283 female, 496 male, 5 other, 5 prefer not to say; $M_{\text{age}} = 26.25$, min = 18, max = 76, $SD = 10.20$).

Study 1: Procedure and materials

As in the pilot study, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of characters (*Sam*, *Alex*, *Francis*, *Robin* from Pilot Study 1). All descriptions included diagnostic information relating to three moral foundations, e.g., *Imagine a person named Robin. Throughout their life they have been known to physically hurt others, treat some people differently to others, and show lack of loyalty.* We programmed our survey to randomly present non-diagnostic information along with two of the descriptions participants read (this was done through blocking, for details on the blocks see full materials at https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67. This

meant that all participants read two descriptions containing diagnostic information only, and two descriptions that additionally included non-diagnostic information. We hypothesized that the descriptions including non-diagnostic information would be rated as less severe than the diagnostic-only descriptions. Study 1 was pre-registered at https://aspredicted.org/DVY_QN3

Study 1: Results

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam*, $M_{\text{MPS-4}} = 2.55$, $SD_{\text{MPS-4}} = 0.86$, *Francis*, $M_{\text{MPS-4}} = 3.05$, $SD_{\text{MPS-4}} = 0.97$, *Alex*, $M_{\text{MPS-4}} = 2.32$, $SD_{\text{MPS-4}} = 0.88$, *Robin*, $M_{\text{MPS-4}} = 2.13$, $SD_{\text{MPS-4}} = 0.91$. There was significant variation depending on the description, $F(3,2280) = 297.82$, $p < .001$, partial $\eta^2 = 0.13$. *Francis* appeared to be rated as more moral than each of the other characters (all ps $< .001$), while *Robin* was rated as less moral than each of the other characters (all ps $< .001$), while *Sam* was rated more favorably than *Alex* ($p < .001$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam*, $M_{\text{MM-1}} = 23.94$, $SD_{\text{MM-1}} = 16.18$; *Francis*, $M_{\text{MM-1}} = 30.12$, $SD_{\text{MM-1}} = 17.86$; *Alex*, $M_{\text{MM-1}} = 20.55$, $SD_{\text{MM-1}} = 16.65$; *Robin*, $M_{\text{MM-1}} = 20.60$, $SD_{\text{MM-1}} = 17.06$. There was significant variation depending on the description, $F(3,2253) = 154.08$, $p < .001$, partial $\eta^2 = 0.05$. *Francis* was rated more favorably than all other characters ($p < .001$), *Sam* was the next most favorably rated character, rated significantly more favorably than both *Alex* and *Robin* (ps $< .001$), there was no difference between *Alex* and *Robin* ($p = = .953$).

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants, and scenario was also included in the model. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 816.91$, $p < .001$. Condition significantly influenced responses to the MPS-4, $F(1, 799.42) = 51.47$, p

$< .001$; and was a significant predictor in the model when controlling for scenario, $b =$ -0.08, $t(799.42) =$ -7.17, $p < .001$, with the diagnostic descriptions being rated as more immoral than the non-diagnostic descriptions Figure 2.

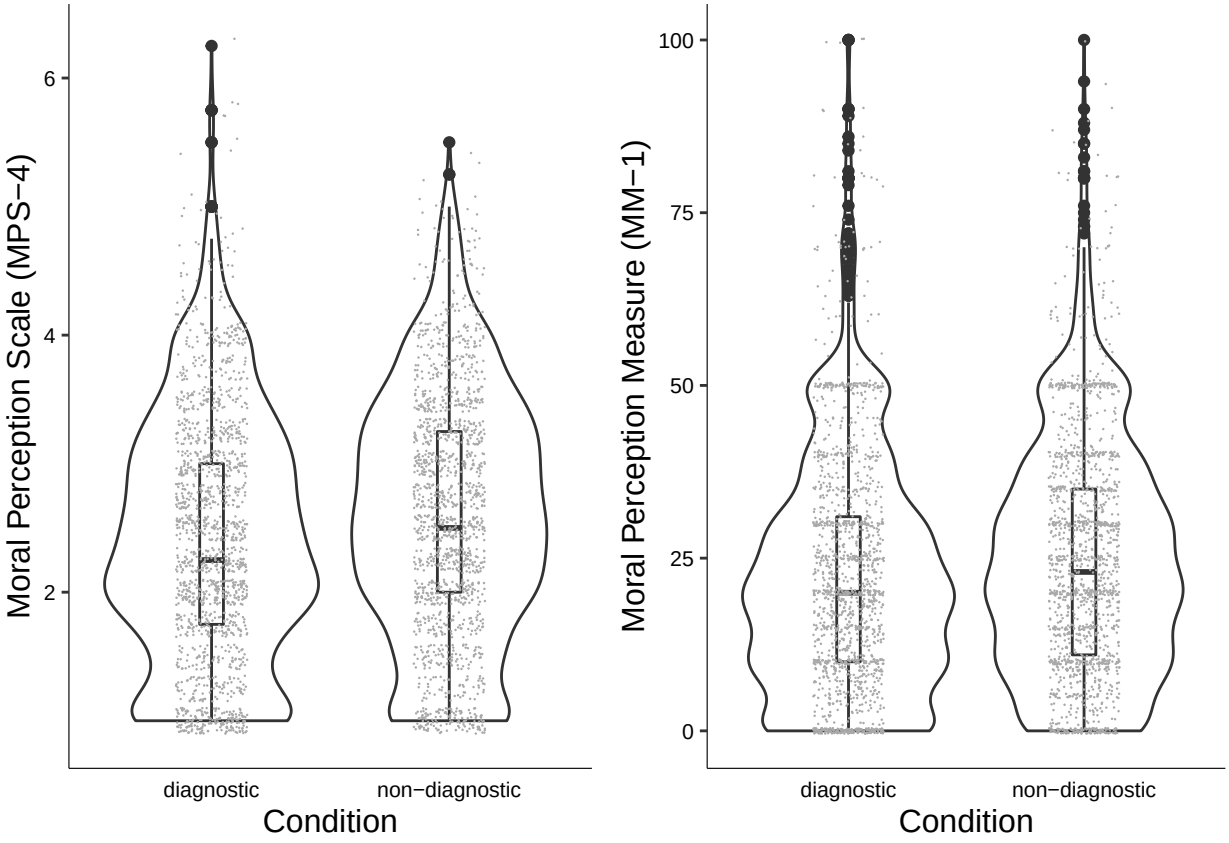

Figure 2

Study 1: Differences in moral perception depending on condition

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 475.52$, $p < .001$. Condition significantly predicted MM-1 responses $F(1, 799.71) = 44.39$, $p < .001$, and when controlling for scenario was a significant predictor in the model $b =$ -1.22, $t(799.71) =$ -6.66, $p < .001$, with the diagnostic descriptions being rated as more immoral than the non-diagnostic descriptions Figure 2.

In the supplementary analyses we report the effect of condition on moral perception for each description individually.

Pilot Study 2

Study 1 showed the moral dilution effect for judgments of *bad* characters. The aim of this pilot study was to develop and test materials that may be used to study the moral dilution effect for judgments of morally *good* characters. As with Pilot Study 1, we developed diagnostic and non-diagnostic descriptions. We hypothesized that evaluations of the diagnostic descriptions would be more extreme (more moral) than for the non-diagnostic descriptions

Pilot Study 2: Method

Pilot 2: Participants and design

The pilot study was a within-subjects design. The independent variable was description type with two levels, *diagnostic* and *non-diagnostic*. We used the same two dependent variables as in previous studies, the four item moral perception scale (MPS-4, $\alpha = 0.84$), and the single item moral perception measure (MM-1).

A total sample of 245 (70 female, 175 male, 0 non-binary, 0 prefer not to say; $M_{\text{age}} = 36.69$, min = 18, max = 71, $SD = 9.57$) started the survey. Participants were recruited from MTurk.

We removed participants who failed both manipulation checks ($n = 30$), leaving a total sample of 215 participants (63 female, 152 male, 0 non-binary, 0 prefer not to say; $M_{\text{age}} = 36.59$, min = 18, max = 71, $SD = 9.59$).

Pilot 2: Procedure and materials

Data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with descriptions of six characters.

Moral character descriptions were developed by combining descriptions relating to three different moral foundations, focusing on upholding the moral foundations (rather than transgressions as in previous studies). A sample description reads: *Imagine a person named Sam. Throughout their life they have been known to always help and care for others, treat everyone fairly and equally, and show a strong sense of loyalty to others.*. Full text of these descriptions can be found in the supplementary materials.

We developed neutral descriptions that included information relating to physical appearance/attributes, hobbies/activities, and a color preference, e.g., *Imagine a person named Charlie. They have blue eyes, drink coffee in the morning, and their favourite colour is green.*

We used the same gender ambiguous names, and we did not specify the gender of the characters. Pilot Study 2 was pre-registered at https://aspredicted.org/W52_VPX.

Pilot 2: Results

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (diagnostic), $M_{\text{MPS-4}} = 6.01$, $SD_{\text{MPS-4}} = 0.91$, *Francis* (diagnostic), $M_{\text{MPS-4}} = 5.89$, $SD_{\text{MPS-4}} = 0.95$, *Alex* (diagnostic), $M_{\text{MPS-4}} = 5.94$, $SD_{\text{MPS-4}} = 0.94$, *Robin* (diagnostic), $M_{\text{MPS-4}} = 5.93$, $SD_{\text{MPS-4}} = 0.92$, *Jackie* (non-diagnostic), $M_{\text{MPS-4}} = 5.60$, $SD_{\text{MPS-4}} = 0.99$, *Charlie* (non-diagnostic), $M_{\text{MPS-4}} = 5.53$, $SD_{\text{MPS-4}} = 1.08$. For the diagnostic descriptions, there was significant variation depending on the description, $F(3,613) = 2.91$, $p = .036$, partial $\eta^2 = 0.00$, *Sam* was viewed significantly more favorably than *Francis* ($p = .040$). For the non-diagnostic descriptions there was no significant difference in ratings depending on description, $t(214) = -1.79$, $p = .075$, $d = 0.12$.

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (diagnostic), $M_{\text{MM-1}} = 79.85$, $SD_{\text{MM-1}} = 15.44$; *Francis* (diagnostic), $M_{\text{MM-1}} = 78.30$, $SD_{\text{MM-1}} = 15.84$; *Alex* (diagnostic), $M_{\text{MM-1}} = 79.78$, $SD_{\text{MM-1}} = 15.71$; *Robin* (diagnostic), $M_{\text{MM-1}} = 79.46$, $SD_{\text{MM-1}} = 15.41$; *Jackie* (non-diagnostic), $M_{\text{MM-1}} = 73.44$, $SD_{\text{MM-1}} =$

15.83; *Charlie* (non-diagnostic), $M_{\text{MM-1}} = 73.07$, $SD_{\text{MM-1}} = 16.22$. For the diagnostic descriptions, we observed no significant variation depending on the description, $F(3,594) = 1.45$, $p = .231$, partial $\eta^2 = 0.002$. For the non-diagnostic descriptions there was no significant difference in ratings depending on description, $t(214) = -0.60$, $p = .552$, $d = 0.04$.

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(2) = 475.42$, $p < .001$. Condition was a significant predictor in the model $b = 0.19$, $t(214.35) = 6.53$, $p < .001$, with the diagnostic descriptions being rated as more moral than the non-diagnostic descriptions of immoral characters Figure 3.

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(2) = 324.13$, $p < .001$. Condition was a significant predictor in the model $b = 3.04$, $t(214.90) = 6.02$, $p < .001$, with the diagnostic descriptions being rated as more moral than the non-diagnostic descriptions, see Figure 3.

Study 2 - Good Characters

The aim of Study 2 is to test if the dilution effect exists in the moral domain for judgments of morally *good* characters. Participants were presented with descriptions of four characters, two descriptions contain diagnostic information only (morally relevant information) and two will additionally contain non-diagnostic information (non morally relevant information) along with the diagnostic information. We hypothesize that moral perceptions of the diagnostic only descriptions will be more extreme (more moral) than for

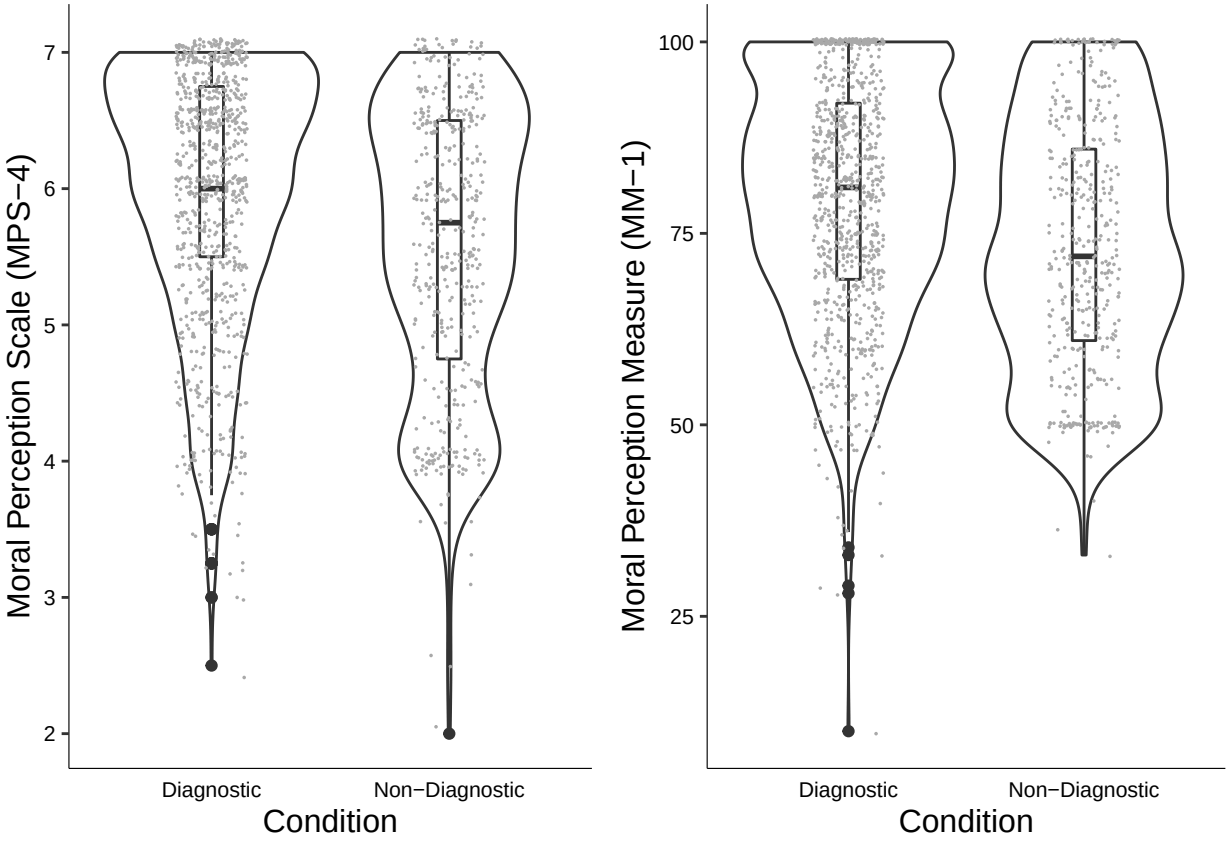

Figure 3

Pilot Study 2: Differences in moral perception depending on condition

the descriptions that also contain non-diagnostic information.

Study 2: Method

Study 2: Participants and design

Study 2 was a within-subjects design. The independent variable was condition with two levels, diagnostic and non-diagnostic. We used the same two dependent variables as in previous studies, the four item moral perception scale (MPS-4, $\alpha = 0.85$), and the single item moral perception measure MM-1.

A total sample of 1068 (418 female, 557 male, 13 non-binary, 2 other; 4 prefer not to say, $M_{\text{age}} = 29.04$, min = 18, max = 74, $SD = 10.66$) started the survey. Participants were

recruited from the student population at University of [BLINDED].

Participants who failed both manipulation checks were removed ($n = 248$), leaving a total sample of 820 participants (337 female, 466 male, 2 other, 2 prefer not to say; $M_{\text{age}} = 29.03$, min = 18, max = 74, $SD = 10.92$).

The majority of participants were from the student body: $n = 533$, (female = 370, male = 147, non-binary/other = 14, prefer not to say 3, $M_{age} = 25.50$, $SD = 9.60$).

In order to reach our pre-registered target sample size we recruited additional participants from MTurk: $n = 287$, (female = 96, male = 190, non-binary/other = 1, prefer not to say 1, $M_{age} = 35.70$, $SD = 10.10$). Participants from MTurk were paid $0.40 for their participation.

Study 2: Procedure and materials

Again, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of characters (*Sam*, *Alex*, *Francis*, *Robin* from Pilot Study 2). All descriptions included diagnostic information relating to three moral foundations, e.g., *Imagine a person named Alex. Throughout their life they have been known to protect and provide shelter to the weak and vulnerable, uphold the rights of others, and show respect for authority.* For each participant, two descriptions additionally included non-diagnostic information (this was randomized through blocking, see https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67. Study 1 was pre-registered at https://aspredicted.org/NX2_HN6

Study 2: Results

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam*, $M_{\text{MPS-4}} = 6.12$, $SD_{\text{MPS-4}} = 0.97$, *Francis*, $M_{\text{MPS-4}} = 5.86$, $SD_{\text{MPS-4}} = 1.07$, *Alex*, $M_{\text{MPS-4}} = 6.13$, $SD_{\text{MPS-4}} = 0.99$, *Robin*, $M_{\text{MPS-4}} = 6.10$, $SD_{\text{MPS-4}} = 0.99$. There was significant

variation depending on the description, $F(3,2356) = 54.47$, $p < .001$, partial $\eta^2 = 0.01$. *Francis* appeared to be rated as less moral than each of the other characters (all ps $< .001$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (diagnostic/moral), $M_{\text{MM-1}} = 84.60$, $SD_{\text{MM-1}} = 14.47$; *Francis* (diagnostic/moral), $M_{\text{MM-1}} = 82.05$, $SD_{\text{MM-1}} = 15.24$; *Alex* (diagnostic/moral), $M_{\text{MM-1}} = 85.02$, $SD_{\text{MM-1}} = 15.01$; *Robin* (diagnostic/moral), $M_{\text{MM-1}} = 84.95$, $SD_{\text{MM-1}} = 13.94$. There was significant variation depending on the description, $F(3,2387) = 24.20$, $p < .001$, partial $\eta^2 = 0.007$. *Francis* was rated less favorably than all other characters (all ps $< .001$).

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants, and scenario was also included in the model. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 160.00$, $p < .001$. Condition did not influence responses to the MPS-4, $F(1, 838.12) = 0.24$, $p = .624$; and was not a significant predictor in the model when controlling for scenario, $b = 0.00$, $t(838) = 0.49$, $p = .624$, see Figure 4.

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 75.69$, $p < .001$. Condition did not influence MM-1 responses $F(1, 2453) = 1.23$, $p = .267$, and was not a significant predictor in the model $b = 0.16$, $t(2453) = 1.11$, $p = .267$, see Figure 4.

In the supplementary analyses we report the effect of condition on moral perception for each description individually.

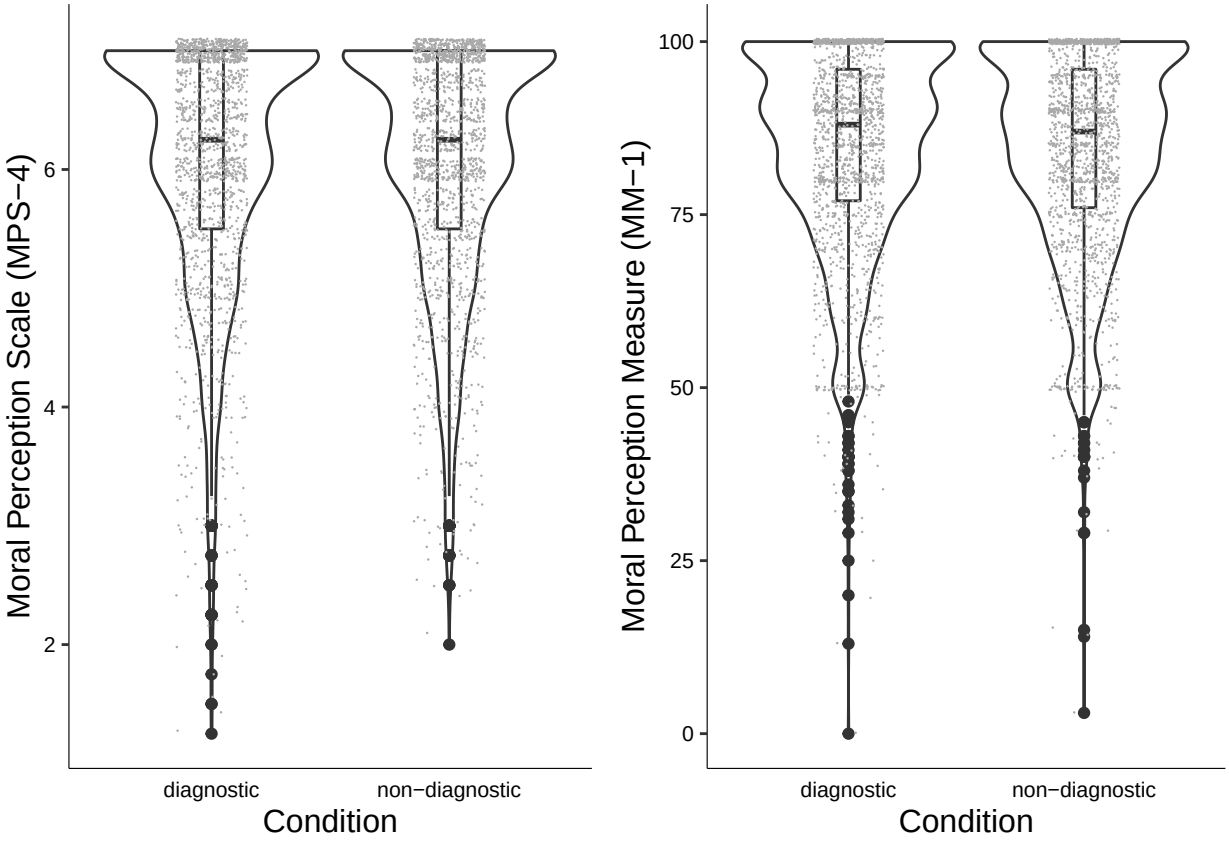

Figure 4

Study 2: Differences in moral perception depending on condition

Study 3 - Good and Bad Characters

In Study 1 we found evidence for the moral dilution effect for judgments of *bad* moral characters. In Study 2 we failed replicate this effect for judgments of *good* moral characters. The aim of Study 3 was to test if valence (good vs. bad) moderates the moral dilution effect. We hypothesized that valence (good vs bad) would interact with condition in producing a dilution effect, such that the dilution effect would be observed for bad characters but not for good characters. Study 3 was pre-registered at https://aspredicted.org/QDF_XT1.

Study 3: Method

Study 3: Participants and design

Study 3 was a 2 × 2 within-subjects factorial design. The first independent variable was condition with two levels, diagnostic and non-diagnostic. The second independent variable was valence of character description, with two levels morally good and morally bad. We used the same two dependent variables as in previous studies, the four item moral perception scale (MPS-4, $\alpha = 0.94$), and the single item moral perception measure MM-1.

A total sample of 1095 (700 female, 386 male, 2 non-binary, 0 other; 2 prefer not to say, $M_{\text{age}} = 36.42$, min = 19, max = 77, $SD = 10.65$) started the survey. Participants were recruited from MTurk and paid $0.40 for their participation.

Participants who failed both manipulation checks were removed ($n = 221$), leaving a total sample of 874 participants (550 female, 320 male, 0 other, 0 prefer not to say; $M_{\text{age}} = 36.37$, min = 19, max = 77, $SD = 10.72$).

Study 3: Procedure and materials

Again, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of characters taken from Studies 1 and 2. To ensure consistency across character judgments, we selected descriptions that related to the same moral foundations (care, fairness, and loyalty). We used the same four character names as in previous studies. The *good* characters were *Sam* and *Robin*, and the *bad* characters were *Francis* and *Alex*, e.g., *Imagine a person named Robin. Throughout their life they have been known to show compassion and empathy for others, act with a sense of fairness and justice, and, never to break their word.* or, *Imagine a person named Alex. Throughout their life they have been known to be cruel, act unfairly, and to betray their own group.* Full descriptions for each character are in the supplementary materials. One description for each the *good* and *bad* characters was

randomly assigned to include non-diagnostic information for each participant thus all participants were exposed to all conditions (see https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67for details of the randomization blocks). Study 3 was pre-registered at https://aspredicted.org/QDF_XT1

Study 3: Results

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (good), $M_{\text{MPS-4}} = 5.90$, $SD_{\text{MPS-4}} = 1.03$, *Francis* (bad), $M_{\text{MPS-4}} = 4.07$, $SD_{\text{MPS-4}} = 2.07$, *Alex* (bad), $M_{\text{MPS-4}} = 4.03$, $SD_{\text{MPS-4}} = 2.03$, *Robin* (good), $M_{\text{MPS-4}} = 5.85$, $SD_{\text{MPS-4}} = 1.05$. There was significant variation depending on the description, $F(1,1080) = 442.71$, $p < .001$, partial $\eta^2 = 0.24$. Both the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than both the *bad* characters (*Alex* and *Francis*; all *p*s $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .366$) or between *Alex* and *Francis* (*bad*; ($p = .648$)).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (good), $M_{\text{MM-1}} = 81.01$, $SD_{\text{MM-1}} = 15.23$; *Francis* (bad), $M_{\text{MM-1}} = 51.49$, $SD_{\text{MM-1}} = 33.18$; *Alex* (bad), $M_{\text{MM-1}} = 50.89$, $SD_{\text{MM-1}} = 32.14$; *Robin* (good), $M_{\text{MM-1}} = 80.81$, $SD_{\text{MM-1}} = 15.16$. There was significant variation depending on the description, $F(1,1080) = 458.92$, $p < .001$, partial $\eta^2 = 0.254$. Again, the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than the *bad* characters (*Alex* and *Francis*; all *p*s $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .776$) or between *Alex* and *Francis* (*bad*; ($p = .683$)).

We conducted a linear-mixed-effects model to test if our predictors influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variables were condition and valence; we allowed intercepts and the effects of condition and valence to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(5) = 3{,}420.34$, $p < .001$. As

expected, on its own, condition did not influence responses to the MPS-4 , $F(1, 1746) = 0.01$, $p = .937$; valence significantly predicted responses, , $F(1, 1746) = 587.37$, $p < .001$; and there was a significant condition × valence interaction, , $F(1, 1746) = 9.13$, $p = .003$. and was not a significant predictor in the model when controlling for scenario, $b = 0.00$, $t(1,746.00) = -0.08$, $p = .937$.

We conducted a linear-mixed-effects model to test if our predictors influenced MM-1 responses. The model was the same as the previous model, with a change to the outcome measure, our outcome measure for this model was MM-1. As above, our predictor variables were condition and valence; we allowed intercepts and the effects of condition and valence to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(5) = 3{,}441.43$, $p < .001$. As expected, on its own, condition did not influence responses to the MPS-4 , $F(1, 1746) = 0.03$, $p = .852$; valence significantly predicted responses, , $F(1, 1746) = 638.14$, $p < .001$; and there was a significant condition × valence interaction, , $F(1, 1746) = 17.23$, $p < .001$. and was not a significant predictor in the model when controlling for scenario, $b = -0.03$, $t(1,746.00) = -0.19$, $p = .852$.

Differences in the *Bad* Descriptions

To interpret the interaction effect, we conducted separate analyses for the Good and Bad descriptions.

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model did not significantly predict participants responses, or provide a better fit for the data than the baseline model, $\chi^2(3) = 5.40$, $p = .145$. Condition did not significantly influence MPS-4 responses $F(1, 872.00) = 3.54$, $p = .060$, and was not a significant predictor in the model $b = -0.03$, $t(872.00) = -1.88$, $p = .060$, see Figure 5.

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(3) = 8.67$, $p = .034$. Condition significantly influenced MM-1 responses $F(1, 872.00) = 7.01$, $p = .008$, and was a significant predictor in the model $b =$ -0.69, $t(872.00) =$ -2.65, $p = .008$, see Figure 5.

Differences in the *Good* Descriptions

To interpret the interaction effect, we conducted separate analyses for the Good and Bad descriptions.

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(3) = 13.66$, $p = .003$. Condition significantly influenced MPS-4 responses $F(1, 872.00) = 6.82$, $p = .009$, and was a significant predictor in the model $b =$ 0.03, $t(872.00) = 2.61$, $p = .009$, see Figure 5.

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(1) = 11.97$, $p < .001$. Condition significantly influenced MM-1 responses $F(1, 873) = 12.04$, $p < .001$, and was a significant predictor in the model $b =$ 0.63, $t(873) = 3.47$, $p < .001$, see Figure 5.

In the supplementary analyses we report the effect of condition on moral perception for each description individually.

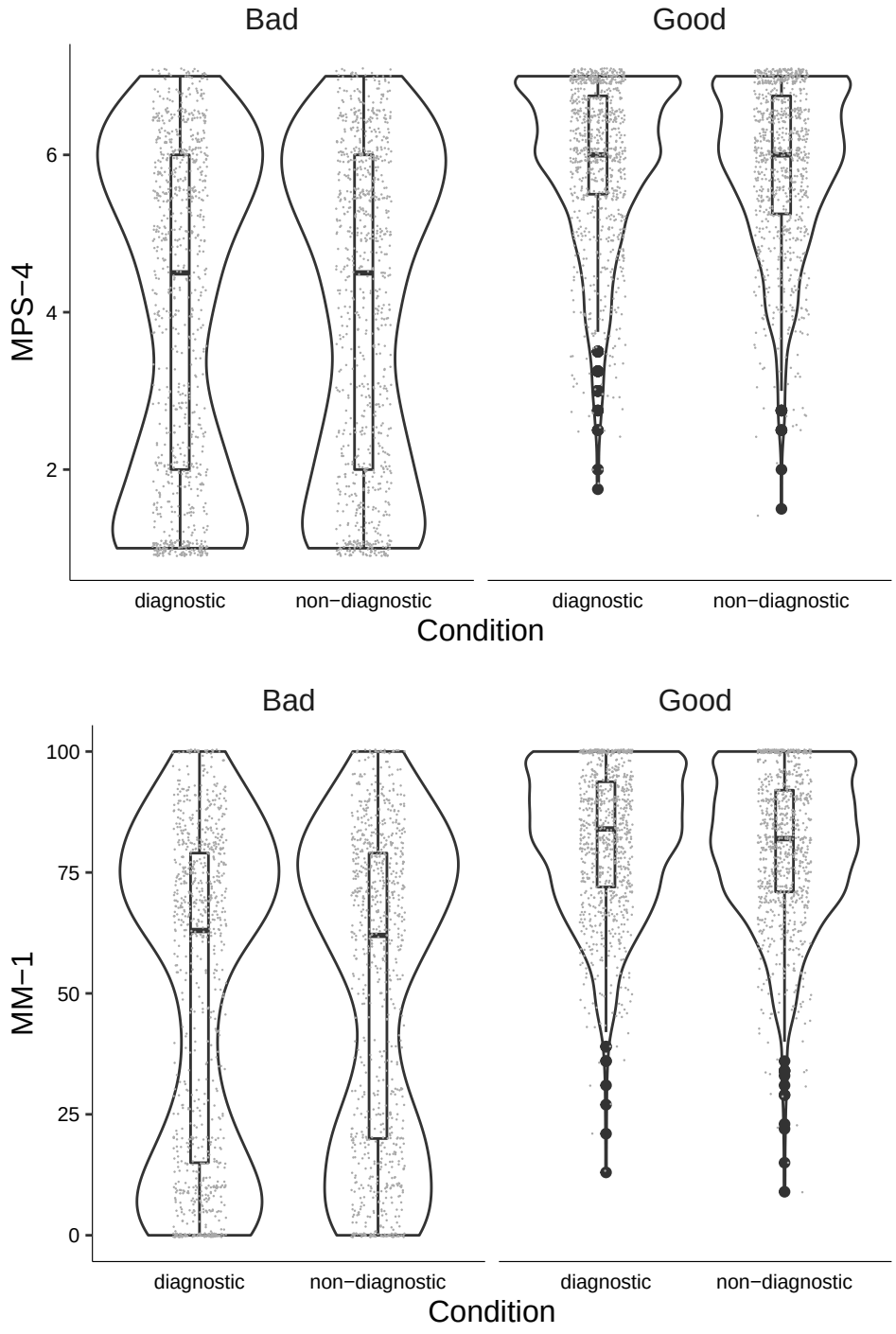

Figure 5

Study 3: Differences in moral perception depending on condition

The aim of Study 3 was to test if the moral dilution effect was moderated by valence of description. Based on the results of Studies 1 and 2 we hypothesized that a dilution effect would be observed for judgments of *bad* characters, but not for judgments of *good* characters. Interestingly, in Study 2 we found a dilution effect for *good* characters across both measures, however we only found a dilution effect for *bad* characters when using the single item measure of moral perception.

Study 4 - Good Characters

Following the divergent results for morally good characters between Studies 2 and 3, Study 4 was an attempted replication of Study 2.

Study 4: Method

Study 4: Participants and design

The design, materials, and procedure for Study 4 were the same as for Study 2, the only change from Study 2 was that all participants in Study 4 were recruited from MTurk. Study 4 was a within-subjects design. The independent variable was condition with two levels, diagnostic and non-diagnostic. We used the same two dependent variables as in previous studies, the four item moral perception scale (MPS-4, $\alpha = 0.81$), and the single item moral perception measure MM-1.

A total sample of 1118 (642 female, 445 male, 2 non-binary, 3 other; 1 prefer not to say, $M_{\text{age}} = 37.44$, min = 19, max = 84, $SD = 11.08$) started the survey. Participants were recruited from MTurk and paid $0.40 for their participation.

Participants who failed both manipulation checks were removed ($n = 262$), leaving a total sample of 856 participants (507 female, 347 male, 0 other, 0 prefer not to say; $M_{\text{age}} = 37.12$, min = 19, max = 84, $SD = 11.04$).

Study 4: Procedure and materials

All materials and procedures were the same as in Study 2.

Study 4: Results

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam*, $M_{\text{MPS-4}} = 5.95$, $SD_{\text{MPS-4}} = 0.93$, *Francis*, $M_{\text{MPS-4}} = 5.89$, $SD_{\text{MPS-4}} = 0.91$, *Alex*, $M_{\text{MPS-4}} = 5.94$, $SD_{\text{MPS-4}} = 0.96$, *Robin*, $M_{\text{MPS-4}} = 5.95$, $SD_{\text{MPS-4}} = 0.94$. There was significant variation depending on the description, $F(3,2527) = 3.30$, $p = .020$, partial $\eta^2 = 0.00$. Pairwise comparisons did not reveal any significant differences between individual descriptions (all ps $> .05$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (diagnostic/moral), $M_{\text{MM-1}} = 81.34$, $SD_{\text{MM-1}} = 14.14$; *Francis* (diagnostic/moral), $M_{\text{MM-1}} = 80.65$, $SD_{\text{MM-1}} = 14.16$; *Alex* (diagnostic/moral), $M_{\text{MM-1}} = 81.15$, $SD_{\text{MM-1}} = 14.42$; *Robin* (diagnostic/moral), $M_{\text{MM-1}} = 81.63$, $SD_{\text{MM-1}} = 14.15$. There was significant variation depending on the description, $F(3,2518) = 2.89$, $p = .035$, partial $\eta^2 = 0.000623$. Pairwise comparisons did not reveal any significant differences between individual descriptions (all ps $> .05$).

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants, and scenario was also included in the model. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 17.86$, $p = .022$. Condition did not influence responses to the MPS-4, $F(1, 866.60) = 2.80$, $p = .095$; and was not a significant predictor in the model when controlling for scenario, $b = 0.01$, $t(867) = 1.67$, $p = .095$, see Figure 6.

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data

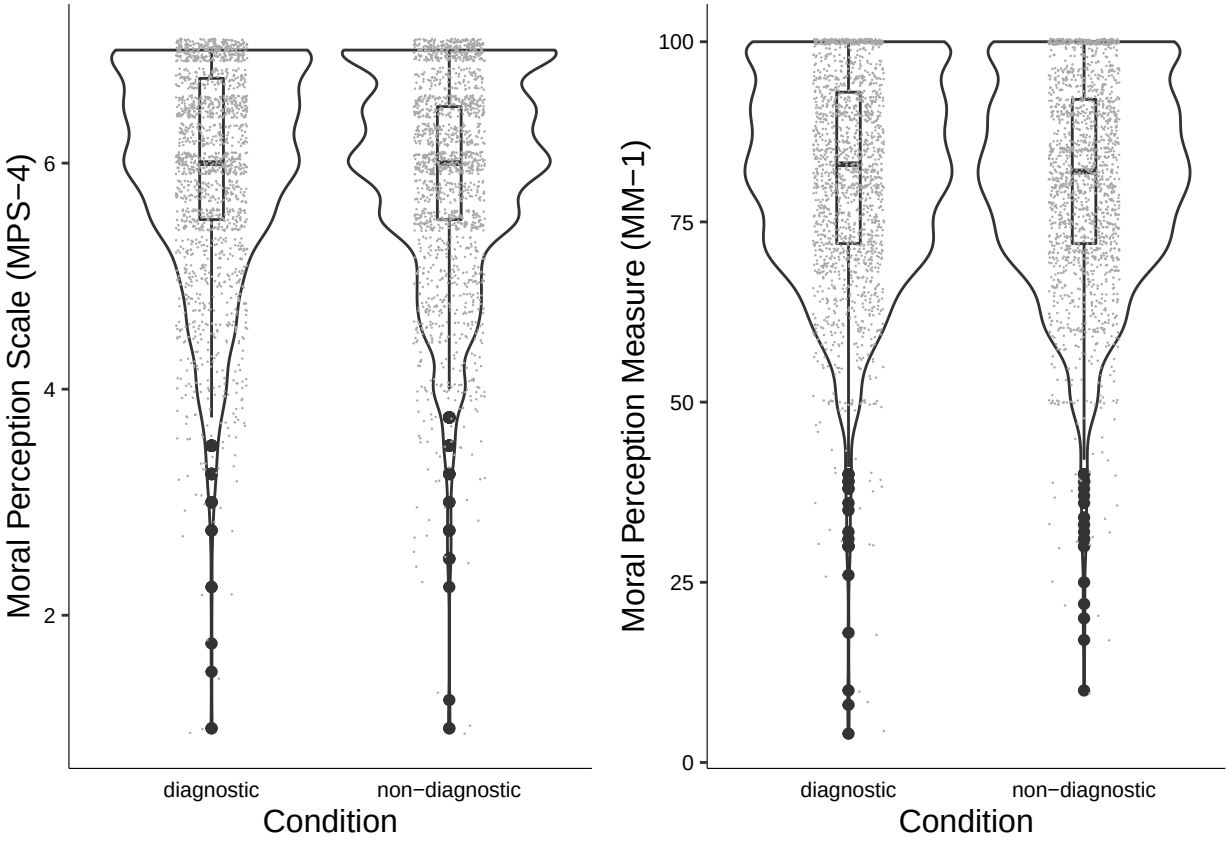

Figure 6

Study 2: Differences in moral perception depending on condition

than the baseline model, $\chi^2(8) = 40.10$, $p < .001$. Condition significantly influenced MM-1 responses $F(1, 864) = 4.79$, $p = .029$, and was a significant predictor in the model $b = 0.29$, $t(864) = 2.19$, $p = .029$, see Figure 6.

In the supplementary analyses we report the effect of condition on moral perception for each description individually.

Study 5 - Good and Bad Characters

Studies 1-4 showed some evidence for the presence of the moral dilution effect. The effect appears to vary depending on specific factors. One factor we identified is the presence, and valence, of other descriptions. The dilution effect is observed for *bad* characters when participants are presented with only *bad* characters (Study 1). However

this effect for *bad* characters is less reliably observed (depending on the measure) when participants are additionally presented with descriptions of *good* characters (Study 3). The reverse appears to be the case for *good* characters, such that when *good* characters are presented along with descriptions of *bad* characters, a moral dilution effect is observed (Study 3). However, when descriptions of *bad* characters are not present the moral dilution effect is not observed (Study 2), or less reliably observed (depending on the measure, Study 4). The aim of Study 5 was to test for the moral dilution effect in both good and bad characters, while attempting to eliminate the confounding influence of the presence of other descriptions by adopting a between-subjects design.

Study 5: Method

Study 5: Participants and design

Study 5 was a 2 × 2 between-subjects factorial design. As in Study 3, the first independent variable was condition with two levels, diagnostic and non-diagnostic. The second independent variable was valence of character description, with two levels morally good and morally bad. We used the same two dependent variables as in previous studies (MPS-4, $\alpha = 0.97$, and MM-1).

A total sample of 2126 (1107 female, 1004 male, 5 non-binary, 2 other; 8 prefer not to say, $M_{\text{age}} = 38.79$, min = 2, max = 1995, $SD = 45.07$) started the survey. Participants were recruited from MTurk and paid $0.10 for their participation.

Participants who failed both manipulation checks were removed ($n = 376$), leaving a total sample of 1750 participants (879 female, 858 male, 1 other, 1 prefer not to say; $M_{\text{age}} = 37.82$, min = 2, max = 454, $SD = 15.78$).

Study 5: Procedure and materials

The materials for Study 5 were the same as those used in Study 3. Participants were randomly presented with a single character description: *Sam*, *Robin* (*good* characters),

Francis and *Alex* (*bad* characters), and were randomly assigned to the diagnostic condition (containing diagnostic information only), or the non-diagnostic condition (where the character description additionally included non-diagnostic information). Study 5 was not pre-registered however our predictions were the same as those for Study 3.

Study 5: Results

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (good), $M_{\text{MPS-4}} = 6.13$, $SD_{\text{MPS-4}} = 0.87$, *Francis* (bad), $M_{\text{MPS-4}} = 3.72$, $SD_{\text{MPS-4}} = 2.16$, *Alex* (bad), $M_{\text{MPS-4}} = 3.73$, $SD_{\text{MPS-4}} = 2.08$, *Robin* (good), $M_{\text{MPS-4}} = 6.19$, $SD_{\text{MPS-4}} = 0.85$. There was significant variation depending on the description, $F(3,1746) = 333.96$, $p < .001$, partial $\eta^2 = 0.36$. Both the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than both the *bad* characters (*Alex* and *Francis*; all *p*s $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .938$) or between *Alex* and *Francis* (*bad*; ($p > .999$)).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (good), $M_{\text{MM-1}} = 84.52$, $SD_{\text{MM-1}} = 15.49$; *Francis* (bad), $M_{\text{MM-1}} = 44.51$, $SD_{\text{MM-1}} = 35.08$; *Alex* (bad), $M_{\text{MM-1}} = 45.85$, $SD_{\text{MM-1}} = 34.36$; *Robin* (good), $M_{\text{MM-1}} = 85.15$, $SD_{\text{MM-1}} = 14.61$. There was significant variation depending on the description, $F(3,1746) = 324.34$, $p < .001$, partial $\eta^2 = 0.36$. Both the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than both the *bad* characters (*Alex* and *Francis*; all *p*s $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .985$) or between *Alex* and *Francis* (*bad*; ($p = .882$)).

Testing for the Interaction

We conducted a 2 × 2 between subjects ANOVA to test for an interaction between valence and condition in predicting MPS-4. As expected, on its own, condition did not influence responses to the MPS-4, $F(1, 1746) = 0.01$, $p = .907$; valence significantly predicted responses, $F(1, 1746) = 1{,}004.46$, $p < .001$; and there was a significant condition

× valence interaction, $F(1, 1746) = 5.45$, $p = .020$.

We conducted a 2 × 2 between subjects ANOVA to test for an interaction between valence and condition in predicting responses to MM-1. As expected, on its own, condition did not influence responses to MM-1, $F(1, 1746) = 0.21$, $p = .650$; valence significantly predicted responses, $F(1, 1746) = 977.37$, $p < .001$; and there was a significant condition × valence interaction, $F(1, 1746) = 9.63$, $p = .002$.

To interpret the interaction effect, we conducted separate analyses for the Good and Bad descriptions.

Differences in the *Bad* Descriptions

For the *bad* characters, there was no significant difference in responses to MPS-4 between the diagnostic condition ($M = 3.63$, $SD = 2.21$) and the non-diagnostic condition ($M = 3.81$, $SD = 2.03$) depending on condition, $t(834.36) = -1.23$, $p = .221$, $d = 0.08$.

For the *bad* characters, there was no significant difference in responses to MPS-4 between the diagnostic condition ($M = 43.35$, $SD = 35.56$) and the non-diagnostic condition ($M = 46.86$, $SD = 33.85$) depending on condition, $t(843.09) = -1.48$, $p = .140$, $d = 0.10$.

Differences in the *Good* Descriptions

For the *good* characters, there was no significant difference in responses to MPS-4 between the diagnostic condition ($M = 6.25$, $SD = 0.84$) and the non-diagnostic condition ($M = 6.07$, $SD = 0.87$) depending on condition, $t(886.55) = 3.16$, $p = .002$, $d = 0.21$.

For the *good* characters, there was no significant difference in responses to MPS-4 between the diagnostic condition ($M = 87.11$, $SD = 13.75$) and the non-diagnostic condition ($M = 82.74$, $SD = 15.89$) depending on condition, $t(883.93) = 4.40$, $p < .001$, $d = 0.29$.

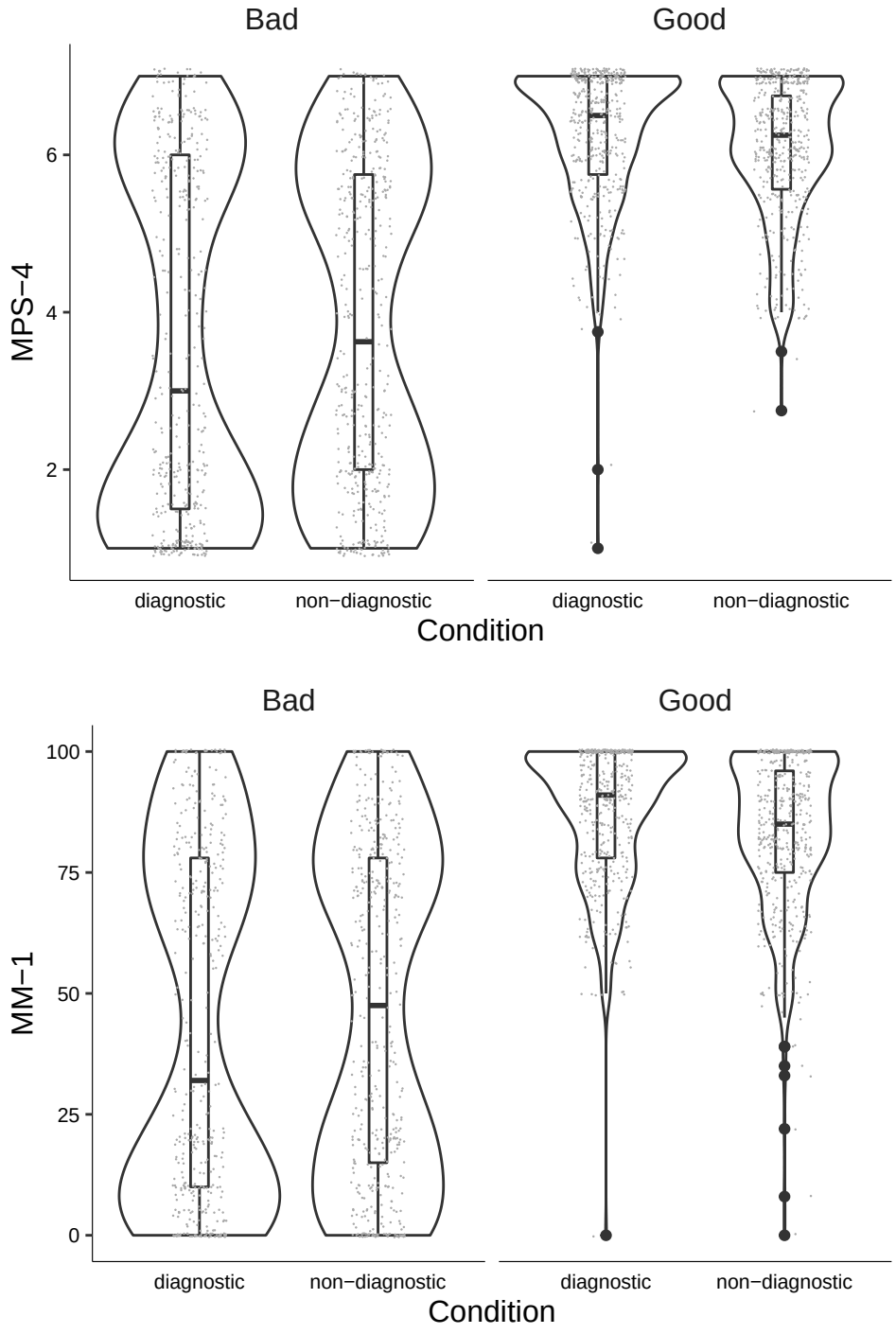

Figure 7

Study 3: Differences in moral perception depending on condition

In the supplementary analyses we report the effect of condition on moral perception for each description individually.

Discussion

Accessibility Statement

All data and analysis code are publicly available on this project's OSF page at https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67.

References

Grizzard, M., Fitzgerald, K., Francemone, C. J., Ahn, C., Huang, J., Walton, J., ... Eden, A. (2020). Validating the extended character morality questionnaire. *Media Psychology*, *23*(1), 107–130. https://doi.org/10.1080/15213269.2019.1572523

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2022). Moral Judgment as Categorization (MJAC). *Perspectives on Psychological Science*, *17*(1), 131–152. https://doi.org/10.1177/1745691621990636

Schein, C., & Gray, K. J. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Walker, A. C., Turpin, M. H., Fugelsang, J. A., & Białek, M. (2021). Better the two devils you know, than the one you don't: Predictability influences moral judgments of immoral actors. *Journal of Experimental Social Psychology*, *97*, 104220. https://doi.org/10.1016/j.jesp.2021.104220