Does Conceptual Abstraction Moderate Whether Past Moral Deeds Motivate Consistency or Compensatory Behavior? A Registered Replication and Extension of Conway and Peetz (2012)

Jareef Martuza[1], Olivia Kim[1]

[1]Department of Strategy and Management, Norwegian School of Economics, Helleveien 30, Bergen, Norway

Final version of the authors' manuscript, subject to final editing. Suggested citation:

Martuza, J., & Kim, O. (2024). Does Conceptual Abstraction Moderate Whether Past Moral Deeds Motivate Consistency or Compensatory Behavior? A Registered Replication and Extension of Conway and Peetz (2012). *Personality and Social Psychology Bulletin*, 01461672241238420. https://doi.org/10.1177/01461672241238420

Author notes

Correspondence concerning this article should be sent to Jareef Martuza, Department of Strategy and Management, Norwegian School of Economics, Helleveien 30, Bergen 5045. Email: jareef.martuza@nhh.no. Phone: +47 9251 3344

We (the authors) have no conflicts of interest to disclose. This manuscript has not been published elsewhere and is not under consideration by another journal. Our studies comply with all relevant ethical regulations regarding human research participants, including the guidelines from the Helsinki Declaration. As Norwegian laws and regulations do not require review by an institutional review board for anonymous, non-medical, low-risk research with human participants, we did not submit the project to such a review. All codes, data, and study materials from the two planned experiments are publicly available on the Open Science Framework (OSF) repository here.

We thank the editor Michael D. Robinson and three reviewers for detailed feedback during both Stage 1 and Stage 2. The comments and suggestions helped us a lot to improve the manuscript both before and after we conducted the study. We also thank Paul Conway for sharing materials from the original study which enabled us to conduct a close replication.

We received funding for the experiment from the Distribusjonsøkonomiske forskningsprosjekter (Distribution economics research project) thanks to Professor Sven Arne Haugland. The funding did not influence the study or our interpretations of the results. We also thank Professor Helge Thorbjørnsen for his feedback during the early stages of the project, and participants of the 2023 Johan Arndt Conference.

Abstract

A long-standing debate in psychology concerns whether doing something good or bad leads to more of the same or the opposite. Conway and Peetz (2012) proposed that *conceptual abstraction* moderates if past moral deeds lead to consistent or compensatory behavior. Although cited 384 times across disciplines, we did not find any direct replications. It was also unclear how increases or decreases from one's baseline prosociality might underlie the effect. A large-scale experiment (N=5091) in the registered report format tested Conway and Peetz's original hypothesis. The hypothesized interaction was *not* replicated: conceptual abstraction did not moderate the effect of recalling moral vs. immoral behavior on prosocial intentions. Our results show that recalling moral behavior led to higher prosocial intentions than recalling either immoral or neutral behavior, irrespective of recalling from the recent or distant past. Thus, the current research found no evidence for compensatory moral behavior, only for positive moral consistency.

Keywords: conceptual abstraction; moral consistency; moral licensing; moral cleansing.

Does Conceptual Abstraction Moderate Whether Past Moral Deeds Motivate Consistency or Compensatory Behavior? A Registered Replication and Extension of Conway and Peetz (2012)

Does doing something good or bad lead to more of the same or the opposite? Behavioral consistency theories propose that individuals tend to maintain a steady course of action (Festinger, 1954; Heider, 1946). For example, individuals who heavily invested in failing projects doubled down and invested more (Arkes & Blumer, 1985), people who had previously agreed to a small request were more likely to agree to a larger one later (Freedman & Fraser, 1966), and shoppers who brought their own bags to the store purchased more organic products (Karmarkar & Bollinger, 2015). Conversely, the *moral licensing* literature proposes that engaging in ethical behavior can give individuals a perceived license to act less ethically afterward (Effron et al., 2009; Effron & Conway, 2015; Monin & Miller, 2001; Mullen & Monin, 2016). For example, when individuals expressed disagreement with racist statements (Monin & Miller, 2001) or endorsed Barack Obama, they subsequently displayed a preference for hiring a white person (Effron et al., 2009), individuals recalling past moral actions showed higher future prosocial intentions (Jordan et al., 2011), and individuals who purchased environmentally friendly products subsequently acted less altruistically and were more likely to cheat or steal (Mazar & Zhong, 2010).

To reconcile the debate between sequential moral consistency and compensatory behavior, several studies have examined potential moderating factors determining *when* each is more likely to occur. Research shows compensatory (consistency) moral behavior is more likely when people adopt a concrete (abstract) mental construal (Brown et al., 2011; Conway & Peetz, 2012), are oriented toward an outcome-based (rule-based) ethical mindset (Cornelissen et al., 2013), have a goal progress (commitment) mindset (Susewind & Hoelzl,

2014), (do not) feel social recognition of the initial good behavior (Susewind & Walkowitz, 2020), focus on the past (future) organizational citizenship behavior (Griep et al., 2021), or have prevention (promotion) regulatory focus (Lalot et al., 2022). This stream of research suggests that the moral consistency vs. compensatory behavior debate may not be about *which* is more likely but *when* is each more likely.

In that vein, Conway and Peetz (2012) hypothesized that *conceptual abstraction* moderates whether past moral or immoral behavior leads to consistency or compensatory behavior. Specifically, Conway and Peetz's (2012) seminal Study 1 (hereafter C&P) found that recalling moral/immoral behavior in the distant past leads to consistency, whereas recalling moral/immoral behavior in the recent past leads to compensatory behavior. C&P informed later studies across several areas: including prosocial time giving (Reed et al., 2016), energy conservation (Tiefenbeck et al., 2013), workplace incivility (Rosen et al., 2016), CSR and employee misconduct (List & Momeni, 2021), and pro-environmental behavior (van der Werff et al., 2014). While C&P's Studies 2 and 3 also supported their main hypothesis, most other researchers who investigated moderators refer to the findings of C&P's Study 1- recalling distant vs. recent past behavior- when contextualizing their own results (Gholamzadehmir et al., 2019; Lalot et al., 2022; Susewind & Hoelzl, 2014).

Although cited 384 times across disciplines since publication in the *Personality and Social Psychology Bulletin*, to the best of our knowledge, no direct replications of C&P's hypothesized moderator of conceptual abstraction existed. This proposed moderator was based on construal level theory (Trope & Liberman, 2003, 2010). Considering the recent findings that highlight publication bias and the tendency to overestimate the effects of construal level theory, as well as calls for high-powered registered replications (Maier et al., 2022), it is crucial to also revisit the role of conceptual abstraction (based on temporal construal) as a *moderator*. Moreover, given the lack of control conditions in C&P, in which

participants would have recalled *neutral* behavior, it remains unclear how the differences after recalling moral or immoral behavior can be attributed to an increase and/or decrease from one's *baseline* prosocial intentions, a common shortcoming of existing literature on sequential moral behavior (Mullen & Monin, 2016).

Given the general need to independently revisit key psychological findings using current standards (Open Science Collaboration, 2015) and improve understanding of sequential moral behavior with recommended best practices (Blanken et al., 2015; Effron & Conway, 2015; Mullen & Monin, 2016), we conducted a registered replication and extension of Conway and Peetz's (2012) Study 1. We followed the general advice for using larger samples (Chambers & Tzavella, 2021) in replications, resulting in, to our knowledge, the largest single-lab study in the sequential moral behavior literature. Furthermore, we added two control conditions to delineate how the effects of recalling past moral and immoral behavior can increase and/or decrease prosocial intentions from baselines.

Replicating and Extending Conway and Peetz's (2012) Study 1

Conway and Peetz (2012) proposed that level of *conceptual abstraction* can moderate whether past moral behavior leads to moral consistency or compensatory behavior. They posited that moral behaviors from the distant past may be construed in more abstract terms, motivating individuals to act consistently with their salient moral identity (Blasi, 1980; Reed et al., 2007). Conversely, moral behaviors from the recent past, being more concretely construed, might lead individuals to feel they have made sufficient progress toward their moral goals, leading them to engage in compensatory behavior. Indeed, Conway and Peetz's (2012) findings revealed that participants who recalled moral (immoral) behavior from the *recent* past reported lower (higher) prosocial intentions, suggesting compensatory moral

behavior. Conversely, participants who recalled moral (immoral) behavior from the *distant* past reported higher (lower) prosocial intentions, suggesting consistent moral behavior.

Our main motivations for the current replication and extension were two-fold. First, despite the widespread influence of Conway and Peetz's Study 1, we found no independent direct replications in the literature. While related studies exist, such as Rotella and Barclay (2020) and Griep et al. (2021), their methods varied significantly from C&P. Rotella and Barclay (2020) conducted an online experiment that did not replicate either moral licensing or moral cleansing. This is not a direct replication of C&P because it had a three-condition design *without* any temporal specifications (recent vs. distant past) to the behaviors participants were asked to recall. Nonetheless, Rotella and Barclay's (2020) inclusion of a control condition (recalling neutral behavior) inspired our design and instructions. Additionally, although Griep et al. (2021) examined the moderating role of *temporal focus*, they measured individual differences in how people focus on their past, present, and future. So, their investigation cannot be directly compared to *experimentally* manipulated levels of conceptual abstraction as in C&P.

Further, recent null findings cast doubt on the generalizability and robustness of compensatory moral behavior altogether. For example, writing about positive traits (Sachdeva et al., 2009) did not lead to lower donations (Blanken et al., 2014), green consumption (Mazar & Zhong, 2010) did not increase subsequent cheating (Urban et al., 2019), and exposure to organic food (Eskine, 2013) did not reduce altruistic intentions (Moery & Calin-Jageman, 2016). However, another way of interpreting these null findings may be that compensatory (consistent) moral behavior may be more likely in *some* conditions than others. As C&P found, compensatory (consistent) moral behavior was more likely to manifest under concrete (abstract) conceptualization of the initial behavior. So, a large-scale replication can bolster our confidence in this key moderator.

Second, it is an open question if the compensatory moral behavior found by C&P was driven by moral licensing, cleansing, or both. In this paper, compensatory moral behavior refers to when an initial moral or immoral act leads to a subsequent behavior of the opposite moral valence (Mullen & Monin, 2016). Within this framework, moral licensing refers to a decrease in subsequent moral behavior following an initial moral behavior (Merritt et al., 2010), while moral cleansing refers to an increase in subsequent moral behavior following an initial immoral behavior (Zhong & Liljenquist, 2006).

In their study, C&P asked participants to recall and write about past moral or immoral behavior from the recent or distant past, and then measured prosocial intentions. C&P's main hypothesis was that recalling behaviors from the recent (distant) past should lead to compensatory (consistent) moral behavior. However, without any baseline (recalling neutral behavior) in the recent past conditions, it is unclear if prosocial intentions distinctly *decreased* after recalling moral behavior (licensing) or *increased* after past recalling immoral behavior (cleansing). As moral licensing and cleansing can rely on different psychological processes, neutral conditions can avoid *conflating* one with the other (Mullen & Monin, 2016), and so give us clearer insights into compensatory moral behaviors. Additionally, without any baseline in the distant past conditions, we cannot say if prosocial intentions distinctly increased after recalling moral behavior (positive consistency) or decreased after recalling immoral behavior (negative consistency). Therefore, extending C&P by adding two baseline conditions, that is also asking participants to recall neutral behavior from both the recent and distant past, can substantially improve our understanding of sequential moral behavior.

The Current Research

We conducted a large-scale (N = 5091) replication and extension of Conway and Peetz's (2012) original Study 1. We used the original materials and measures and closely followed C&P's procedures and materials for an independent direct replication. In line with Conway and Peetz's (2012) findings, we primarily hypothesized that recalling behaviors from the recent (distant) past should lead to compensatory (consistent) moral behavior.

All our hypotheses and analyses were pre-registered and can be accessed at: https://aspredicted.org/it26q.pdf. The study was programmed using Qualtrics (*Qualtrics*, 2022). All analyses were done using the statistical software jamovi (The jamovi project, 2022) and R (R Core Team, 2013). All codes, data, and study materials from the two planned experiments are publicly available on the Open Science Framework (OSF) repository here.

Method

Recruitment and Data Quality. We recruited participants from Prolific Academic, diverging from the original study which employed Amazon's Mechanical Turk (MTurk). This decision was based on Prolific having functions built specifically for academic research (Buhrmester et al., 2018), more stringent pre-screening for participants (Palan & Schitter, 2018), and greater naivety among participants (Peer et al., 2017) than MTurk. Moreover, Prolific has shown superior performance in terms of participant attention, comprehension, honesty, and reliability compared to MTurk (Peer et al., 2022). Given that both MTurk and Prolific are online crowdsourcing platforms with many shared characteristics (Goodman & Paolacci, 2017), choosing Prolific over Mturk should not bias results.

Participants. To minimize attrition, we stated in the recruitment post that the survey involved written communication, requesting participation only from those comfortable with typing one or two paragraphs (Zhou & Fishbach, 2016). Out of 5713 participants who began the study, 5165 participants completed the study (548 participants started but did not

complete it). The attrition was slightly higher in the moral (205, 10.7%) and immoral (213, 11.2%) than in the neutral (130, 6.8%) conditions. To test for differences in attrition, we conducted a binary logistic regression with Attrition (1 = did not complete, 0 = completed) as the outcome variables, and Event Valence, Event Distance, and the interaction between Event Valence and Event Distance as predictors. The results indicated that compared to the neutral conditions, attrition rates were significantly higher in both the moral (estimate = .032, p = .035; OR = 1.033, 95% CI [1.048, 1.080]) and immoral (estimate = .042, p = .006; OR = 1.042, 95% CI [1.002, 1.064]) conditions. Neither Event Distance nor any of the interaction terms were significant (estimates < .024, ps > .117).

Further, to rigorously assess the impact of missing data on our findings, we conducted sensitivity analyses by condition (Valence: Moral, Immoral, Neutral) using Manski bounds. Manski bounds allow us to estimate the range of possible outcomes for our dependent variables—Willingness to Volunteer (WTV) and Willingness to Help (WTH)—under two contrasting scenarios. The 'lower bound' represents a conservative 'worst case' scenario, assuming that all missing data would have resulted in the lowest possible outcomes. Conversely, the 'upper bound' reflects an optimistic 'best case' scenario, assuming missing data would have resulted in the highest possible outcomes. The calculated bounds for both WTV (Moral [4.20, 4.77], Immoral [3.94, 4.55], Neutral [4.06, 4.43]) and WTH (Moral [5.22, 5.81], Immoral [5.12, 5.76], Neutral [5.20, 5.59]) across the three valence conditions were within narrow ranges. This indicates that irrespective of the assumptions made about the nature of the missing data, the observed differences in attrition rates have minimal impact on our study’s conclusions.

We now turn to exclusions from the 5165 complete responses. As per our pre-registered exclusion criteria, we excluded 59 responses that were flagged by Qualtrics’ fraud detection measure (ReCaptchaScore <= .5), and an additional 15 responses where participants

failed two out of three attention checks. The final dataset for analyses comprised responses from 5091 participants (50.87% female; M_{age}= 39.3, SD = 14.0). For the direct replication analyses, a subset of N = 3339 participants was used (51.4% female; M_{age}= 39.2, SD = 14.1).

Due to a technical glitch in Qualtrics, there were unequal cell sizes with nearly three times as many participants in the "distant" conditions as in the "recent" conditions (as embedded in Qualtrics). Also, we mislabeled conditions such that neutral/distant and neutral/recent were reversed when programming Qualtrics, and so we relabeled those before data analyses. Although the stimuli in terms of descriptions shown to participants were unaffected, our coding error led to three times more participants in the neutral/recent than neutral/distant. Exact cell sizes in the final dataset are reported in Table 1.

Table 1. Achieved cell sizes per condition (exclusions in parentheses) in the final dataset.

Participants/ condition	Recent	Distant
Moral	447 (4)	1226 (23)
Immoral	455 (4)	1221 (19)
Neutral	1316 (15)	436 (9)

Unequal cell sizes may raise questions about whether homogeneity of variances assumptions hold. For the 2 (Event Valence: Moral vs. Immoral) X 2 (Event Distance: Recent X Distant) ANOVA for the replication analyses, the assumptions of equal variances were not violated for either of the dependent variables: Willingness-to-volunteer: $F(3, 3335) = 1.63$, $p = .180$, or Willingness-to-help: $F(3, 3335) = 2.50$, $p = .058$. For the 3 (Event Valence: Moral vs. Immoral) X 2 (Event Distance: Recent X Distant) ANOVA for the extension analyses, the assumptions of equal variances were violated for both of the dependent variables: willingness-to-volunteer: $F(5, 5085) = 2.40$, $p = .035$), and WTV: $F(3, 3335) = 2.30$, p

= .042). Nonetheless, Levene's test is quite sensitive in large samples because even small deviations from the homogeneity of variances assumption would be statistically significant, even though not meaningful in terms of the effect. Indeed, the ratios between the largest and smallest variation across cells for both dependent variables, WTV (ratio = 1.24) and WTH (ratio =1.44), were under 1.5, and therefore, unproblematic (Please see Table 4 to compare standard deviations). So, a statistically significant result in a large sample like ours may not imply a practically significant difference in variances. We also conducted robust linear regressions (non-preregistered) on both outcome variables. The results did not change in either the direction of effects or their significance levels. Please see Table S5C in section 6.4 in the Online Supplement for exact estimates.

Detectable effect sizes. Sensitivity analyses using G*Power (Faul et al., 2007) showed that the original C&P study, with 90% power, could detect an effect size of f = .326 or greater at $p < .05$ for the Event Valence X Event Distance interaction term in a 2 X 2 ANOVA, and an effect size of d = .651 or greater at $p < .05$ (two-tailed) for the planned contrast in a two-tailed t-test. At 90% power and $p < .05$ (two-tailed), our conducted replication can detect effect sizes more than five times smaller (f = .056 and d = .112 respectively) than the original C&P study can detect.

Considering the recent null results and possible publication bias inflating the moral licensing effect observed in meta-analyses (d = .31: Blanken et al., 2014; d = .32: Simbrunner & Schlegelmilch, 2017), the classical moral licensing effect may have been overestimated (Kuper & Bott, 2019). A "many-labs" study suggests a more modest effect size of d = .14 (Ebersole et al., 2016). Our sample size (N = 3339), which is more than 30 times that of C&P's original study (N = 101), allows us to detect effect sizes as small as d = .112 for C&P's specific a priori contrast (H1a), as well as for contrasts with neutral baselines (H2a

and H2c). Please see Table 2 for a summary of minimal detectable effect sizes. Please see Figures S1.1 to S8 in the Online Supplement for power calculations.

Table 2. Summary of hypotheses and minimal detectable effect sizes (MDES) at 90% power for two-tailed tests at $p < .05$.

Effect	#	Hypothesis	MDES
Original Interaction		Event Valence X Event Distance Interaction (2 X 2 ANOVA)	f = .056 (Figure S1)
Contrast	H1	Participants in the moral/distant and immoral/recent conditions will exhibit higher prosocial intentions than participants in the moral/recent and immoral/distant conditions.	d = .112 (Figure S2)
	H2a	Participants in the moral/distant and immoral/recent conditions will exhibit higher prosocial intentions than participants in the neutral/recent and neutral/distant conditions.	d = .111 (Figure S3)
	H2b	Participants in the moral/recent and immoral/distant conditions will exhibit lower[1] prosocial intentions than participants	d = .111 (Figure S4)

[1] In our original pre-registration on Aspredicted, we made a typo, whee we wrote "higher" instead of "lower". H2a and H2b proposes to test if prosocial intentions increase and/or decrease from the baseline, with H2a and H2b mirroring each other such that H2a hypothesizes higher (increase) and H2b hypothesizes lower (decrease).

		in the neutral/recent and neutral/distant conditions.	
Licensing	H3a	Participants in the moral/recent condition will exhibit lower prosocial intentions than participants in the neutral/recent condition.	d = .200 (Figure S5)
Cleansing	H3b	Participants in the immoral/recent condition will exhibit higher prosocial intentions than participants in the neutral/recent condition.	d = .196 (Figure S6)
Positive consistency	H3c	Participants in the moral/distant condition will exhibit higher prosocial intentions than participants in the neutral/distant condition.	d = .201 (Figure S7
Negative consistency	H3d	Participants in the immoral/distant condition will exhibit lower prosocial intentions than participants in the neutral/distant condition.	d = .201 (Figure S8)

Procedure. In a 3 (Event Valence: Moral vs. Immoral vs. Neutral) X 2 (Event Distance: Recent vs. Distant) between-participants design, participants were randomly assigned to one of six conditions. They were instructed to recall and describe a moral, immoral, or neutral behavior from either the recent or distant past. Subsequently, participants responded to outcome measures of willingness-to-help (WTH) and willingness-to-volunteer (WTV), consistent with the original study.

Behavior Recall manipulation. We manipulated Event Valence by instructing participants to recall a moral, immoral, or neutral event. Participants in the moral condition were asked to recall a time when they acted in such a way that they felt righteous or honorable. Participants in the immoral condition were asked to recall a time when they acted

in such a way that they felt guilty or ashamed. Participants in the neutral condition were asked to recall a time when they went shopping by themselves.

Event Distance was manipulated by asking participants to describe an event that occurred either in the past week (recent conditions) or over one year ago (distant conditions). For example, the prompt in the moral/recent condition was "Please recall a time within the past week when you acted in such a way that you felt righteous or honorable. Perhaps you were loyal to a friend, were generous when you could have been selfish, were kind to someone for no particular reason, or caring toward someone who needed you." To elicit elaborate responses and strengthen the manipulations, participants were also told, "Please provide as much detail as you can, and write at least a paragraph with complete sentences." on the same page as the response box in all conditions. Please see Table S2 in the Online Supplement for the recall prompts across conditions in exact words.

Prosocial Intention Measures. As dependent variables, we measured willingness to volunteer and help others right after participants completed the behavioral recall task. First, participants completed a 5-item willingness-to-volunteer measure (DeVoe & Pfeffer, 2007; $\alpha = .88$) on a 7-point scale anchored at 1 (completely disagree) to 7 (completely agree) comprising a randomized order of items such as "Volunteering is a worthwhile use of my time even if I do not get paid".

Then, participants read four vignettes in a randomized order, each depicting other people needing small everyday help (e.g., paying a few extra cents for someone else's restaurant bill). Participants indicated their willingness to help on a 7-point scale anchored at 1 (very unlikely) to 7 (very likely) for each scenario. These responses were aggregated into an index of willingness-to-help ($\alpha = .64$).

Manipulation checks. After the main outcome measures, participants rated event positivity, "The event I wrote about made me feel good about myself", and perceived

temporal distance, “The event I wrote about happened a long time ago”) on a 7-point scale anchored at 1 (completely agree) to 7 (completely disagree). At the very end of the study, participants were also asked to respond to an additional direct manipulation check of whether the event they recalled happened (a) within the last week, (b) more than a year ago, or (c) in between a week and a year.

PANAS. Participants completed the 20-item positive and negative affect schedule (PANAS; Watson et al., 1988) after the main measures. Participants indicated the extent to which they currently felt ten positive (e.g., interested, excited) and ten negative (e.g., depressed, upset) emotions on a 5-point scale anchored at 1 (not at all) to 5 (very strongly). Respective items were averaged into a positive emotion subscale ($\alpha = .93$) and a negative emotion subscale ($\alpha = .92$). Finally, participants responded to individual-level questions including indicating their age, gender, education, and income. Please see Table S1 in the Online Supplement for a list of all measures and items of the survey.

Differences between C&P and the current research

In Table 3, we list the known differences between the original and the replication study concerning design, materials, measures, and psychometric properties. Two notable differences may stand out. First, event positivity ratings in the moral conditions are much higher in our study. Second, perceived event distance ratings in the distant conditions are also higher in our replication. However, we argue that this perhaps suggests greater strengths in our manipulations and should not matter for replicating C&P’s interaction effect.

Table 3. Comparisons of design, sample characteristics, and psychometric properties of key variables between Conway and Peetz (2012) and the current study.

	C&P	Current Study

Design	4 conditions 2 (Event Valence: Moral vs. Immoral) X 2 (Event Distance: Recent vs. Distant) between-participants.	6 conditions 3 (Event Valence: Moral vs. Immoral vs. Neutral) X 2 (Event Distance: Recent vs. Distant) between-participants.
Materials: Recall prompt format (Illustration with Moral/Recent condition)	Please recall a time within the last week when you acted in such a way that you felt righteous or honorable. Perhaps you were loyal to a friend, were generous when you could have been selfish, were kind to someone for no particular reason, or caring toward someone who needed you.	Please recall a time within the last week when you acted in such a way that you felt righteous or honorable. Perhaps you were loyal to a friend, were generous when you could have been selfish, were kind to someone for no particular reason, or caring toward someone who needed you. **Please provide as much detail as you can, and write at least a paragraph with complete sentences.**
Outcome measures	willingness-to-volunteer, willingness-to-help	willingness-to-volunteer, willingness-to-help
Manipulation check measures	perceived event positivity, perceived event distance	perceived event positivity, perceived event distance, **specific time of recalled event**.
Sample Characteristics	Mturk participants N = 101	Prolific participants N = 5091

	Female: 68% Age: M = 43.91, SD = 14.19	Female: 50.9% Age: M = 39.3, SD = 14.0
Cronbach's alpha (a) of measured constructs	Willingness-to-volunteer: α = .92 Willingness-to-help: α = .56 Negative emotion: α = .88 Positive emotion: α = .88	Willingness-to-volunteer: α = .89 Willingness-to-help: α = .64 Negative emotion: α = .92 Positive emotion: α = .93
Relevant means (M) and standard deviations (SD) of manipulation checks	Event positivity: Moral: M = 4.25, SD = .84 Immoral: M = 1.70, SD = .69. Event distance: Recent: M = 1.40, SD = .54 Distant: M = 3.02, SD = 1.09	Event positivity: Moral: M = 5.81, SD = 1.30 Immoral: M = 1.81, SD = 1.64 Neutral: M = 4.0, SD = 1.64. Event distance Recent: M = 1.54, SD = 1.17 Distant: M = 4.07, SD = 1.86

Results

First, we compare descriptive statistics between the original and current study for the two dependent variables listed in Table 4 while also detailing results from manipulation checks. Then, we replicate C&P's exact analyses on a subset of the data (N = 3339) comprising the four original conditions of C&P. Following that, we report analyses based on the full dataset (N = 5091) with all six conditions. Although our current sample reported lower willingness-to-volunteer than C&P *on average*, this may merely be due to differences in possible sampling error between C&P (N = 101) and the current research (N = 5091).

Table 4. Comparing means (M) and standard deviations (SD) of dependent variables (D) between C&P and current study across conditions.

DV	Condition	C&P	Current Study
Willingness-to-volunteer	Moral/Recent	M = 5.42, SD = 1.37	M = 4.51, SD = 1.39
	Immoral/Recent	M = 5.97, SD = 1.30	M = 4.29, SD = 1.38
	Moral/Distant	M = 5.94, SD = 1.02	M = 4.63, SD = 1.32
	Immoral/Distant	M = 5.26, SD = 1.49	M = 4.37, SD = 1.39
	Neutral/Recent	NA	M = 4.32, SD = 1.42
	Neutral/Distant	NA	M = 4.26, SD = 1.47
Willingness-to-help	Moral/Recent	M = 5.81, SD = .70	M = 5.69, SD = 0.89
	Immoral/Recent	M = 6.00, SD = .73	M = 5.56, SD = 0.99
	Moral/Distant	M = 6.01, SD = .78	M = 5.68, SD = 1.01
	Immoral/Distant	M = 5.56, SD = .78	M = 5.63, SD = 1.02
	Neutral/Recent	NA	M = 5.48. SD = 1.07
	Neutral/Distant	NA	M = 5.53, SD = 1.07

Manipulation checks

With respect to the full dataset consisting of all six conditions, a 3 (Event Valence: Moral vs. Immoral vs. Neutral) X 2 (Event Distance: Recent vs. Distant) between-participants ANOVA on recalled event positivity revealed the expected effect of Event Valence on event positivity, F (2, 5085) = 2568.82, p < .001, $\eta^2_p = .502$. Event Distance did not significantly moderate the effect of valence, F (2, 5085) = .102, p = .903, $\eta^2_p < .001$. Participants instructed to recall and write about a moral event (M = 5.81, SD = 1.30) reported their event to be significantly more positive than those in the immoral conditions (M = 1.81, SD = 1.31), t (5086) = 71.6, $p_{Tukey} < 0.001$, d = 2.79, 95% CI [2.70, 2.88]. Participants

instructed to recall and write about a neutral event (M = 4.00, SD = 1.64) reported their event to be significantly *less* positive than participants in the moral conditions (M = 5.81, SD = 1.30), t (5085) = -33.3, $p_{Tukey} < .001$, d = -1.30, 95% CI [-1.38, 1.-22], and significantly *more* positive than participants in the immoral conditions (M = 1.81, SD = 1.31), t (5085) = 38.2, $p_{Tukey} < .001$, d = 1.49, 95% CI [1.41, 1.57].

A 2 (Event Distance: Recent vs. Distant) X 2 (Event Valence: Moral vs. Immoral) between-participants ANOVA on perceived distance of the recalled event revealed the expected effect of Event Distance on perceived distance, F (1, 5085) = 2472.46, p < .001, $\eta^2_p = .327$. Event Valence did not moderate the effect of Event Distance, F (2, 5085) = .249, p = .780, $\eta^2_p < .001$. Participants instructed to recall and write about a distant event (M = 3.17, SD = 1.84) perceived the event to be significantly more distant than participants in the recent conditions (M = 2.39, SD = 2.05), t (5085) = 49.72, $p_{Tukey} < .001$, d = 1.58, 95% CI [1.51, 1.65]. Please see Tables S4A and S4B in the Online Supplement for comparisons of test statistics of manipulation checks between the original and replication.

Our direct manipulation check measure for Event Distance showed that a significant majority of participants in the 'Recent' condition perceived the events as occurring 'within the last week' (90.7%) than otherwise (9.03%), Chi-squared statistic = 882.91, p < .001. Moreover, a significant proportion of participants in the 'Distant' condition perceived the events as occurring 'more than a year ago' (63.8%) than otherwise (36.2%), Chi-squared statistic = 203.37, p < .001. Including or excluding participants who answered incorrectly with respect to this check did not materially change results. Please see Tables S5A and S5B in the Online Supplement for an overview and comparisons of test statistics between datasets.

Replicating C&P's analyses

Here, we report the same analyses as in C&P to test if support for their original hypothesis (H1) can be replicated. For comparisons of the test statistics between C&P and our study, please see Tables S4C and S4D in the Online Supplement.

Willingness-to-volunteer (WTV). A 2 (Event Valence: Moral vs. Immoral) X 2 (Event Distance: Recent vs. Distant) between-participants ANOVA on willingness to volunteer (WTV) revealed a significant effect of Event Valence, F (1, 3335) = 20.332, p <.001, $\eta^2_p = .006$; and a non-significant effect of Event Distance, F (1, 3335) = 3.817, p = .051, $\eta^2_p = .001$. Crucially, the Event Valence X Event Distance interaction effect on WTV was **not** significant, F (1, 3335) = .176, p = .675, $\eta^2_p < .001$, rendering the planned comparison of H1 irrelevant.

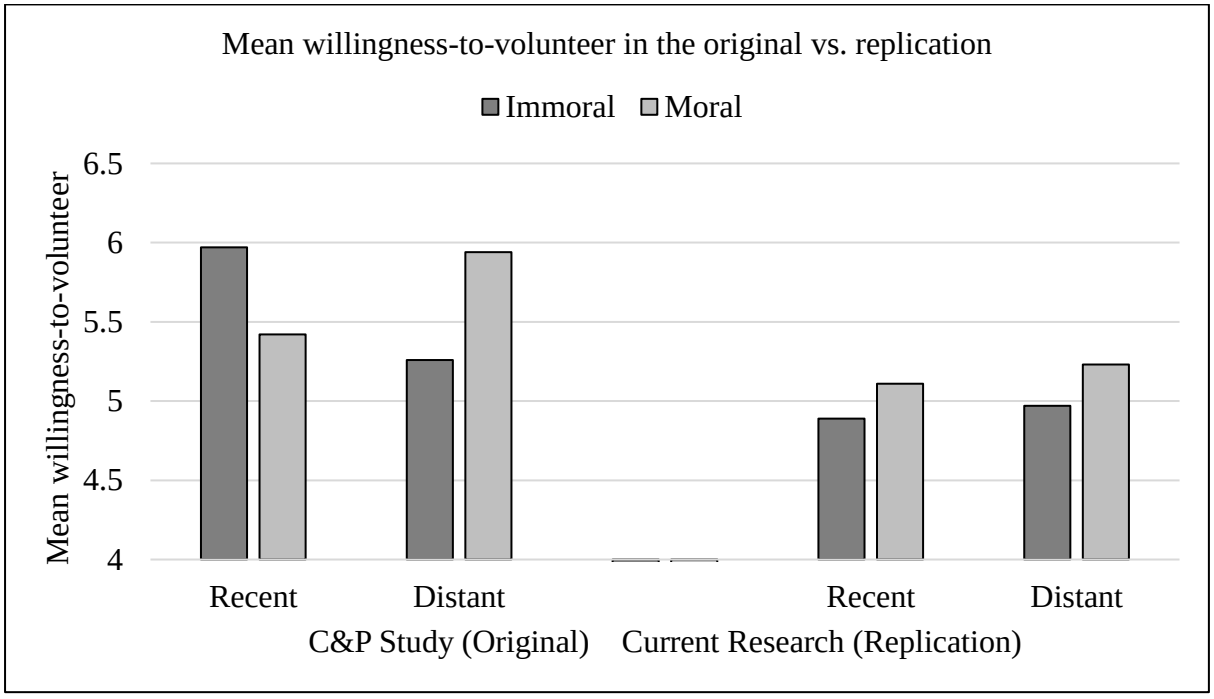


Figure 1A. Mean willingness-to-volunteer across conditions in the original versus replication. Responses were measured on 7-point scales anchored at 1 and 7.

Tukey's post hoc comparisons showed that WTV was higher in the moral (M = 4.60, SD = 1.34) than immoral (M = 4.35, SD = 1.39) conditions, t (3335) = 4.51, $p_{Tukey} < .001$, d

= .176, 95% CI [.099, .252]. A closer look also revealed that WTV was non-significantly higher in the moral/recent (M = 5.11, SD = 1.39) than immoral/recent (M = 4.89, SD = 1.38) condition, t (3335) = 2.39, p_{Tukey} = .079, d = .159, 95% CI [.029, .290]). Similarly, WTV was higher in the moral/distant (M = 5.23, SD = 1.32) than immoral/distant (M = 4.97, SD = 1.39) condition, t (3335) = 4.74, p_{Tukey} < .001, d = .192, 95% CI [.113, .272]).

So, contrary to C&P, our results suggest an overall presence of moral consistency in both recent and distant conditions, and the absence of compensatory moral behavior in the recent conditions, together suggesting that the original interaction hypothesis (H1) is not supported.

Willingness-to-help. A 2 (Event Valence: Moral vs. Immoral) X 2 (Event Distance: Recent vs. Distant) between-participants ANOVA on willingness to help (WTH) revealed a significant effect of Event Valence, F (1, 3335) = 5.643, p = .017, $\eta^2_p = .002$; and a non-significant effect of Event Distance, F (1, 3335) = 0.519, p = .47, $\eta^2_p < .001$. Crucially, the Event Valence X Event Distance interaction effect on WTH was **not** significant, F (1, 3335) = 1.052, p = .305, $\eta^2_p < .001$, rendering the planned comparisons with respect to H1 superfluous.

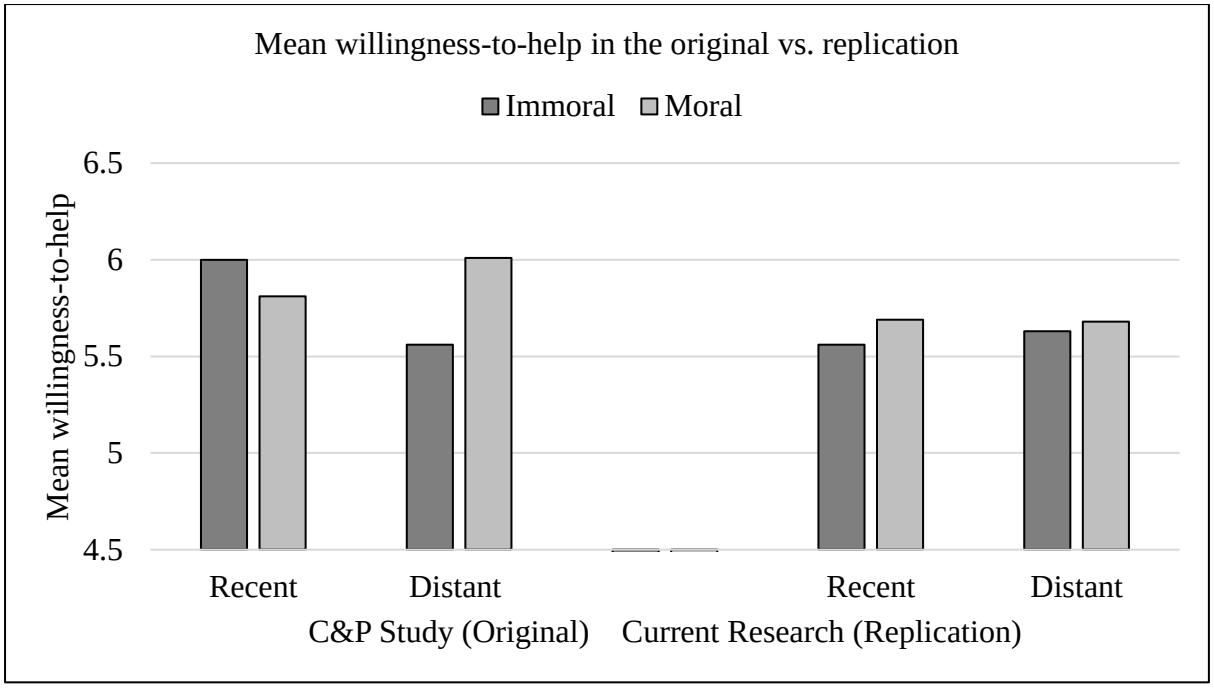


Figure 1B. Mean willingness-to-help across conditions in the original versus replication. Responses were measured on 7-point scales anchored at 1 and 7.

Tukey's post hoc comparisons showed that WTH was higher in the moral (M = 5.68, SD = 0.98) than immoral (M = 5.61, SD = 1.01) conditions, t (3335) = 2.38, p_{Tukey} = .017, d = .093, 95% CI [.016, .169]. A closer look also revealed that WTH was non-significantly higher in the moral/recent (M = 5.69, SD = .089) than immoral/recent (M = 5.56, SD = 0.99) condition, t (3335) = 1.99, p_{Tukey} = .190, d = .133, 95% CI [.002, .263]). However, WTH was non-significantly different in the moral/distant (5.68, SD = 1.01) and immoral/distant (M = 5.63, SD = 1.02) conditions, t (3335) = 1.31, p_{Tukey} = .560, d = .068, 95% CI [-.040, .176]).

So, contrary to C&P, our results suggest a minimal effect of moral consistency (d = .093) when the recent and distant conditions are combined, and the absence of compensatory moral behavior in the recent conditions, again suggesting that the original interaction hypothesis (H1) is not supported.

Analyses with the full dataset

We now turn to our analyses with the full dataset to test hypotheses H2a-b and H3a-d. For all non-significant planned comparisons, we conducted equivalence tests (Lakens, 2017) to examine if the null was conclusive or inconclusive. These were not pre-registered and should be seen as exploratory when interpreting null results. As opposed to how traditional hypothesis testing looks for significant differences, an equivalence test examines if the difference between two conditions is smaller than a pre-specified smallest effect size of interest (SESOI). Here, the null hypothesis posits that the true difference between the conditions is greater in magnitude (that is either above the upper bound or below the lower bound). The alternative hypothesis posits that the true differences lie within the bound, hence suggesting equivalence between the groups, and so statistical equivalence. For each TOST analysis, we reported the p-value against the bound (lower or upper) depending on the direction of the hypothesis. For a primer and tutorial, please see Lakens (2017) and Lakens et al. (2018).

We determined our smallest effect sizes of interest (SESOI) to d = .14, informed by Ebersole et al.'s (2016) multi-lab study finding the average effect of moral licensing effect to be d = -.14. Accordingly, we set our upper bounds as d = .14 when equivalence testing for moral cleansing and positive moral consistency, and lower bounds as d = -.14 when equivalence testing for moral licensing and negative moral consistency.

Willingness-to-volunteer. A 3 (Event Valence: Moral vs. Immoral vs. Neutral) X 2 (Event Distance: Recent vs. Distant) between-participants ANOVA on willingness-to-volunteer (WTV) revealed a significant effect of Event Valence, $F(2, 5085) = 15.54$, $p < .001$, $\eta^2_p = .006$; but a non-significant effect of Event Distance, $F(1, 5085) = 1.16$, $p = .281$, $\eta^2_p < .001$. The Event Valence X Event Distance interaction effect on WTV was **not**

significant, F (2, 5085) = 1.69, p = .184, $\eta^2_p < .001$, rendering planned comparisons for H2a-b superfluous.

Tukey's post hoc pairwise comparisons revealed that although WTV was lower in the neutral (M = 4.31, SD = 1.43) than in the moral (M = 4.60, SD = 1.34) conditions, t (5085) = -4.36, $p_{Tukey} < .001$, d = -.170, 95% CI [-.247, -.094], WTV in the neutral conditions was not significantly different than in the immoral (M = 4.35, SD = 1.39) conditions, t (5085) = -.73, $p_{Tukey} = .746$, d = -.028, 95% CI [-.105, .048].

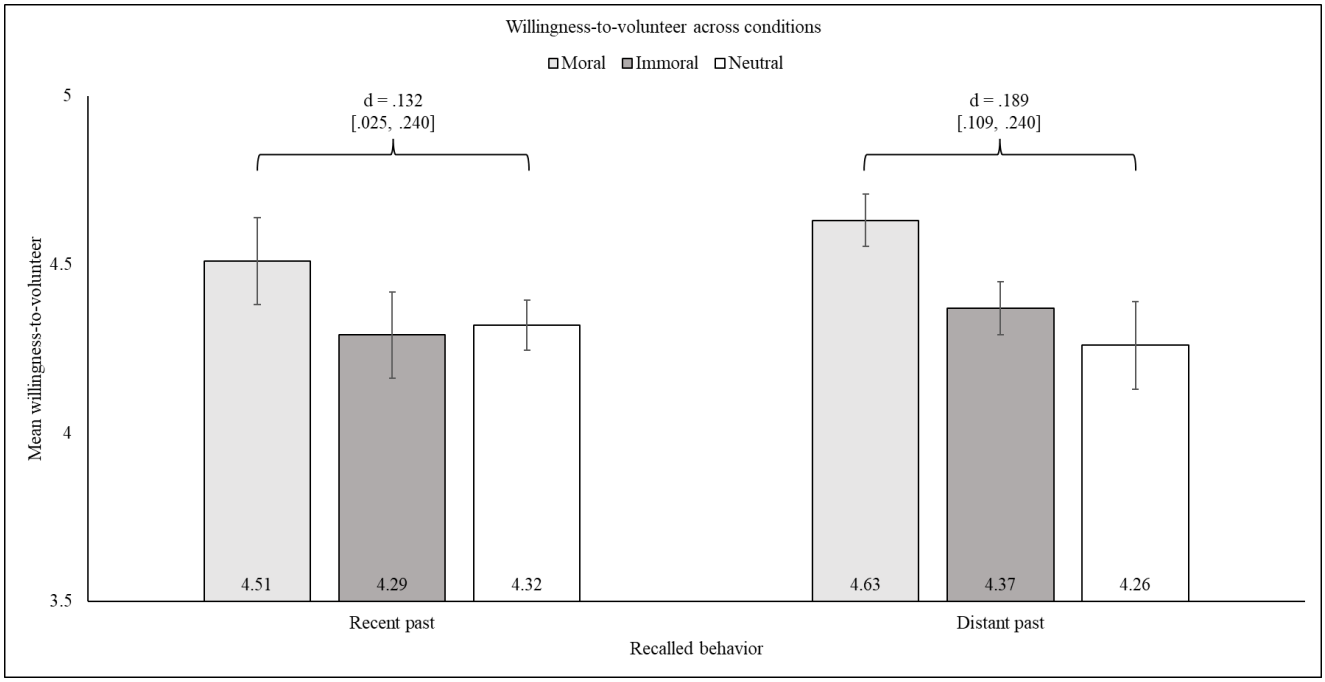


Figure 2A. Mean willingness-to-volunteer (WTV) across conditions. Error bars refer to 95% confidence intervals.

To test H3a (moral licensing) and H3b (moral cleansing), we turn to planned comparisons to test for the presence of *compensatory* moral behavior in the "recent" event conditions. WTV was non-significantly higher in the moral/recent (M = 4.51, SD = 1.39) than neutral/recent (M = 4.32, SD = 1.42) condition, t (5085) = -2.41, $p_{Tukey} = .150$, d

= .132, 95% CI [.025, .240]. The opposite direction of the significant effect suggests strong evidence *against* a moral licensing effect (H3a) in the recent conditions.

However, WTV was not significantly different in the immoral/recent (M = 4.29, SD = 1.38) than in the neutral/recent (M = 4.32, SD = 1.42) condition, t (5085) = -.44, p_{Tukey} = .998, d = .024, 95% CI [-.082, .131]. An equivalence test with an upper bound of d = .14 was significant, t (809) = 3.04, p = .001. So, the standardized mean difference between the immoral and neutral conditions lies between 0 and .14, a magnitude smaller than our minimum effect size of interest, and so suggesting conclusive evidence of a null effect of moral cleansing (H3b).

To test H3c (positive moral consistency) and H3d (negative moral consistency), we turn to planned comparisons to test for *consistent* moral behavior in the "distant" event conditions. WTV was significantly higher in the moral/distant (4.63, SD = 1.32) than neutral/distant (M = 4.26, SD = 1.47) condition, t (5085) = 4.84, p_{Tukey} < .001, d = .189, 95% CI [.109, .268], suggesting a positive moral consistency effect in the distant conditions.

However, WTV was not significantly different in the immoral/distant (M = 4.37, SD = 1.39) than in neutral/distant (M = 4.26, SD = 1.47) condition, t (5085) = 1.45, p_{Tukey} = .695, d = .081, 95% CI [.028, -.191]. An equivalence test with a lower bound of d = -.14 was significant, t (733) = -3.86, p < .001. So, the standardized mean difference between the immoral and neutral conditions lies between 0 and .14, suggesting conclusive evidence of a null effect of negative moral consistency (H3d).

Willingness-to-help. A 3 (Event Valence: Moral vs. Immoral vs. Neutral) X 2 (Event Distance: Recent vs. Distant) between-participants ANOVA on willingness-to-help (WTH) revealed a significant effect of Event Valence, F (2, 5085) = 9.53, p < .001, $\eta^2_p = .004$, and non-significant effect of Event Distance, F (1, 5085) = 1.20, p = .273, $\eta^2_p < .001$. The Event

Valence X Event Distance interaction effect on WTH was **not** significant, F (2, 5085) = .555, p = .574, $\eta^2_p < .001$, making the planned comparisons of H2a-b redundant.

Tukey's post hoc pairwise comparison tests revealed that WTH was significantly lower in the neutral (M = 5.50, SD = 1.07) than moral (M = 5.68, SD = .98) conditions, t (5085) = -4.36, p < .001, d = -.17, 95% CI [-.247, -.094], and not significantly different than immoral (M = 5.61, SD = 1.01) condition, t (5085) = -2.05, p_{Tukey} = .10, d = -.080, 95% CI [-.003, -.156]. This again hints at the presence of an overall positive moral consistency effect.

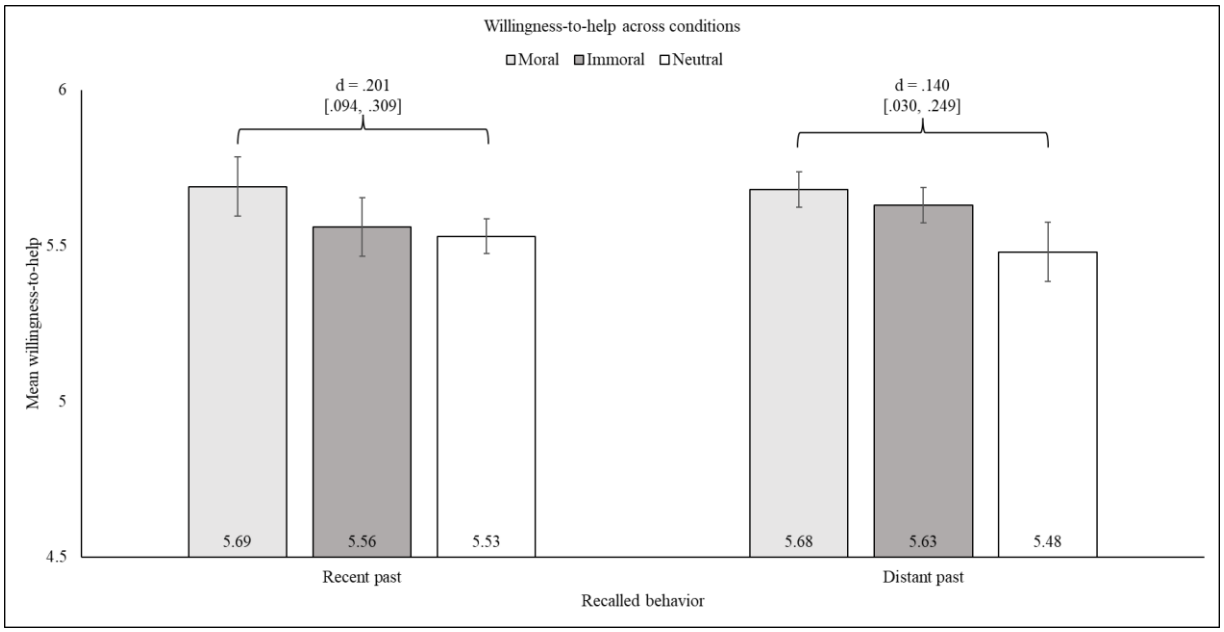


Figure 2B. Mean willingness-to-help (WTH) across conditions. Error bars refer to 95% confidence intervals.

To test H3a (moral licensing) and H3b (moral cleansing), we turn to planned comparisons to test for the presence of *compensatory* moral behavior in the "recent" distant conditions. WTH was significantly higher in the moral/recent (M = 5.69, SD = .89) than neutral/recent (M = 5.53, SD = 1.07) condition, t (5085) = -3.67, p_{Tukey} = .003, d = .201,

95% CI [.094, .309]. The opposite direction of the significant effect provides strong evidence *against* moral licensing (H3a).

However, WTH was not significantly different in the immoral/recent (M = 5.56, SD = .99) than neutral/recent (M = 5.53, SD = 1.07) condition, t (5085) = -.44, p_{Tukey} = .998, d = .024, 95% CI [-.082, .131]. An equivalence test with an upper bound of d = .14 was not significant, t (841) = 1.29, p = .098, thereby suggesting the standardized mean difference between the immoral and neutral condition may lie outside 0 and -.14, potentially having a magnitude bigger than our minimum effect size of interest. This presents inconclusive evidence for a null effect of moral cleansing (H3b) when comparing the "recent" conditions.

To test H3c (positive moral consistency) and H3d (negative moral consistency), we turn to planned comparisons to test for *consistent* moral behavior in the "distant" event conditions. WTH was not significantly higher in the moral/distant (5.68, SD = .98) than in the neutral/distant (M = 5.48, SD = 1.07) condition, t (5085) = 2.51, p_{Tukey} = .122, d = .140, 95% CI [.030, .249]. Note that the p-value *without* correcting for multiple post hoc comparisons, p = .012, is below the conventional significance level. Furthermore, an equivalence test with an upper bound of d = .14 was not significant, t (731) = .052, p = .479, thereby suggesting the standardized mean difference between the immoral and neutral condition may lie outside 0 and .14, a magnitude greater than our minimum effect size of interest. This presents inconclusive evidence for a null effect of positive moral consistency (H3c) the distant conditions.

However, WTH was not significantly different in the immoral/distant (M = 5.63, SD = 1.02) than in neutral/distant (M = 5.48, SD = 1.07) condition, t (5085) = 1.58, p_{Tukey} = .612, d = .088, 95% CI [.021, -.198]. An equivalence test with a lower bound of d = -.14 was significant, t (733) = -4.01, p < .001. So, the standardized mean difference between the immoral and neutral conditions may lie between 0 and -.14, a magnitude smaller than our

minimum effect size of interest, suggesting conclusive evidence of a null effect of negative moral consistency (H3d).

Discussion

Our study, as the first independent replication of C&P and the largest single-lab investigation in this domain, tested the hypothesis proposed by Conway and Peetz (2012) that recalling moral or immoral deeds from the recent (distant) past should lead to compensatory (consistent) moral behavior. The results did *not* support conceptual abstraction being a moderator. Instead, the results suggest a positive moral consistency effect. Specifically, recalling past moral behavior (versus immoral or neutral behavior), regardless of whether from the recent or distant past, increased prosocial intentions. These results challenge the prevailing theories on compensatory moral behavior and highlight the need for reevaluation in this domain.

Given that we could not replicate C&P's proposed moderation by conceptual abstraction, it may also be important to reconsider other hypothesized moderators in sequential moral behavior (e.g., Brown et al., 2011; Griep et al., 2021; Lalot et al., 2022; Susewind & Hoelzl, 2014; Susewind & Walkowitz, 2020). Understanding the factors that regulate whether doing good or bad leads to similar or opposite behavior afterward is crucial not only in moral psychology but also in domains such as organizational (List & Momeni, 2021) and consumer (Juhl et al., 2017) behavior.

Building on these broader implications, our findings of a robust positive moral consistency effect resonate with some existing literature in the field. For instance, a study tracking purchases of 8704 randomly selected Danish retail consumers over 20 months found that those who bought organic products at one point in time were more likely to buy *more* organic products (Juhl et al., 2017), supporting the positive moral consistency perspective.

Moreover, Rotella and Barclay (2020) found that participants donated significantly more after recalling past moral behavior than neutral behavior, and *marginally* after recalling past moral than immoral behavior. While we find effects in the same direction as Rotella and Barclay (2020), ours are more robust given the much larger sample size (N = 519 vs. N = 5091). Together with other null effects of compensatory moral behavior (e.g., Blanken et al., 2014; Eskine, 2013), our study supports the relevance of classical behavioral consistency theories (Festinger, 1954; Heider, 1946).

Our findings should be interpreted with some caveats. First, empirical findings from a single-lab online experiment conducted on a single population (participants based in the USA) limit cross-cultural generalizability. Interestingly, a culture-moderated meta-analysis examining the moral licensing effect (Simbrunner & Schlegelmilch, 2017) found moral licensing to be stronger among North American than Western European and Asian participants. Finding conclusive null effects with U.S. American participants, as we do, may suggest a need to rethink how we understand compensatory moral behavior because this was a cultural sample where the effects are supposedly stronger (Simbrunner & Schlegelmilch, 2017).

Second, we do not claim to rule out compensatory moral behavior altogether. A meta-analysis on interpersonal and intrapsychic mechanisms finds strong evidence in favor of compensatory behavior only when individuals are *observed* (Rotella et al., 2023). Because our study was conducted online and anonymously, recalling past behavior may not have granted a license or need to behave oppositely after (Rotella & Barclay, 2020). This may explain why we did not observe a robust licensing or cleansing effect.

Third, because our replication was conducted at least 11 years after Conway and Peetz (2012), experimental participants may have become less sensitive to recall-and-write tasks. Nonetheless, our manipulations were successful in that there were clear differences with

respect to both manipulation check measures and automated text analyses (Berger et al., 2020) also indicated clear differences by condition.

Future research could explore whether this positive moral consistency is observed in settings with aligned incentives, such as donation or dictator games, or in the context of dishonesty measures, to see if recalling moral/immoral deeds influences cheating behavior. Although we find robust evidence of positive moral consistency across one prior and one sequential behavior, it can be interesting to test how consistent behaviors are across *multiple* subsequent behaviors.

Conclusion

Our large-scale replication and extension of Conway and Peetz's (2012) study challenge longstanding assumptions in the field of moral psychology, particularly regarding the role of conceptual abstraction in moderating moral behavior. Our findings, which did not support the hypothesized interaction between moral behavior recollection and its temporal distance, suggest a more prevalent influence of positive moral consistency across different temporal contexts. This highlights a potential need to reevaluate the theoretical frameworks underpinning compensatory moral behavior. By revealing the robustness of moral consistency, irrespective of whether past behavior is recalled from the recent or distant past, our study supports the classical consistency perspective in moral decision-making and underscores the importance of registered replications in psychology.

References

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes, 35*, 124–140. https://doi.org/10.1016/0749-5978(85)90049-4

Berger, J., Sherman, G., & Ungar, L. (2020). *TextAnalyzer* [Computer software]. http://textanalyzer.org

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin, 41*, 540–558. https://doi.org/10.1177/0146167215572134

Blanken, I., van de Ven, N., Zeelenberg, M., & Meijers, M. H. C. (2014). Three attempts to replicate the moral licensing effect. *Social Psychology, 45*, 232–238. https://doi.org/10.1027/1864-9335/a000189

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin, 88*, 1–45. https://doi.org/10.1037/0033-2909.88.1.1

Brown, R. P., Tamborski, M., Wang, X., Barnes, C. D., Mumford, M. D., Connelly, S., & Devenport, L. D. (2011). Moral credentialing and the rationalization of misconduct. *Ethics & Behavior, 21*, 1–12. https://doi.org/10.1080/10508422.2011.537566

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science, 13*, 149–154. https://doi.org/10.1177/1745691617706516

Chambers, C. D., & Tzavella, L. (2021). The past, present and future of Registered Reports. *Nature Human Behaviour, 6*, 29–42. https://doi.org/10.1038/s41562-021-01193-7

Conway, P., & Peetz, J. (2012). When does feeling moral actually make you a better person? Conceptual abstraction moderates whether past moral deeds motivate consistency or

compensatory behavior. *Personality and Social Psychology Bulletin*, *38*, 907–919. https://doi.org/10.1177/0146167212442394

Cornelissen, G., Bashshur, M. R., Rode, J., & Le Menestrel, M. (2013). Rules or consequences? The role of ethical mind-sets in moral dynamics. *Psychological Science*, *24*, 482–488. https://doi.org/10.1177/0956797612457376

DeVoe, S. E., & Pfeffer, J. (2007). When time is money: The effect of hourly payment on the evaluation of time. *Organizational Behavior and Human Decision Processes*, *104*, 1–13. https://doi.org/10.1016/j.obhdp.2006.05.003

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Effron, D. A., Cameron, J. S., & Monin, B. (2009). Endorsing Obama licenses favoring Whites. *Journal of Experimental Social Psychology*, *45*, 590–593. https://doi.org/10.1016/j.jesp.2009.02.001

Effron, D. A., & Conway, P. (2015). When virtue leads to villainy: Advances in research on moral self-licensing. *Current Opinion in Psychology*, *6*, 32–35. https://doi.org/10.1016/j.copsyc.2015.03.017

Eskine, K. J. (2013). Wholesome foods and wholesome morals?: Organic foods reduce prosocial behavior and harshen moral judgments. *Social Psychological and Personality Science*, *4*, 251–254. https://doi.org/10.1177/1948550612447114

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117–140. https://doi.org/10.1177/001872675400700202

Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, *4*, 195–202. https://doi.org/10.1037/h0023552

Gholamzadehmir, M., Sparks, P., & Farsides, T. (2019). Moral licensing, moral cleansing and pro-environmental behaviour: The moderating role of pro-environmental attitudes. *Journal of Environmental Psychology*, *65*, 101334. https://doi.org/10.1016/j.jenvp.2019.101334

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research*, *44*, 196–210. https://doi.org/10.1093/jcr/ucx047

Griep, Y., Germeys, L., & Kraak, J. M. (2021). Unpacking the relationship between organizational citizenship behavior and counterproductive work behavior: Moral licensing and temporal focus. *Group & Organization Management*, *46*, 819–856. https://doi.org/10.1177/1059601121995366

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology*, *21*, 107–112. https://doi.org/10.1080/00223980.1946.9917275

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin*, *37*, 701–713. https://doi.org/10.1177/0146167211400208

Juhl, H. J., Fenger, M. H. J., & Thøgersen, J. (2017). Will the consistent organic food consumer step forward? An empirical analysis. *Journal of Consumer Research, 44*, 519–535. https://doi.org/10.1093/jcr/ucx052

Karmarkar, U. R., & Bollinger, B. (2015). BYOB: How bringing your own shopping bags leads to treating yourself and the environment. *Journal of Marketing, 79*, 1–15. https://doi.org/10.1509/jm.13.0228

Kuper, N., & Bott, A. (2019). Has the evidence for moral licensing been inflated by publication bias? *Meta-Psychology, 3*. https://doi.org/10.15626/MP.2018.878

Lalot, F., Falomir-Pichastor, J. M., & Quiamzade, A. (2022). Regulatory focus and self-licensing dynamics: A motivational account of behavioural consistency and balancing. *Journal of Environmental Psychology*, *79*, 101731. https://doi.org/10.1016/j.jenvp.2021.101731

List, J. A., & Momeni, F. (2021). When corporate social responsibility backfires: Evidence from a natural field experiment. *Management Science*, *67*, 8–21. https://doi.org/10.1287/mnsc.2019.3540

Maier, M., Bartoš, F., Oh, M., Wagenmakers, E.-J., Shanks, D., & Harris, A. J. L. (2022). *Adjusting for publication bias reveals that evidence for and size of construal level theory effects is substantially overestimated* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/r8nyu

Mazar, N., & Zhong, C.-B. (2010). Do green products make us better people? *Psychological Science, 21*, 494–498. https://doi.org/10.1177/0956797610363538

Moery, E., & Calin-Jageman, R. J. (2016). Direct and conceptual replications of Eskine (2013): Organic food exposure has little to no effect on moral judgments and prosocial behavior. *Social Psychological and Personality Science, 7*, 312–319. https://doi.org/10.1177/1948550616639649

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*, 33–43. https://doi.org/10.1037/0022-3514.81.1.33

Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology*, *67*, 363–385. https://doi.org/10.1146/annurev-psych-010213-115120

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. https://doi.org/10.1126/science.aac4716

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*, 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Qualtrics (Version 2022). (2022). [Computer software]. Qualtrics. https://www.qualtrics.com

R Core Team. (2013). *R* [Computer software]. R Foundation for Statistical Computing. www.R-project.org/

Reed, A., Aquino, K., & Levy, E. (2007). Moral identity and judgments of charitable behaviors. *Journal of Marketing*, *71*, 178–193. https://doi.org/10.1509/jmkg.71.1.178

Reed, A., Kay, A., Finnel, S., Aquino, K., & Levy, E. (2016). I don't want the money, I just want your time: How moral identity overcomes the aversion to giving time to

prosocial causes. *Journal of Personality and Social Psychology*, *110*, 435–457. https://doi.org/10.1037/pspp0000058

Rosen, C. C., Koopman, J., Gabriel, A. S., & Johnson, R. E. (2016). Who strikes back? A daily investigation of when and why incivility begets incivility. *Journal of Applied Psychology*, *101*, 1620–1634. https://doi.org/10.1037/apl0000140

Rotella, A., & Barclay, P. (2020). Failure to replicate moral licensing and moral cleansing in an online experiment. *Personality and Individual Differences*, *161*, 109967. https://doi.org/10.1016/j.paid.2020.109967

Rotella, A., Jung, J., Chinn, C., & Barclay, P. (2023). *Observation moderates the moral licensing effect: A meta-analytic test of interpersonal and intrapsychic mechanisms* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/tmhe9

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-Regulation. *Psychological Science*, *20*, 523–528. https://doi.org/10.1111/j.1467-9280.2009.02326.x

Simbrunner, P., & Schlegelmilch, B. B. (2017). Moral licensing: A culture-moderated meta-analysis. *Management Review Quarterly*, *67*, 201–225. https://doi.org/10.1007/s11301-017-0128-0

Susewind, M., & Hoelzl, E. (2014). A matter of perspective: Why past moral behavior can sometimes encourage and other times discourage future moral striving. *Journal of Applied Social Psychology*, *44*, 201–209. https://doi.org/10.1111/jasp.12214

Susewind, M., & Walkowitz, G. (2020). Symbolic moral self-completion – social recognition of prosocial behavior reduces subsequent moral striving. *Frontiers in Psychology*, *11*, 560188. https://doi.org/10.3389/fpsyg.2020.560188

The jamovi project. (2022). *Jamovi.* (2.3) [Computer software]. https://www.jamovi.org

Tiefenbeck, V., Staake, T., Roth, K., & Sachs, O. (2013). For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign. *Energy Policy, 57*, 160–171. https://doi.org/10.1016/j.enpol.2013.01.021

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review, 110*, 403–421. https://doi.org/10.1037/0033-295X.110.3.403

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*, 440–463. https://doi.org/10.1037/a0018963

Urban, J., Bahník, Š., & Kohlová, M. B. (2019). Green consumption does not make people cheat: Three attempts to replicate moral licensing effect due to pro-environmental behavior. *Journal of Environmental Psychology, 63*, 139–147. https://doi.org/10.1016/j.jenvp.2019.01.011

van der Werff, E., Steg, L., & Keizer, K. (2014). Follow the signal: When past pro-environmental actions signal who you are. *Journal of Environmental Psychology, 40*, 273–282. https://doi.org/10.1016/j.jenvp.2014.07.004

Wang, Y., Rodríguez De Gil, P., Chen, Y.-H., Kromrey, J. D., Kim, E. S., Pham, T., Nguyen, D., & Romano, J. L. (2017). Comparing the performance of approaches for testing the homogeneity of variance assumption in ane-Factor ANOVA Models. *Educational and Psychological Measurement, 77*, 305–329. https://doi.org/10.1177/0013164416645162

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of*

Personality and Social Psychology, *111*, 493–504.

https://doi.org/10.1037/pspa0000056

ONLINE SUPPLEMENT

For the manuscript

Does Conceptual Abstraction Moderate Whether Past Moral Deeds Motivate Consistency or Compensatory Behavior? A Registered Replication and Extension of Conway and Peetz (2012)

Table of Contents

1. Codebook of all measurements used in the study

Table S1. List of measures and codebook for the data file.

Variable	# of items (a)	Item	Data File Column Name	Scale
Attention Check 1	1	Caution This text is about the following issue. In surveys like ours, sometimes participants do not carefully read the instructions and just click randomly to finish the survey. This leads to several random responses and that can compromise the results. To confirm that you read our instructions carefully, please select "A great deal" as your answer to the question "How much do you like sports?" below. How much do you like sports?	Atttention_Check1	Multiple Choice: A great deal A lot A moderate amount A little None at all
Participants' text response prompt		Please describe the event you recalled by typing in the text box below. Please provide as much detail as you can, and write at least a paragraph with complete sentences. Your responses are completely anonymous.	Response	
Willingness to Volunteer (WTV)	5	1. Volunteering is a worthwhile use of my time even if I do not get paid. 2. I am willing to volunteer for an organization I care about without financial compensation for me.	1. Vol1 2. Vol2 3. Vol3RC 4. VolRC4 5. VolRC5	Strongly disagree to Strongly Agree (7-point Likert scale)

		3. Even for an organization I care about, I am unwilling to work without getting paid (R). 4. Without some financial compensation, it is not worth doing volunteer work (R). 5. I am unlikely to undertake any type of work without being paid (R).		
Attention Check 2		PLEASE SELECT "Strongly disagree" here.	Attention_Check2	Strongly disagree to Strongly Agree (7-point Likert scale)
Willingness to Help (WTH) 4 scenarios, each with two questions, with the latter pertaining to the deservingness of help	8	1. (Cindy) Imagine you are a customer in the restaurant that Cindy visits for lunch. You notice that when the bill comes she is embarrassed to find that she does not have quite enough local currency to pay for her meal (they accept cash only). She apologizes to the waiter and asks for directions to the nearest bank machine, which turns out to be quite far away. You could save her the trouble and embarrassment by giving her the small amount she needs (approximately 50 cents). How likely would you be to give Cindy a small amount of money to save her the hassle of walking many blocks to a bank machine and back? How much does Cindy deserve your help? 2. (Susan) Imagine that you are one of Susan's co-workers. At some	The help items. 1. Cindy_Help 2. Susan_Help 3. Bill_Help 4. Jim_Help The deservingness items 1. Cindy_Deserve 2. Susan_Deserve 3. Bill_Deserve 4. Jim_Deserve	Not at all to Very likely/much (7-point Likert scale)

		point, you are chatting and she asks a favor of you. She needs to deliver a large heavy parcel to the local post office. She does not have a car, but you do, so she asks if you would mind taking it for her on your way home. The post office is in the opposite direction as your home, so helping Susan will add at least 10-15 minutes to your evening commute. How likely would you be to agree to deliver Susan's parcel? How much does Susan deserve your help? 3. (Bill) Imagine you are one of Bill's neighbors. One day you see him working in his garden using a small hand-tool where a bigger one would be much easier to use. You happen to have the exact tool he needs, but you were planning to use it yourself later that day. Bill asks if you have any tools that would help make his gardening chores easier. How likely are you to offer your superior garden tool to Bill even though it would prevent you from using it on the same day? How much does Bill deserve your help? 4. (Jim) Imagine you work in the same building as Jim and sometimes see him in line to buy food at the		

		building food court. One day, the line is particularly slow and you realize that you may not have long to eat your lunch before you must go back to work. Then Jim asks if he could move ahead of you in line because he has only a few minutes before his next meeting and might not get to buy lunch otherwise. How likely are you to let Jim move ahead of you in line even though it will make your short lunch even shorter? How much does Jim deserve your help?		
Attention Check 3		Please indicate your agreement with the statement below. I swim across the Atlantic Ocean to get to work every day.	Attention_Check2.0	Multiple Choice: Strongly disagree Disagree Agree Strongly agree
Event Valence (manipulation check)	1	The event I wrote about made me feel good about myself.	Feel_Good	Completely disagree to completely agree (7-point Likert scale)
Event Distance (manipulation check)	1	The event I wrote about happened a long time ago.	Happened_Long	Completely disagree to completely agree (7-point Likert scale)
Specific Event Time (manipulation check)	1	Specifically, the event I wrote about happened	Happened_Specific	Multiple Choice: Within last week More than a year ago In between a week and a year
Feeling	20	Interested Excited Strong Enthusiastic Proud Alert Inspired	Feeling_1 Feeling_2 Feeling_3 Feeling_4 Feeling_5 Feeling_6 Feeling_7	Very slightly or not at all to Extremely (5-point Likert scale)

		Determined Attentive Active Distressed Upset Guilty Scared Hostile Irritable Ashamed Nervous Jittery Afraid	Feeling_8 Feeling_9 Feeling_10 Feeling_11 Feeling_12 Feeling_13 Feeling_14 Feeling_15 Feeling_16 Feeling_17 Feeling_18 Feeling_19 Feeling_20	
Age	1	What is your age in years	Age_Own	Blank box
Gender	1	What is your gender?	Gender_Own	Multiple Choice: Male Female Other/Prefer not to say
Education	1	What is the highest level of education you have completed?	Education	Multiple Choice: Some high school or less High school diploma or GED Some college, but no degree Associates or technical degree Bacherlor's degree Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.) Prefer not to say
Income	1		What is your total household income before taxes during the past 12 months?	Multiple Choice: Less than $25,000 $25,000-$49,999 $50,000-$74,999 $75,000-$99,999 $100,000-$149,999 $150,000 or more Prefer not to say
Zip Code	1	What is your US Zip Code?		Open-ended question.
Moral_Identity	10	Lastly, visualize a person who is caring,	1. Identity_1_I 2. Identity_2_I	Strongly disagree to Strongly agree

		compassionate, fair, friendly, generous, helpful, hardworking, honest, and kind. Form a clear image of how that person would think, feel, and act and then indicate the extent to which you agree or disagree with the following statements. 1. It would make me feel good to be a person who has these characteristics. 2. Being someone who has these characteristics is an important part of who I am. 3. I often wear clothes that identify me as having these characteristics. 4. I would be ashamed to be a person who had these characteristics. 5. The types of things I do in my spare time (e.g., hobbies) clearly identify me as having these characteristics 6. The kinds of books and magazines that I read identify me as having these characteristics. 7. Having these characteristics is not really important to me (R). 8. The fact that I have these characteristics is communicated to others by my membership in certain organizations. 9. I am actively involved in activities that communicate to others	3. Identity_3_S 4. Identity_4_I_R 5. Identity_5_S 6. Identity_6_S 7. Identity_7_I_R 8. Identity_8_S 9. Identity_9_S 10. Identity_10_I	(7-point Likert scale)

		that I have these characteristics. 10. I strongly desire to have these characteristics.		

2. Experimental Manipulations

Table S2. Exact prompts for the recall tasks.

Morality X Construal Level	Prompt
Moral/Recent	Please recall a time within the last week when you acted in such a way that you felt righteous or honorable. Perhaps you were loyal to a friend, were generous when you could have been selfish, were kind to someone for no particular reason, or caring toward someone who needed you.
Moral/Distant	Please recall a time over a year ago when you acted in such a way that you felt righteous or honorable. Perhaps you were loyal to a friend, were generous when you could have been selfish, were kind to someone for no particular reason, or caring toward someone who needed you.
Immoral/Recent	Please recall a time within the last week when you acted in such a way that you felt guilty or ashamed. Perhaps you were disloyal to a friend, were greedy when you should have shared, were mean to someone for no particular reason, or uncaring toward someone who needed you.
Immoral/Distant	Please recall a time over a year ago when you acted in such a way that you felt guilty or ashamed. Perhaps you were disloyal to a friend, were greedy when you should have shared, were mean to someone for no particular reason, or uncaring toward someone who needed you.
Neutral/Recent	Please recall a time over a year ago when you acted in such a way that you felt neutral or indifferent. Perhaps you were doing something by yourself (e.g., shopping for groceries, doing chores) and your actions did not affect anyone in any way.
Neutral/Distant	Please recall a time within the last week when you acted in such a way that you felt neutral or indifferent. Perhaps you were doing something by yourself (e.g., shopping for groceries, doing chores) and your actions did not affect anyone in any way.

3. Descriptive Statistics

Table S3A. Means (M) and standard deviations (SD) of the two manipulation checks across conditions.

Dependent Variable	Recalled Event Valence	Temporal Distance	Cell size	M	SD
Feel Good	Moral	Recent	447	5.86	1.22
		Distant	1226	5.79	1.33
		Recent + Distant	1673	5.81	1.30
	Immoral	Recent	455	1.89	1.31
		Distant	1211	1.78	1.39
		Recent + Distant	1666	1.81	1.32
	Neutral	Recent	1316	4.03	1.65
		Distant	436	3.91	1.62
		Recent + Distant	1752	4.00	1.64
Happened Long	Recent	Moral	447	4.51	1.39
		Immoral	455	4.29	1.38
		Neutral	1316	4.32	1.42
		Moral + Immoral + Neutral	2218	1.54	1.17
	Distant	Moral	1226	4.63	1.32
		Immoral	1211	4.37	1.39
		Neutral	436	4.26	1.47
		Moral + Immoral + Neutral	2218	4.07	1.86

Table S3B. Means (M) and standard deviations (SD) of the dependent variables across conditions.

Dependent Variable	Recalled Event Valence	Temporal Distance	Cell size	M	SD
Willingness to Volunteer	Moral	Recent	447	4.51	1.39
		Distant	1226	4.63	1.32
		Recent + Distant	1673	4.60	1.34
	Immoral	Recent	455	4.29	1.38
		Distant	1211	4.37	1.39
		Recent + Distant	1666	4.35	1.39
	Neutral	Recent	1316	4.32	1.42
		Distant	436	4.26	1.47
		Recent + Distant	1752	4.31	1.43
Willingness to help	Moral	Recent	447	5.69	0.89

		Distant	1226	5.68	1.01
		Recent + Distant	1673	5.68	0.98
	Immoral	Recent	455	5.56	0.99
		Distant	1211	5.63	1.02
		Recent + Distant	1666	5.60	1.01
	Neutral	Recent	1316	5.53	1.07
		Distant	436	5.48	1.07
		Recent + Distant	1752	5.50	1.07

Table S3C. Means (M) and standard deviations (SD) of the dependent variables across contrasts.				
Dependent Variable	Combined cell	Cell size	M	SD
Willingness to Volunteer	Moral/Distant + Immoral/ Recent	1681	4.54	1.34
	Moral/Recent + Immoral/ Distant	1658	4.41	1.39
	Neutral/Distant + Neutral/Recent	1752	4.31	1.43
Willingness to help	Moral/Distant + Immoral/Recent	1681	5.65	1.01
	Moral/Recent + Immoral/Distant	1658	5.64	.99
	Neutral/Distant + Neutral/Recent	1752	5.50	1.07

4. Sensitivity Analyses (power & minimum detectable effect sizes)

4.1. Minimum detectable effect sizes for the replication analyses (compared to C&P).

Figure S1.1. Minimum detectable effect size in C&P's original study for the interaction effect of event valence X event distance with 90% power: f = .326.

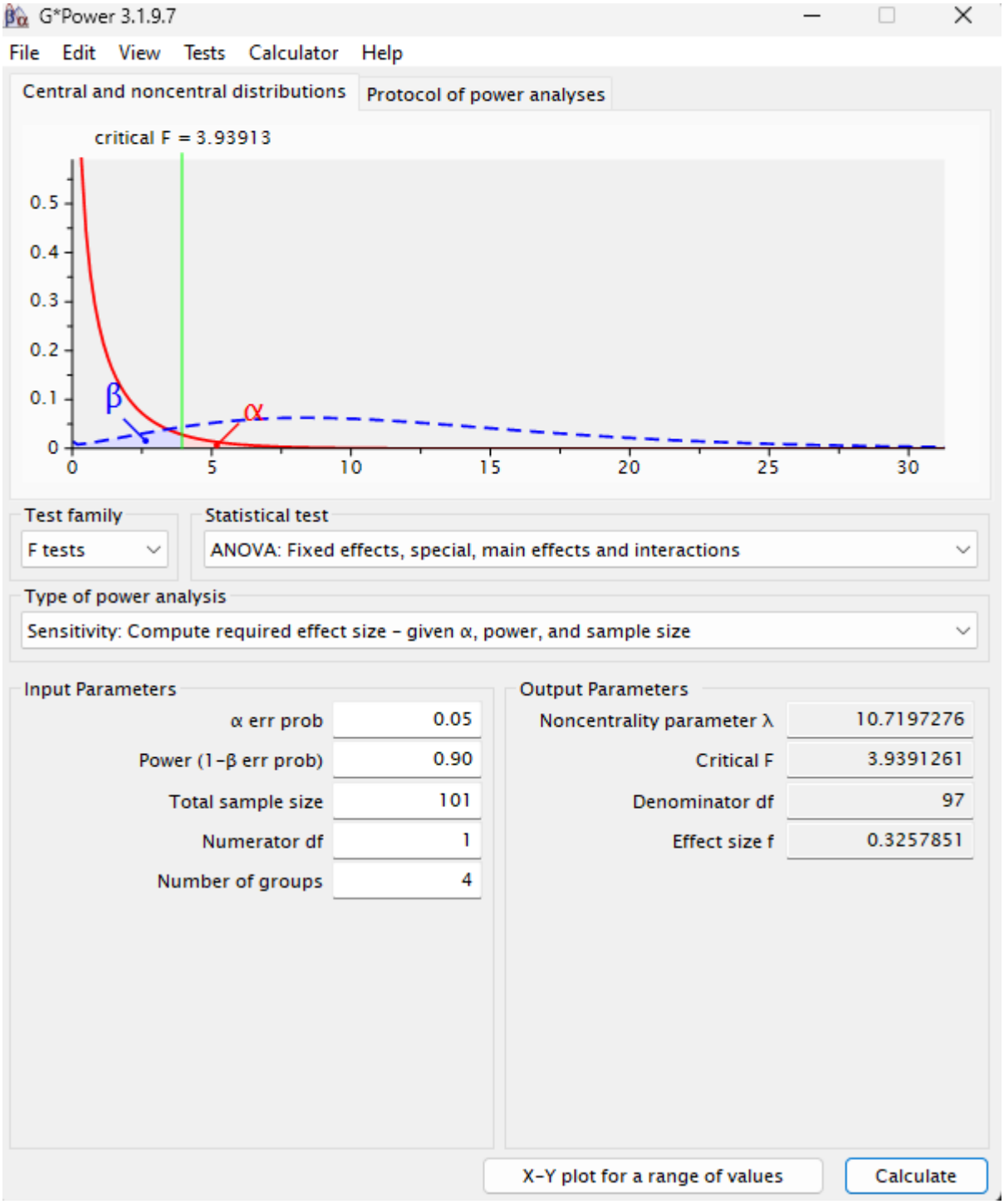


Figure S1.2. Minimum detectable effect size for the interaction effect of event valence X event distance with 90% power: f = .056.

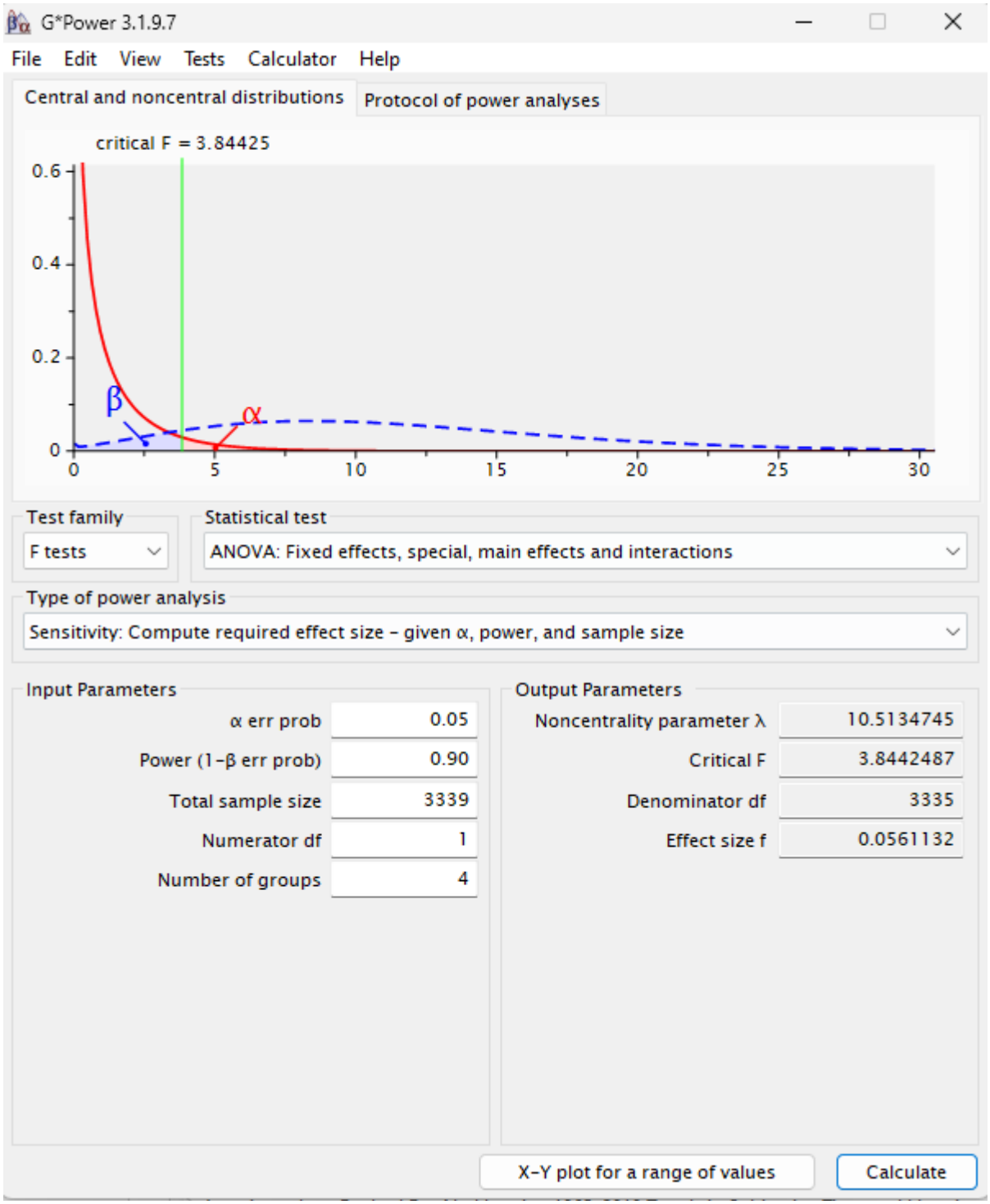


Figure S2.1. Minimum detectable effect size in C&P's original study for the planned contrast (H1: moral/distant + immoral/recent vs. moral/recent + immoral/distant): d = .651.

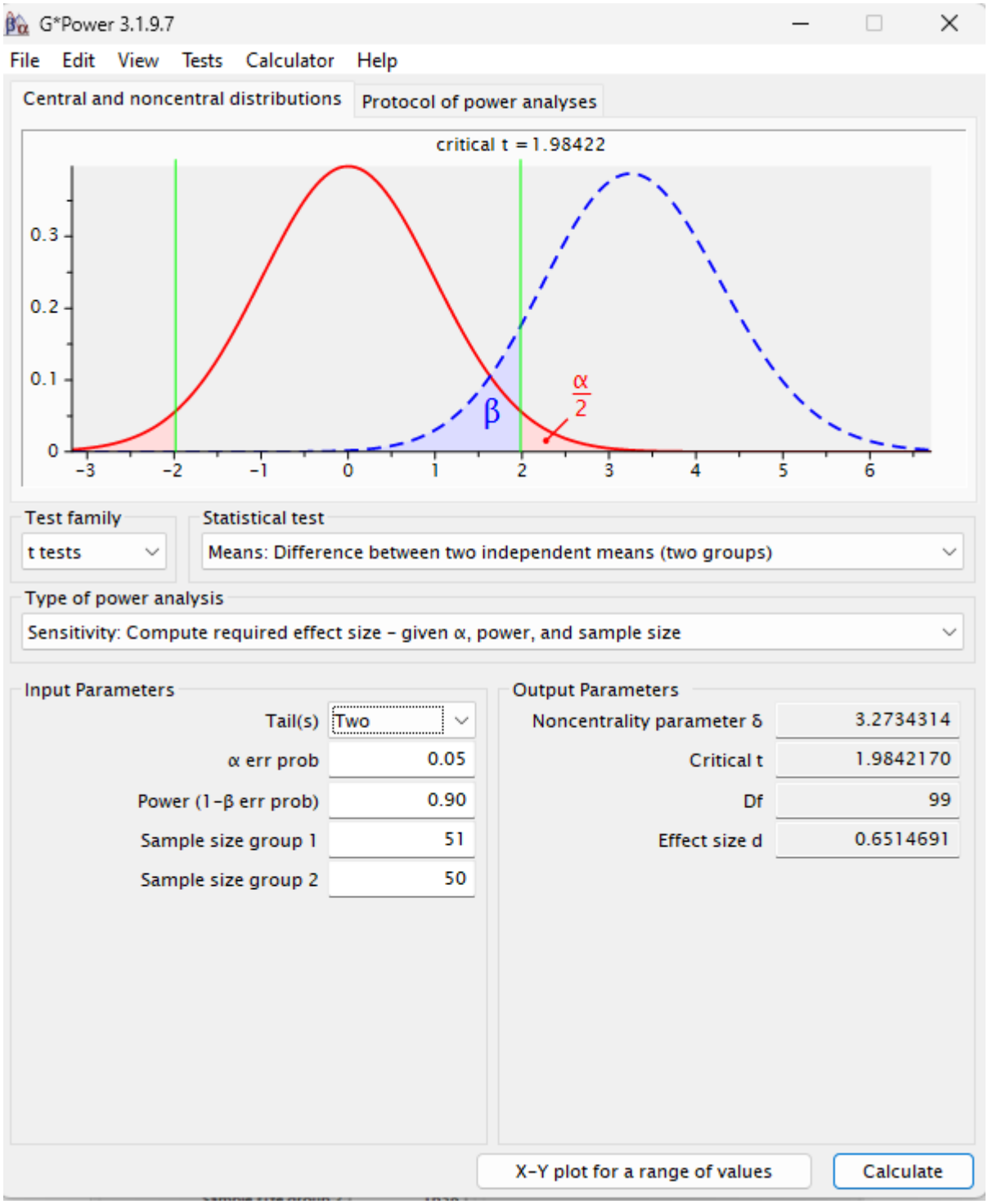
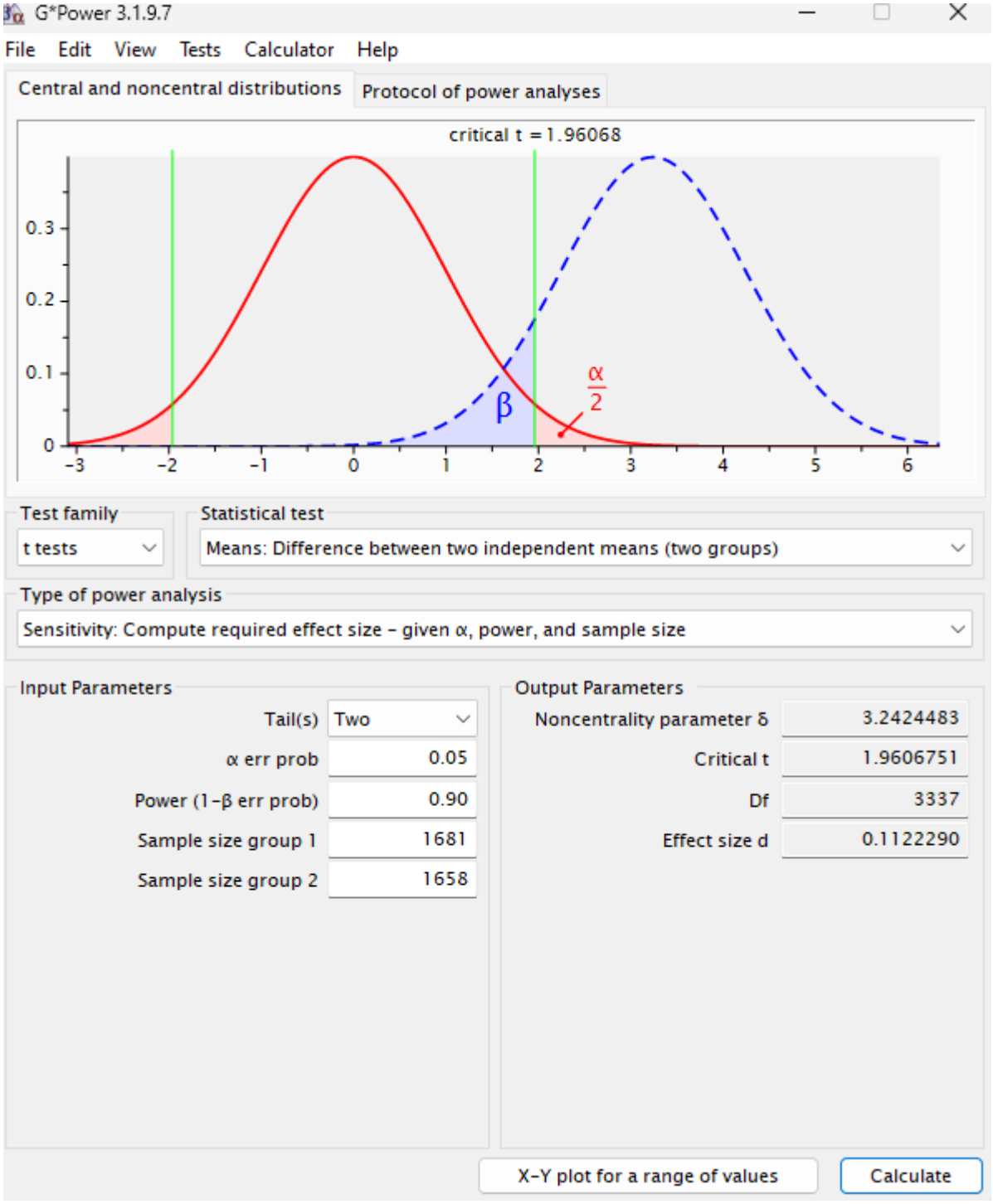


Figure S2.2. Minimum detectable effect size for the planned contrast (H1: moral/distant + immoral/recent vs. moral/recent + immoral/distant): d = .112.



4.2. Minimum detectable effect sizes for the extension analyses.

Figure S3. Minimum detectable effect size for the planned contrast (H2a: moral/distant + immoral/recent vs. neutral/recent + neutral/distant): d = .111

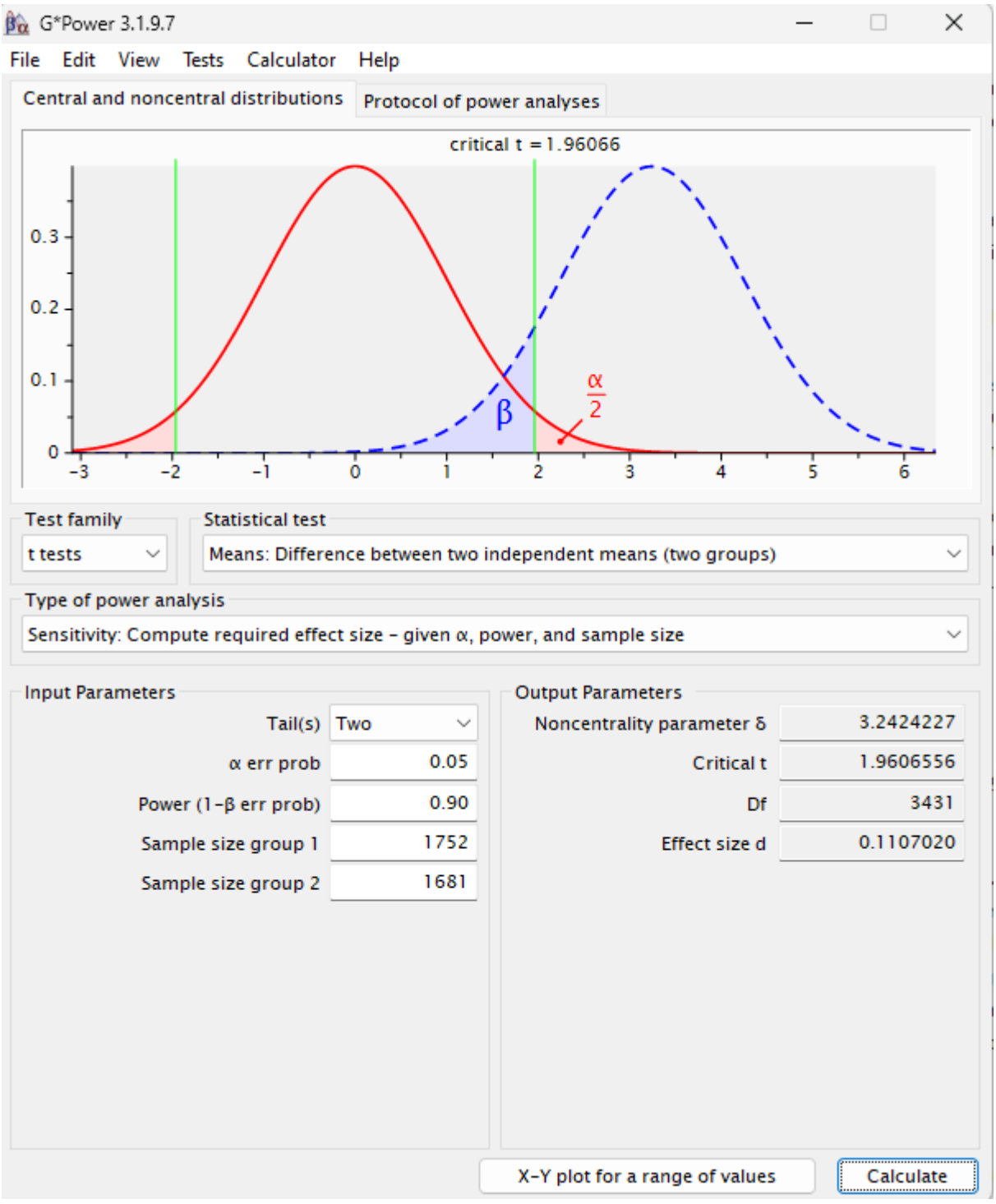


Figure S4. Minimum detectable effect size for the planned contrast (H2b: moral/recent + immoral/distant vs. neutral/recent + neutral/distant): d = .111

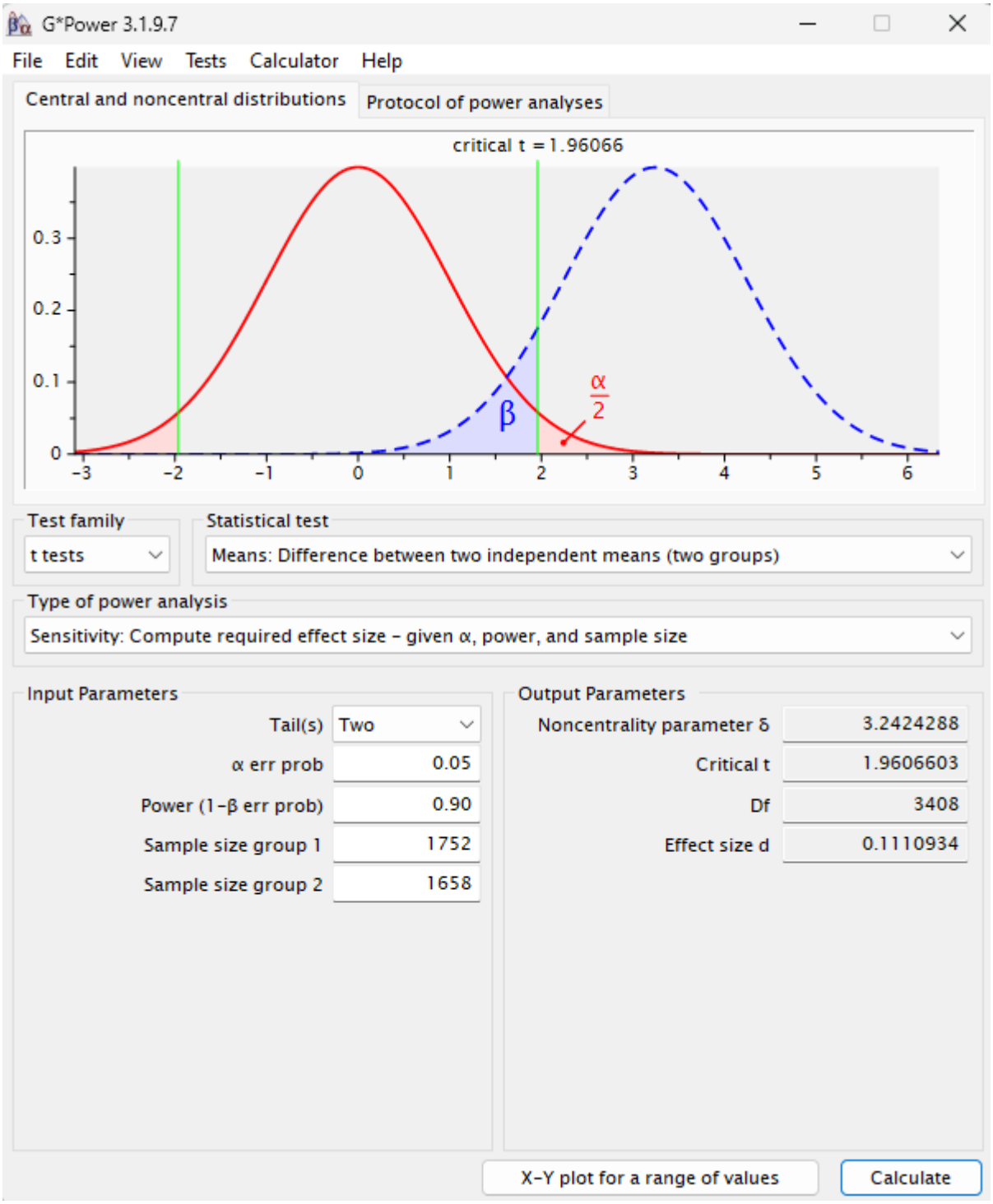


Figure S5. Minimum detectable effect size for moral licensing (H2c: moral/recent vs. neutral/recent): d =.200

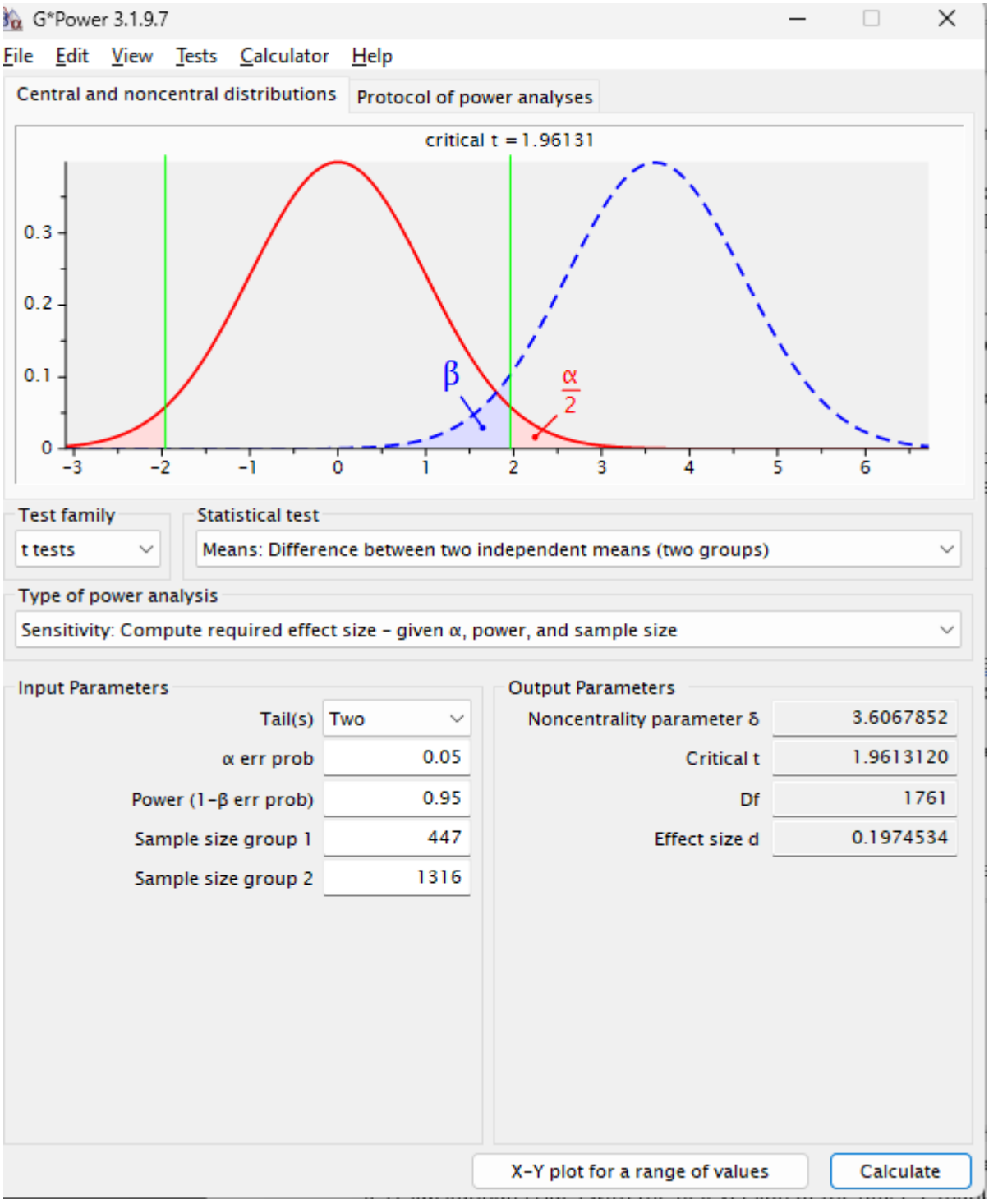


Figure S6. Minimum detectable effect size for moral cleansing (H2d: immoral/recent vs. neutral/recent): d = .196

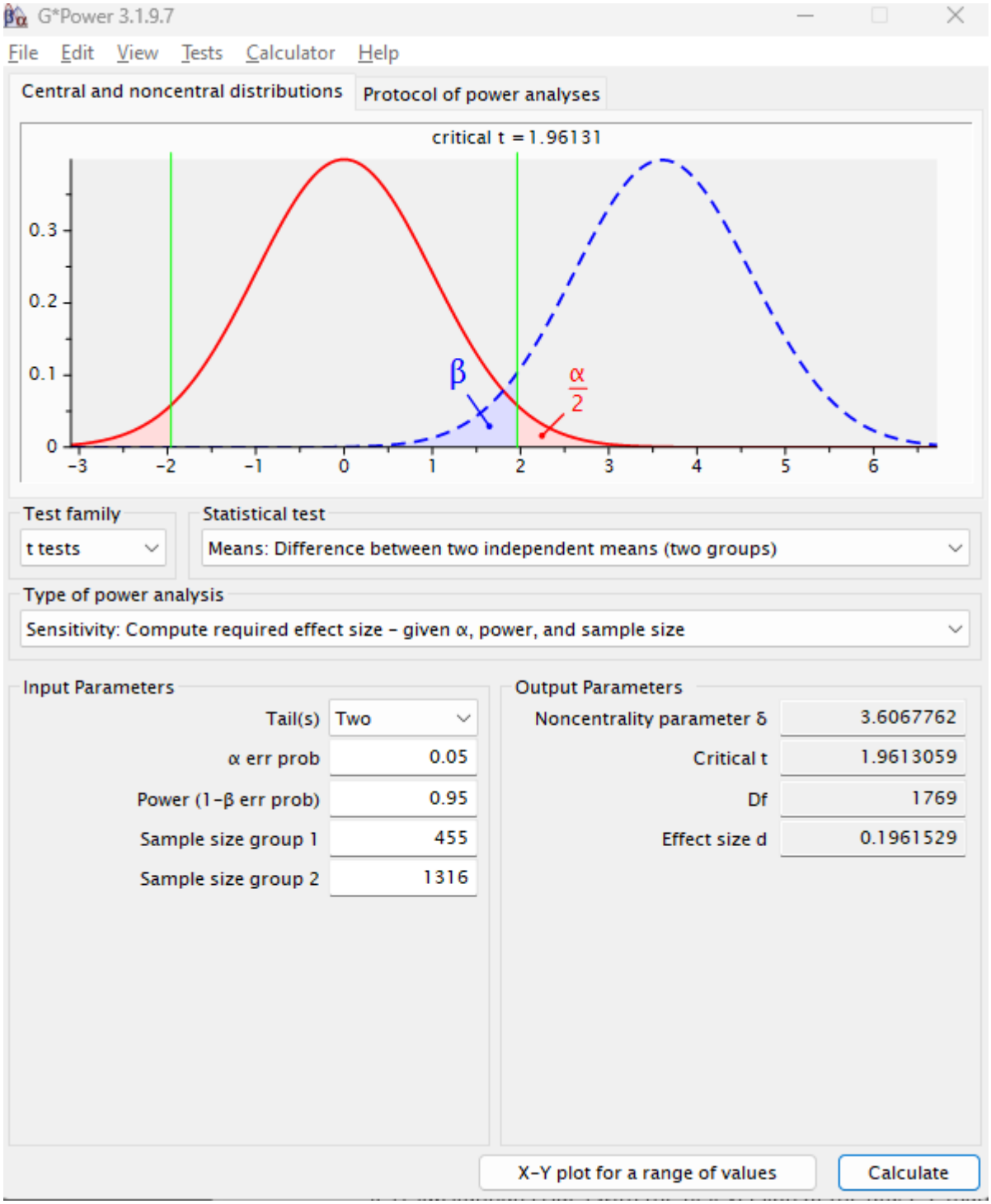


Figure S7. Minimum detectable effect size for positive moral consistency (moral/distant vs. neutral/distant): d = .201

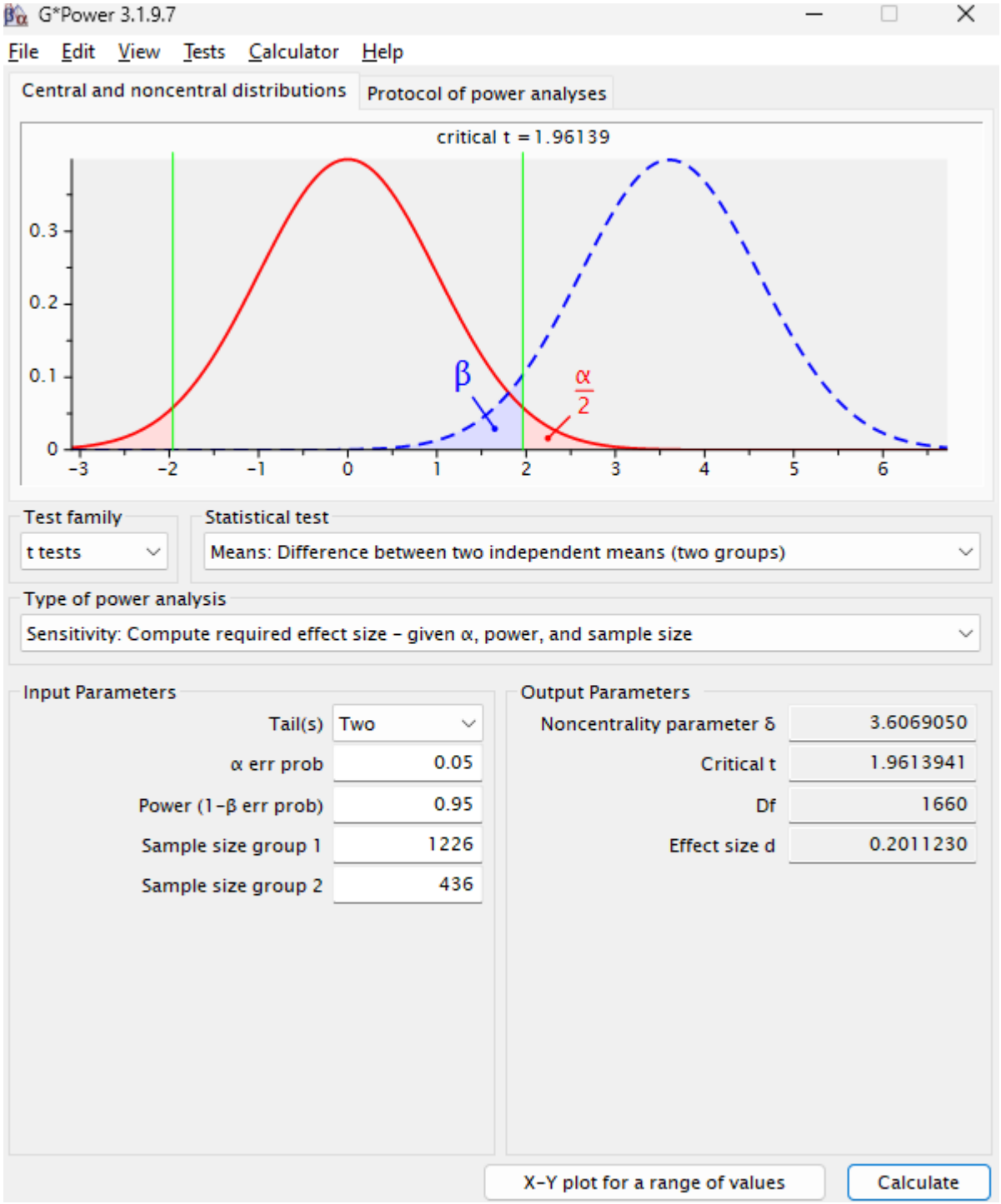
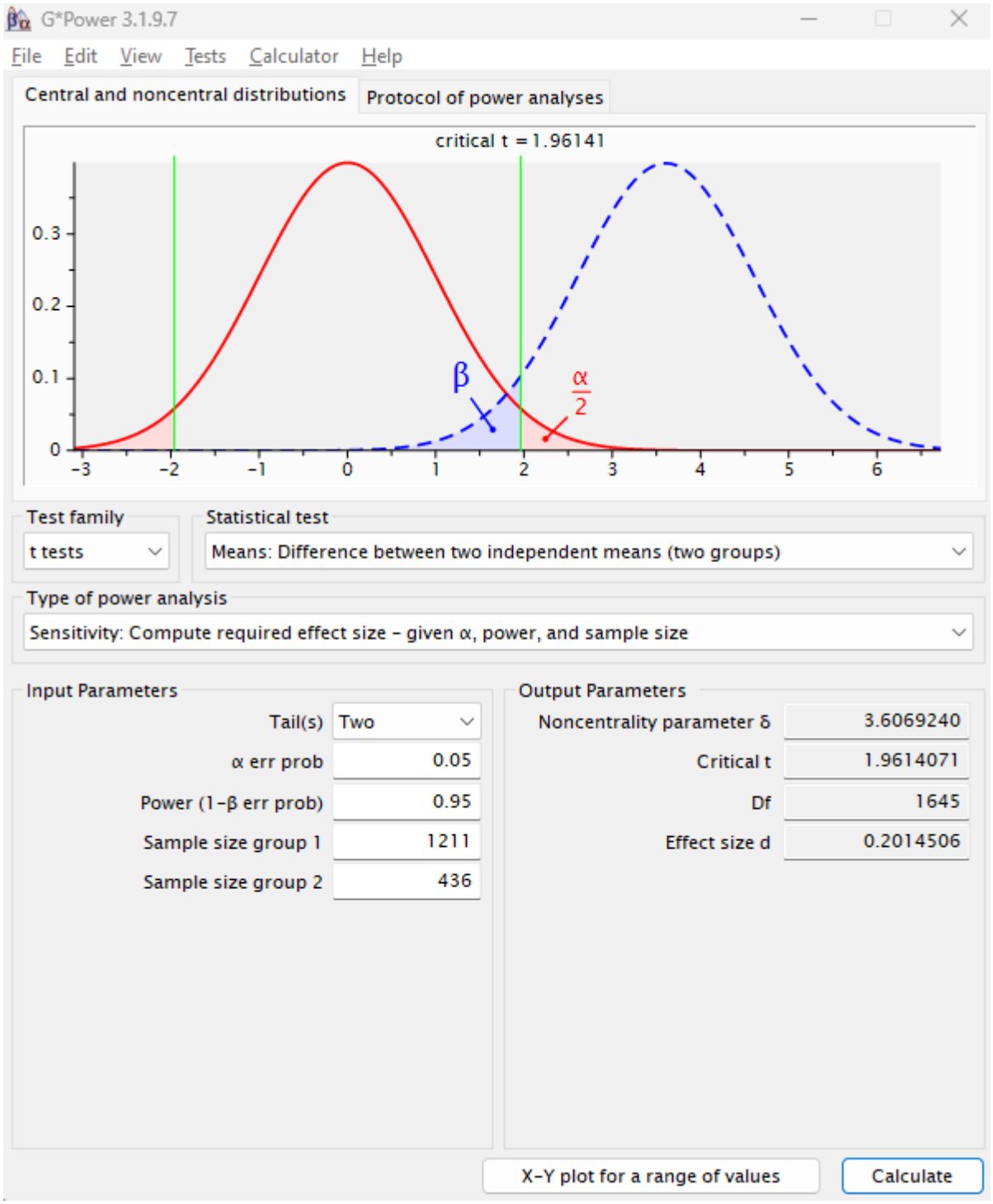


Figure S8. Minimum detectable effect size for negative moral consistency (immoral/distant vs. neutral/distant): d = .201



5. Comparing test statistics between the original and the replication

5.1. Manipulation checks

On a subset of the four original conditions of C&P, we first conducted a 2 (Event Valence: Moral vs. Immoral) X 2 (Event Distance: Recent vs. Distant) between-participants ANOVA on recalled event positivity. The results revealed the expected effect of event valence on event positivity, F (1, 3335) = 6144.85, p < .001, $\eta^2_p = .648$. Event distance did not moderate the effect of event valence, F (1, 3335) = 0.079, p = .778, $\eta^2_p < .001$. Participants instructed to recall and write about a moral event (M = 5.81, SD = 1.30) reported their event to be significantly more positive than those in the immoral conditions (M = 1.81, SD = 1.31), t (3335) = 78.4, $p_{Tukey} < 0.001$, d = 3.06, 95% CI [2.95, 3.16].

Table S4A: Comparison of test statistics of the Event Valence manipulation check between C&P and the current study.

Predictor	Study	df	F	p	eta-squared
Valence	C&P	1	287.96	<.001	.75
	Current	1	6144.85	<.001	.648
Distance	C&P	NA			
	Current	1	2.97	.085	.00
Valence/Distance	C&P	1	.70	.406	.01
	Current	1	.08	.78	.00

A 2 (Event Valence: Moral vs. Immoral) X 2 (Event Distance: Recent vs. Distant) between-participants ANOVA on recalled event distance revealed the expected effect of event distance on perceived distance, F (1, 3335) = 1415.76, p < .001, $\eta^2_p = .298$. Event valence did not moderate the effect of event distance, F (1, 3335) = 0.061, p = .805, $\eta^2_p < .001$. Participants instructed to recall and write about a distant event (M = 4.06, SD = 1.87) perceived the event to be significantly more distant than participants in the recent conditions (M = 1.55, SD = 1.15), t (3335) = 37.6, $p_{Tukey} <$ 0.001, d = 1.47, 95% CI [1.38, 1.55].

Table S4B: Comparison of test statics of the Event Distance manipulation check between C&P and the current study.

Predictor	Study	df	F	p	eta-squared
Valence	C&P	NA			
	Current	1	4.54	.033	.001
Distance	C&P	1	83.72	<.001	.47
	Current	1	1415.76	<.001	.298
Valence/Distance	C&P	1	1.43	.236	.02
	Current	1	.061	.805	<.001

5.2. Dependent variables

Next, we compare exact test statistics for the dependent variables between C&P and the current study.

Table S4C: Comparison of test statics of willingness-to-volunteer (WTV) between C&P and the current study.

Predictor	Study	df	F	p	eta-squared
Valence	C&P	1	.13	.716	<.01
	Current	1	20.33	<.001	.006
Distance	C&P	1	.07	.798	<.01
	Current	1	3.817	0.051	0.001
Valence X Distance	C&P	1	5.37	.023	.05

	Current	1	0.176	0.675	<.001

Table S4D: Comparison of test statics of willingness-to-help (WTH) between C&P and the current study.

Predictor	Study	df	F	p	eta-squared
Valence	C&P	1	1.26	.265	.01
	Current	1	5.67	.018	.002
Distance	C&P	1	.36	.550	<.01
	Current	1	.522	.470	<.001
Valence X Distance	C&P	1	5.97	.016	.06
	Current	1	1.052	.305	<.001

6. Statistical Considerations

6.1. Comparing test statistics between including versus excluding participants who responded incorrectly to the additional Event Distance manipulation check.

Around 20% of participants reported that the incident they recalled was "between one week and one year", thereby potentially undermining the Event Distance manipulation.

Table S5A. Responses to the specific manipulation check question for Event Distance.			
	within last week	in between a week and a year	more than a year ago
Recent (total = 2218)	2012; 90.7%	160; 7.2%	46; 2.1%
Distant (total = 2873)	190; 6.6%	851; 29.6%	1832; 63.8

So, we conducted all our analyses on a dataset where we excluded those participants. Overall, the results do not significantly change.

Table S5B. Comparing test statistics between including versus excluding participants who responded incorrectly to the additional Event Distance manipulation check.			
		Final Dataset	Dataset excluding wrong responders
Replication	Willingness to volunteer	Event Valence: F (1, 3335) = 20.332, p <.001, $\eta^2_p = .006$	Event Valence: F (1, 2325) = 17.80, p <.001, $\eta^2_p = .008$
		Event Distance: F (1, 3335) = 3.817, p = .051, $\eta^2_p = .001$	Event Distance: F (1, 2325) = .46, p = .497, $\eta^2_p < .001$
		Event Valence X Event Distance interaction: F (1, 3335) = .176, p = .675, $\eta^2_p < .001$.	Event Valence X Event Distance interaction: F (1, 2492) = .02, p = .886, $\eta^2_p < .001$.
		Moral vs. Immoral post hoc: t (2327) = 4.51, p < .001, d = .176 [.099, .252]	Moral vs. Immoral post hoc: t (2327) = 4.41, p < .001, d = .183 [.101, .264]
	Willingness to help	Event Valence: F (1, 3335) = 5.64, p = 018, $\eta^2_p = .002$	Event Valence: F (1, 2325) = 2.36, p = .125, $\eta^2_p = .001$
		Event Distance: F (1, 3335) = .519, p = .47, $\eta^2_p < .001$	Event Distance: F (1, 2325) = .015, p = .901, $\eta^2_p < .001$
		Event Valence X Event Distance interaction: F (1, 3335) = 1.052, p = .305, $\eta^2_p < .01$.	Event Valence X Event Distance interaction: F (1, 2325) = .884, p = .347, $\eta^2_p < .001$.
Extension	Willingness to volunteer	Event Valence: F (2, 5085) = 15.54, p < .001, $\eta^2_p = .006$	Event Valence: F (2, 3838) = 11.67, p < .001, $\eta^2_p = .006$
		Event Distance: F (1, 5085) = 1.16, p = .281, $\eta^2_p = .000$.	Event Distance: F (1, 3838) = 0.501, p = .479, $\eta^2_p < .001$
		F (2, 5085) = 1.69, p = .184, $\eta^2_p = .001$	Event Valence X Event Distance interaction: F (2, 3838) = 0.018, p = .982, $\eta^2_p < .001$
		Moral vs. Neutral post hoc: t (5085) = 5.15, p < .001, d = .201 [.124, .278]	Moral vs. Neutral post hoc: t (3838) = 4.21, p < .001, d = .191 [.102, .279]
		Immoral vs. Neutral post hoc:	Immoral vs. Neutral post hoc:

		t (5085) = 0.730, p = .746, d = .028 [-.048, .105]	t (3838) = 0.207, p = .977, d = .009 [-.079, .097]
	Willingness to help	Event Valence: F (2, 5085) = 9.53, p < .001, $\eta^2_p = .004$	Event Valence: F (2, 3838) = 8.61, p < .001, $\eta^2_p = .004$
		Event Distance: F (1, 5085) = 1.20, p = .273, $\eta^2_p < .000$	Event Distance: F (1, 3838) = 0.020, p = .888, $\eta^2_p < .001$
		Event Valence X Event Distance interaction: F (2, 5085) = .555, p = .574, $\eta^2_p < .001$	Event Valence X Event Distance interaction: F (2, 3838) = 0.787, p = .455, $\eta^2_p < .001$
		Moral vs. Neutral post hoc: t (5085) = 4.36, p < .001, d = .171 [.094, .247]	Moral vs. Neutral post hoc: t (3838) = 4.10, p < .001, d = .186 [.097, .274]
		Immoral vs. Neutral post hoc: t (5085) = 2.05, p = .101, d = .080 [.003, .156]	Immoral vs. Neutral post hoc: t (3838) = 2.69, p = .020, d = .121 [.033, .209]

6.2 Robust linear regressions

Table S5C. Outputs from robust linear regressions.

	Helping Intention		Volunteering Intention	
Predictors	*Estimates*	*std. Beta*	*Estimates*	*std. Beta*
(Intercept)	5.58 *** (-0.01)	-0.01 (-0.07 – 0.04)	4.40 *** (-0.01)	-0.01 (-0.07 – 0.05)
Distance2 [Distant]	0.05 (0.05)	0.05 (-0.06 – 0.16)	-0.03 (-0.02)	-0.02 (-0.14 – 0.09)
Valence [Moral]	0.15 ** (0.14)	0.14 (0.05 – 0.24)	0.19 * (0.14)	0.14 (0.03 – 0.25)
Valence [Immoral]	0.04 (0.04)	0.04 (-0.07 – 0.14)	-0.06 (-0.04)	-0.04 (-0.15 – 0.07)
Distance2 [Distant] X Valence [Moral]	-0.02 (-0.02)	-0.02 (-0.17 – 0.13)	0.15 (0.11)	0.11 (-0.05 – 0.27)
Distance2 [Distant] X Valence [Immoral]	0.05 (0.04)	0.04 (-0.11 – 0.20)	0.14 (0.10)	0.10 (-0.06 – 0.26)
Observations	5091		5091	
R^2 / R^2 adjusted	0.005 / 0.004		0.009 / 0.008	

** p<0.05 ** p<0.01 *** p<0.001*

7. Heterogeneity in prosocial intentions

It may be that participants' responses in willingness to volunteer and the four scenarios of willingness to help may have a common source of error terms. So, it may be worth examining them as five within-participant measures to get more precise estimates of the effects. Accordingly, we perform two linear mixed-effects models using the lme4 package in R (Bates et al., 2014) with fixed effects of event valence, event distance, and event valence X event distance, with random intercepts of participant ID and Prosociality Category (willingness to volunteer and the five scenarios).

We find only event valence to be a significant predictor. First, compared to the reference level of neutral event, participants reported higher prosocial intentions in the moral event condition (β = .11, SE = .04, p = .007). Second, compared to the reference level of moral event, participants reporter lower prosocial intentions in both the neutral (β = -.11, SE = .04, p = .007) and immoral (β = -.09, SE = .04, p = .019) conditions. Please see Tables S6A and S6B for full outputs from the models.

Table S6A. Outputs from linear mixed-effects model. We combined the composite measure of Volunteering intention and the Helping intentions of the four helping scenarios to have five measures of "Prosocial Intentions" per participant. (*Reference levels: Valence = Neutral, Distance = Recent)*

Predictors	*Estimates (SE)*	*std. Beta (SE)*	*Estimates CI*	*standardized CI*	*Statistic*	*p*
(Intercept)	5.28 (0.36)	-0.05 (0.22)	4.58 – 5.98	-0.48 – 0.38	14.85	**<0.001**
Valence (Moral)	0.17 (0.06)	0.11 (0.04)	0.05 – 0.30	0.03 – 0.18	2.68	**0.007**
Valence (Immoral)	0.02 (0.06)	0.01 (0.04)	-0.10 – 0.15	-0.06 – 0.09	0.37	0.710
Distance (Distant)	-0.03 (0.05)	-0.02 (0.03)	-0.13 – 0.08	-0.08 – 0.05	-0.52	0.600
Valence (Moral) X Distance (Distant)	0.04 (0.08)	0.03 (0.05)	-0.10 – 0.19	-0.06 – 0.12	0.58	0.561
Valence (Immoral) X Distance (Distant)	0.10 (0.07)	0.06 (0.05)	-0.05 – 0.25	-0.03 – 0.15	1.32	0.188
Random Effects						
σ^2	1.52					
$\tau_{00\ participant\ ID}$	0.62					
$\tau_{00\ Prosocial\ Category}$	0.62					
ICC	0.45					
$N_{Prosocial\ Category}$	5					
$N_{participant\ ID}$	5091					
Observations	25455					

Marginal R^2 / Conditional R^2	0.003 / 0.451

Table S6B. Outputs from linear mixed-effects model. We combined the composite measure of Volunteering intention and the Helping intentions of the four helping scenarios to have five measures of "Prosocial Intentions" per participant. (*Reference levels: Valence = Moral, Distance = Recent)*

Predictors	*Estimates (SE)*	*std. Beta*	*Estimates CI*	*standardized CI*	*Statistic*	*p*
(Intercept)	5.45 (0.36)	0.06 (0.22)	4.76 – 6.15	-0.37 – 0.49	15.34	**<0.001**
Valence (Neutral)	-0.17 (0.06)	-0.11 (0.04)	-0.30 – -0.05	-0.18 – -0.03	-2.68	**0.007**
Valence (Immoral)	-0.15 (0.06)	-0.09 (0.04)	-0.27 – -0.02	-0.17 – -0.01	-2.34	**0.019**
Distance (Distant)	0.02 (0.05)	0.01 (0.03)	-0.09 – 0.12	-0.05 – 0.07	0.30	0.766
Valence (Neutral) X Distance (Distant)	-0.04 (0.08)	-0.03 (0.05)	-0.19 – 0.10	-0.12 – 0.06	-0.58	0.561
Valence (Immoral) X Distance (Distant)	0.05 (0.07)	0.03 (0.05)	-0.09 – 0.20	-0.06 – 0.12	0.73	0.463
Random Effects						
σ^2	1.52					
$\tau_{00\ \text{Participant ID}}$	0.62					
$\tau_{00\ \text{Prosocial Category}}$	0.62					
ICC	0.45					
$N_{\text{Prosocial Category}}$	5					
$N_{\text{Participant ID}}$	5091					
Observations	25455					
Marginal R^2 / Conditional R^2	0.003 / 0.451					

8. ANCOVAs with the importance of moral identity internalization and symbolization as controls.

8.1 Willingness-to-volunteer: Direct outputs from jamovi

ANCOVA - Volunteering_Intention

	Sum of Squares	df	Mean Square	F	p	η^2p
Distance2	3.11	1	3.11	1.94	0.164	0.000
Valence	52.83	2	26.42	16.50	< .001	0.006
Moral_Internalization	1303.57	1	1303.57	813.98	< .001	0.138
Moral_Symbolization	71.54	1	71.54	44.67	< .001	0.009
Distance2 ✻ Valence	16.08	2	8.04	5.02	0.007	0.002
Residuals	8140.26	5083	1.60			

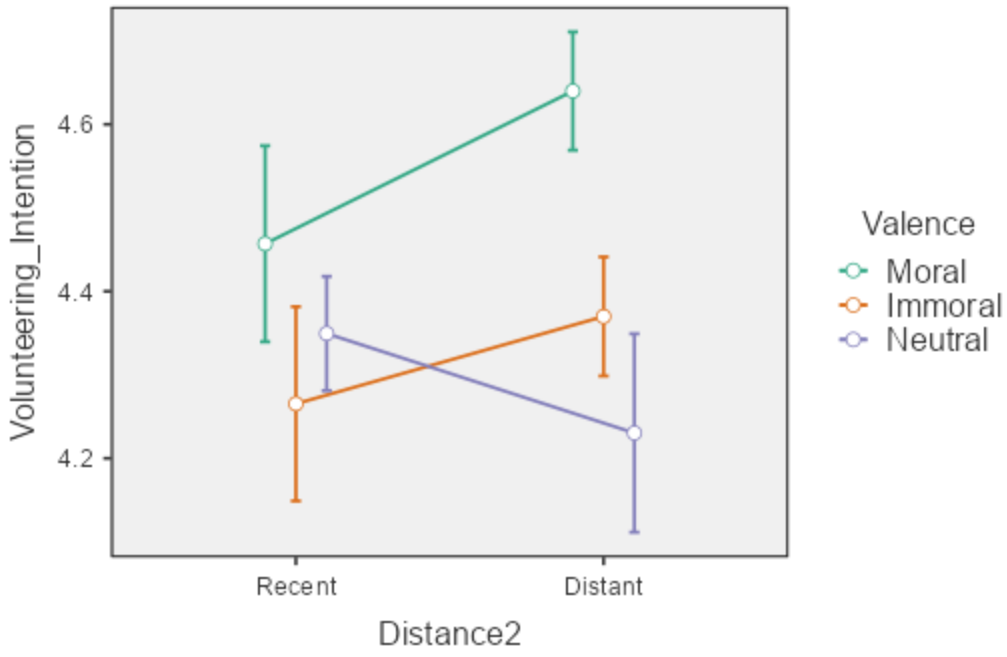
Post Hoc Comparisons - Distance2 ✻ Valence

Comparison									
Distance2	Valence		Distance2	Valence	Mean Difference	SE	df	t	p_{Tukey}
Recent	Moral	-	Recent	Immoral	0.1917	0.0843	5083	2.274	0.205
		-	Recent	Neutral	0.1075	0.0693	5083	1.550	0.632
		-	Distant	Moral	-0.1829	0.0699	5083	-2.615	0.094
		-	Distant	Immoral	0.0870	0.0701	5083	1.242	0.816
		-	Distant	Neutral	0.2266	0.0852	5083	2.660	0.084
	Immoral	-	Recent	Neutral	-0.0842	0.0689	5083	-1.223	0.826
		-	Distant	Moral	-0.3746	0.0695	5083	-5.391	< .001
		-	Distant	Immoral	-0.1046	0.0696	5083	-1.503	0.662
		-	Distant	Neutral	0.0349	0.0848	5083	0.412	0.998
	Neutral	-	Distant	Moral	-0.2904	0.0502	5083	-5.780	< .001
		-	Distant	Immoral	-0.0204	0.0504	5083	-0.406	0.999
		-	Distant	Neutral	0.1191	0.0699	5083	1.703	0.530
Distant	Moral	-	Distant	Immoral	0.2700	0.0513	5083	5.265	< .001
		-	Distant	Neutral	0.4095	0.0706	5083	5.803	< .001
	Immoral	-	Distant	Neutral	0.1396	0.0707	5083	1.975	0.357

Note. Comparisons are based on estimated marginal means

Estimated Marginal Means

Distance2 * Valence



8.2 Willingness-to-help: Direct outputs from jamovi

ANCOVA - Helping_Intention

	Sum of Squares	df	Mean Square	F	p	η^2p
Distance2	1.572	1	1.572	1.735	0.188	0.000
Valence	17.373	2	8.686	9.587	< .001	0.004
Moral_Internalization	612.489	1	612.489	676.034	< .001	0.117
Moral_Symbolization	8.139	1	8.139	8.983	0.003	0.002
Distance2 ✻ Valence	0.810	2	0.405	0.447	0.639	0.000
Residuals	4605.211	5083	0.906			

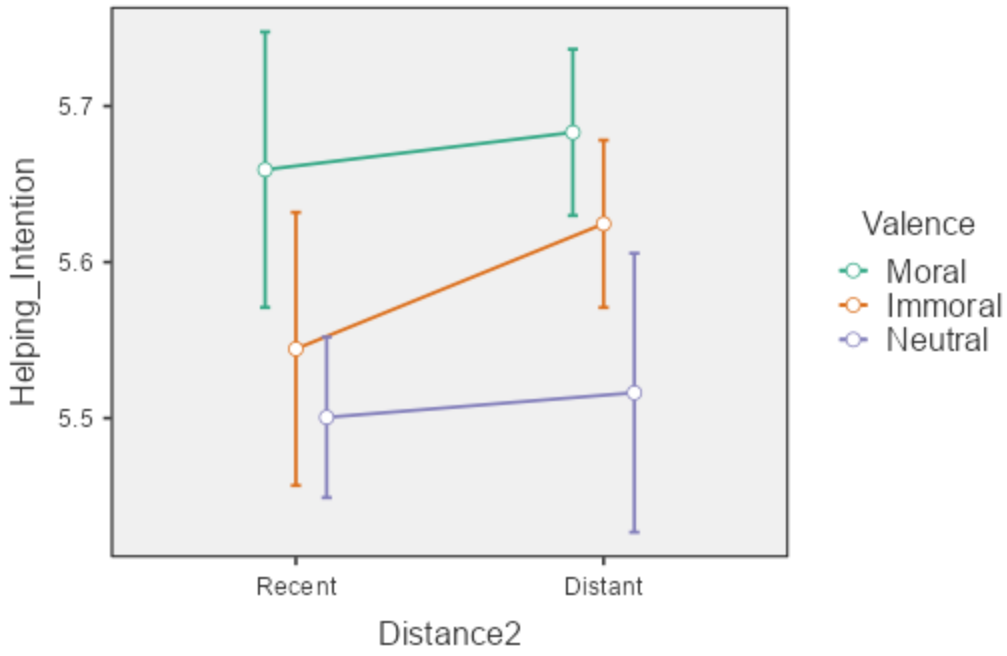
Post Hoc Comparisons - Distance2 ✻ Valence

Comparison									
Distance2	Valence		Distance2	Valence	Mean Difference	SE	df	t	p_{Tukey}
Recent	Moral	-	Recent	Immoral	0.1149	0.0634	5083	1.812	0.458
		-	Recent	Neutral	0.1587	0.0521	5083	3.043	0.028
		-	Distant	Moral	-0.0239	0.0526	5083	-0.454	0.998
		-	Distant	Immoral	0.0347	0.0527	5083	0.659	0.986
		-	Distant	Neutral	0.1429	0.0641	5083	2.231	0.224
	Immoral	-	Recent	Neutral	0.0438	0.0518	5083	0.847	0.959
		-	Distant	Moral	-0.1388	0.0523	5083	-2.655	0.085
		-	Distant	Immoral	-0.0802	0.0524	5083	-1.531	0.644
		-	Distant	Neutral	0.0281	0.0638	5083	0.440	0.998
	Neutral	-	Distant	Moral	-0.1826	0.0378	5083	-4.832	< .001
		-	Distant	Immoral	-0.1240	0.0379	5083	-3.271	0.014
		-	Distant	Neutral	-0.0158	0.0526	5083	-0.300	1.000
Distant	Moral	-	Distant	Immoral	0.0586	0.0386	5083	1.519	0.652
		-	Distant	Neutral	0.1668	0.0531	5083	3.143	0.021
	Immoral	-	Distant	Neutral	0.1082	0.0532	5083	2.036	0.322

Note. Comparisons are based on estimated marginal means

Estimated Marginal Means

Distance2 ✻ Valence


Helping_Intention
5.7
5.6
5.5
Recent
Distant
Distance2
Valence
Moral
Immoral
Neutral

9. References

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

The jamovi project. (2022). *Jamovi.* (2.3) [Computer software]. https://www.jamovi.org

Qualtrics (Version 2022). (2022). Qualtrics. https://www.qualtrics.com

R Core Team. (2022). *R: A language and environment for statistical computing.* (4.1.1). R Foundation for Statistical Computing. https://www.r-project.org/