Most people do not "value the struggle":

Tempted agents are judged as less virtuous than those who were never tempted

Ryan M. McManus[1*]

Helen Padilla Fong[1]

Max Kleiman-Weiner[2]

Liane Young[1]

[1]Department of Psychology and Neuroscience, Boston College, Boston, MA, USA

[2]School of Engineering and Applied Sciences, Harvard University, Boston, MA, USA

*Corresponding author's e-mail:
ryan.m.mcmanus.phd@gmail.com

Author Contributions:

RM developed the research concept. HPF conducted the literature review under the supervision of RM. HPF constructed and modified stimuli and surveys under the supervision of RM and LY. RM, HPF, and LY developed the experimental designs for Studies 1-2, and MKW aided in developing the experimental design for Study 3. HPF collected data and conducted all analyses under the supervision of RM. RM drafted the manuscript, and HPF, MKW, and LY provided critical revisions. All authors approved the final version of the manuscript for submission.

Abstract

Do people judge those who overcome temptation as more virtuous than those who don't feel tempted in the first place? Because prior research provides conflicting answers to this question, the current paper uses an expanded set of methodological and statistical tools to solve this puzzle. First, we replicated results of prior research showing that agents who overcome temptation are seen as less virtuous than non-tempted agents, with 74-78% of people making this judgment (Study 1). Second, we used participant-generated stimuli and one measure from each of two published papers to rule out stimulus and measurement sampling as explanations for the previous opposite effects. We replicated Study 1's results: 72-75% of people judged agents who overcame temptation as less virtuous than non-tempted agents (Study 2). Third, we investigated whether judgments were moderated by relationship context. Again, the majority of people judged agents who overcame temptation–that would harm strangers or close others–as less virtuous than non-tempted agents. Additionally, the following interaction effect was the most common (modal) pattern: While judging tempted agents as less virtuous than non-tempted agents within each relationship context, 39% of people judged agents who were tempted to act in a way that would harm close others as even less virtuous than those agents whose temptations would harm strangers (Study 3). Together, these results provide a detailed moral psychological account of temptation by: resolving a puzzle in the literature, revealing moderation by relationship context, and documenting the pervasiveness of this effect across stimuli, measures, and persons.

Keywords: temptation, virtuosity, moral judgment, obligation, person-level

Introduction

Imagine the following scenario:

> *Gabriella and Katy live in different towns. While each of them are on a walk, they find a wallet on the ground that does not belong to them. The wallets contain cash as well as the owners' ID cards. They could each take the money for themselves, or they could each find the owners and give it all back.*
>
> *Gabriella feels very conflicted about this decision. She wants to steal the cash, and she is tempted to do so. However, even though she is tempted, she decides not to steal and she gives it all back to the owner.*
>
> *Katy does not feel conflicted about this decision. She does not want to steal the cash, and she is not tempted to do so. She gives it all back to the owner.*

Do you "value the struggle," judging Gabriella as more virtuous, as she was tempted to act immorally but ultimately overcame her temptation? Or do you consider her struggle as, in some way, revealing her negative moral character, and therefore judge Katy as more virtuous? These are versions of age-old questions within philosophy. On some philosophical accounts, the most virtuous agents are those who are naturally virtuous–virtue comes easily (e.g., Aristotle, 1985; Taylor, 1981), i.e., non-tempted agents should be seen as more virtuous. On other philosophical accounts, agents deserve credit for working hard to do what is virtuous, and to follow moral rules even and perhaps especially when it doesn't feel easy or natural (Halfon, 1989; Kant, 1998), i.e., agents who overcome temptation should be seen as more virtuous.

Descriptively, previous empirical research also provides opposing answers, with some suggesting that agents who overcome temptation are judged as more virtuous (Starmans & Bloom, 2016 [adult data]; Zhao & Kushnir, 2022 [adult data]), whereas other research suggests that non-tempted agents are judged as more virtuous (Berman & Small, 2018; Critcher, Inbar, & Pizarro, 2012). The current paper seeks to resolve this discrepancy and to improve the

methodological and statistical rigor of studying perceptions of temptation. Moreover, we attempt to provide a more complete picture of the role of temptation in moral psychology by investigating whether and how the social relationship between the tempted agent and their potential target affects these judgments (see Marshall et al., 2021; Marshall et al., 2022; McManus et al., 2020; McManus et al., 2021). For example, people may judge strangers' overcoming temptation to be virtuous, but they may not value the same struggle when it comes to their close friends, partners, or relatives. Intuitively, even if we value the struggle when deciding whether to return a wallet, we might not be equally appreciative of a person who struggles hard to stay faithful to their spouse even if they never cheat. Indeed, Jimmy Carter was widely condemned for his admission: "I've looked on a lot of women with lust. I've committed adultery in my heart many times.". If this is true, then agents who must overcome their temptation to engage in acts that would harm close others may be judged as especially non-virtuous.

In this paper, we first conduct direct replications of prior research showing that overcoming temptation is judged as less virtuous than never being tempted (Berman & Small, 2018; see Study 1). Second, because we initially replicated results of Berman & Small (2018) and therefore did not find evidence in favor of Starmans & Bloom (2016), we aimed to rule out methodological differences as reasons for prior research's discrepancy. Therefore, we expanded the stimulus set in a bottom-up way, using participant-generated (rather than researcher-generated) stimulus bases. Additionally, we used measures from both of the two conflicting publications (Berman & Small, 2018; Starmans & Bloom, 2016). Using participant-generated stimuli and both sets of measures allows us to investigate whether their discrepant results were due to stimulus and measurement sampling differences (see Study 2). Third, we investigate

whether judgments of temptation are more or less severe depending on who would be harmed if the temptation was acted on: close others versus strangers (Study 3). Across replications and extensions, we employ (compared to prior research) more appropriate analytic strategies to enable inferences about person-level—not group-level—moral psychology (i.e., cataloging how many participants show an experimental effect, see Allfield et al., 2016; Birnbaum, 1999; Donhauser et al., 2018; Grice, 2021; Grice et al., 2021; Hamaker, 2012; Ince et al., 2022; Ince et al., 2021; McManus et al., *in press*; Speelman & McGann, 2020).

Literature Review

Many studies have investigated how engaging in effortful moral behavior affects third-party moral character judgments. Some work suggests that people judge agents as more praiseworthy, or attribute a better moral character, when those agents exert effort to act morally. This research can be interpreted as consistent with the idea that agents who overcome temptation (i.e., an effortful task) are more virtuous than agents who never experience temptation. Bigman and Tamir (2016), for example, presented participants with two different situations. As an example of the first situation, consider a man taking a bus to work. While he is on the bus, a woman is about to get off and the man realizes that she dropped her wallet. He picks it up and gives it back to her. This situation is contrasted with one in which the woman gets off the bus so quickly that the man has to also get off the bus and run after her to return the wallet. Bigman & Tamir (2016) find that the man in the second situation (i.e., getting off the bus and running after the woman) is judged as more moral and more deserving of reward, effects driven by inferences of agents' differential goal importance to do the right thing. This effect was recently replicated by Berry & Lucas (2022).

Other research is broadly consistent with the idea that agents who overcome temptation are *less* virtuous than agents who never experience temptation. For example, Critcher, Inbar, & Pizarro (2012) investigated two instances that could be considered similar to those described above. Specifically, they compared a situation in which one agent quickly returned a lost wallet to a situation in which another agent returned a lost wallet after some deliberation. Since the latter condition could be viewed as requiring extra mental effort, some researchers (e.g., Berry & Lucas, 2022; Bigman & Tamir, 2016) might hypothesize that slowly returning the wallet would yield more positive judgments of moral character. However, Critcher et al. (2012) found that agents who slowly returned the wallet were judged as worse than agents who quickly returned the wallet. Critcher et al. (2012) argue that presence of deliberation in this context communicates that the agent at least has a predisposition to consider immoral behavior, and that the agent might eventually act immorally. The absence of deliberation, on the other hand, communicates that the agent may not have such a predisposition.

Additional research is more nuanced on the role of effort in moral judgment, providing evidence in favor of the idea that judgments of agents who experience temptation are sensitive to contextual features of the situation in which they experience conflict. For example, Everett et al. (2016) found that consequentialists who display mental conflict about making a consequentialist decision are judged more positively than those who display no mental conflict. On the other hand, deontologists who display mental conflict about making a deontological decision are judged less positively than those who display no mental conflict.

We focus here on a puzzle within research that has directly investigated the role of temptation in moral character judgments. Some prior work suggests that, while young children judge agents who overcome temptations as less morally good than agents who never experience

temptation, adults show the opposite pattern (Starmans & Bloom, 2016; Zhao & Kushnir, 2022). For example, after both agents promised their parents they would clean up their toys, adults judged the agent who was tempted to go outside and play with her friends (but ultimately cleaned up her toys) as more morally good than the agent who was not tempted to do so. However, other research suggests that adults judge agents who overcome temptation as *less* virtuous than agents who never experience temptation (Berman & Small, 2018). For example, an agent who overcame a temptation to cheat on his wife was judged as less virtuous than an agent who was not tempted to do so.

While these studies as a whole may seem at odds with one another, we note that they vary substantially in their methods, which makes it difficult to know whether differences reflect true differences. While Critcher et al. (2012), Berman & Small (2018), and Starmans & Bloom (2016) might be thought of as manipulating mental effort, the stimuli used by Critcher et al. and Berman & Small vary substantially in content from the stimuli used by Starmans & Bloom (2016). And, in Starmans & Bloom specifically, non-tempted agents are not simply described as lacking a desire to do the thing that tempted agents are to do (e.g., break a promise to clean up their toys), but the non-tempted agents are also described as actively disliking activities that could create a conflict or temptation in the first place (e.g., not only are they not tempted to play outside with their friends, but also they actively dislike playing outside with friends). Therefore, it is important to not treat each of these studies as investigating, in general, how mental effort, or perceived temptation, relates to moral judgment. Studies 1-2 of the current paper attempt to replicate, adjudicate between, and build on the results of studies investigating moral judgment as a function of perceived temptation. Study 2 in particular addresses stimulus sampling by not relying on previous researchers' stimuli and instead relying on participant-generated stimuli.

In an effort to provide a more theoretically complete picture of the role of temptation in moral psychology, we connect these questions to a burgeoning area of research suggesting that people believe others have special obligations to help and protect close others, obligations that they do not necessarily have to distant others such as strangers or acquaintances (Marshall et al., 2021; Marshall et al., 2022; McManus et al., 2020; McManus et al., 2021). Moreover, these beliefs have downstream consequences on moral judgment, where the violation of such relationship-oriented obligations leads to more negative moral judgments (e.g., Everett et al., 2016; Law et al., 2021). Therefore, consistent with Berman & Small (2018), when agents experience a temptation that could result in harm to a stranger, they might be judged as immoral when compared to non-tempted agents, but not nearly as immoral as agents who experience a temptation that could result in harm to close others. On the other hand, consistent with Starmans & Bloom (2016) and Zhao & Kushnir (2022), when agents experience a temptation that could result in a harm to close others, they may be judged as especially moral for overcoming an impulse that would violate their relationship-oriented obligation. As none of the prior research on effort and temptation explicitly investigated the role of social relationships, Study 3 of the current paper addresses these possibilities.

Rationale for General Methodological and Statistical Approaches

As discussed, extant papers make opposing claims about people's judgments of overcoming temptation (e.g., Berman & Small, 2018; Critcher et al., 2012; Starmans & Bloom, 2016; Zhao & Kushnir, 2022). We focus on two of these papers here, but note that our concerns apply in general. Starmans & Bloom (2016) find that adults judge agents who overcome immoral temptations as "more good" than agents who were never tempted in the first place. On the other hand, Berman & Small (2018) find that adults judge agents who never experience temptation as

"more virtuous" than agents who overcame immoral temptations. To solve this discrepancy, we advance two methodological improvements: investigating participant responses at various levels of analysis and addressing potential bias in stimulus/measurement sampling.

Level of Analysis (Group-Level vs Person-Level)

Recently, an issue has been identified with how psychological claims can be undermined given the typical statistical tests psychologists use. Specifically, it is possible for typical group-level tests (e.g., t-tests) to yield results that describe only a minority of participants. Importantly, in even slightly more complicated designs (e.g., with three-level factors or crossed factors), these tests can yield results that do not describe even a single participant's response pattern (McManus et al., *in press*). Therefore, this provides reason to investigate person-level responses rather than (only) group-level responses.

The level of analysis in prior research on temptation was not always one that provided an answer about person-level psychology (see Grice et al., 2020; McManus et al., *in press*; Moore et al., 2022; Speelman & McGann, 2020). Specifically, Berman & Small (2018) conducted group-level analyses (i.e., comparing sample means with t-tests) to make inferences. As has been argued elsewhere (e.g., Richters, 2021), the problem with such analyses is that there is a mismatch between psychological theorizing and the methods used for inference—typical theorizing occurs at the person-level, but when testing those theories, researchers switch to analytic procedures that operate at the group-level. This switch has the consequence of answering questions about nebulous parameters, such as population-level means, rather than the (more important) prevalence of a psychological effect, such as the number of persons whose responses match the theorized pattern. Starmans & Bloom's research (2016), on the other hand, did not have this problem, as their analysis (i.e., comparing the number of participants who made

particular judgments) enabled some inference about prevalence. To address this issue, we employ a variety of analytic techniques, varying from descriptive to inferential at both the person- and group-level to demonstrate equivalence or divergence across methods. Ultimately, person-level techniques provide finer-grained information that psychologists must care about if our goal is to build theories that are universalizable (e.g., how many participants match a theorized pattern).

Specifically, we employed five different analytic techniques across studies. First, as in prior research, we conducted typical group-level analyses to test predicted effects (e.g., t-tests against a null of 0). Second, we calculated the descriptive pervasiveness of the predicted effect (i.e., the proportion of participants' responses that match prediction; see Grice et al., 2020; Speelman & McGann, 2020). For example, if the group-level pattern suggests that agents who overcome temptation are judged as less virtuous than non-tempted agents, this does not necessarily mean that most people's judgments reflect this pattern (and it does not even mean that it is the modal judgment pattern, see Grice et al., 2020; McManus et al., *in press*). Person-level responses need to be investigated and tabulated to understand the pattern's pervasiveness. Third, when possible (i.e., for the current paper's interaction effects), we conducted randomization tests to assess whether the predicted effect occurs in a proportion of the sample that is unlikely to have occurred via repeated random shuffling of the data (i.e., to rule out physical chance, see Grice, 2021). This method computes a chance- or c-value, providing a continuous measure of how likely or unlikely the predicted effect's proportion was to have emerged via chance (with values closer to 0 suggesting that the proportion was unlikely to emerge as a function of participants' randomly responding). Fourth, we conducted frequentist prevalence testing to assess whether the predicted effect occurs in a proportion of the sample

equal to or greater than a theoretical value of interest, allowing a null hypothesis significance testing inference about population prevalence (see Allfield et al., 2016; Donhauser et al., 2018). Here, we test against a majority null (i.e., > 0.50, that at least half of persons in the population are likely to show our predicted effects), as recent evidence suggests that most laypeople and social psychology researchers believe this to be the bare minimum for establishing evidence in favor of a psychological model or theory (McManus et al., *in press*). This method allows a sample-to-population inference, whereas the two previously described approaches (i.e., descriptive pervasiveness and randomization tests) do not allow such an inference. Fifth and finally, we conducted Bayesian prevalence estimation to obtain a posterior distribution, which not only provides information about the most likely population prevalence value, but also enables calculation of probabilities that the population prevalence value is equal to or greater than a theoretical value of interest (e.g., 0.50, see Ince et al., 2022; Ince et al., 2021)[1]. Here, we compute probabilities for the majority null ($Pr(\gamma > 0.50)$, and the global null ($Pr(\gamma > 0)$, i.e., that the predicted effect occurs in at least *some* subset of the population). This method also allows a sample-to-population inference.

Stimulus and Measurement Sampling

In both Berman & Small (2018) and Starmans & Bloom (2016), an additional concern is stimulus and measurement sampling. As has been argued elsewhere (e.g., Yarkoni, 2020), researchers usually intend to generalize over not just participants, but over other features of experiments, such as instruction sets, measures, and stimuli. A problem with many judgment paradigms is that the researchers generate (sometimes only one or a few) stimuli and measures, which can sometimes (either incidentally or intentionally) lead to effects for *only* those stimuli or measures. To address this issue in Studies 2-3 of the current paper, rather than relying on

published stimulus sets from either of these publications, we ask an independent sample of participants to generate scenarios in which people are tempted but ultimately overcome immoral desires, leading to a larger, more diverse, and potentially more generalizable stimulus set than those in prior research. We then visually represent stimulus variability to catalog the generality (or lack thereof) of effects across stimuli. Additionally, we use one measure from each publication to test whether differences in measurement explain the discrepant results between these papers.

Open Practices

All studies in this paper were pre-registered via AsPredicted: https://aspredicted.org/M94_2Q4, https://aspredicted.org/S5X_DZL, https://aspredicted.org/KZF_1G9, and https://aspredicted.org/3B4_X4N. All materials, data, code, and analysis output are provided at https://osf.io/bym4t/?view_only=ab62a6698bb84e34a53892e39576623f. This includes stimuli and measures, raw data, organized and commented RMarkdowns, and RNotebook .html files with all visualizations and analysis outputs that can be compared to the results reported here. In these studies, we report all measures, manipulations, and exclusions.

Study 1

Study 1 had three purposes. First, we attempted to replicate Berman & Small's (2018) finding that agents who overcome temptation are judged as less virtuous than non-tempted agents. Second, we used all five of the previously described analytical techniques to gain precision in our understanding of the psychology claimed in Berman & Small (2018), allowing us to rule out differences in level of analysis as the source of discrepancy between previous research's findings. Third, we conducted two replications, each one with a different

crowdsourcing platform (i.e., CloudResearch and Prolific), to ensure generalizability across populations (Study 1a and Study 1b, respectively).

Method

Participants.

Two U.S. samples were collected (Study 1a and Study 1b). For Study 1a, participants (N = 174) were recruited via CloudResearch's approved participants list and compensated through Amazon's Mechanical Turk. For Study 1b, participants (N = 176) were recruited and compensated via Prolific. As stated in each pre-registration, participants were excluded if they failed a pre-task attention check that was disguised as a task-relevant stimulus, resulting in the following final Ns: Study 1a = 166 (Gender: 103 males, 62 females, 1 non-binary; Age: M = 39.92, SD = 11.95); Study 1b = 168 (Gender: 70 males, 94 females, 2 non-binary, 2 preferred not to disclose; Age: M = 39.39, SD = 13.39).

Design and Procedure.

Both studies used identical designs and procedures. Specifically, each study used a 2 (Domain: Non-Moral vs Moral) x 2 (Temptation: Non-Tempted vs Tempted) within-subjects design. After participants correctly answered an attention check, in random order, they were then presented with six vignettes taken verbatim directly from Berman & Small (2018)[2], three of which Berman & Small deemed as Non-Moral (e.g., cheating on a diet)[3], and three of which were deemed as Moral (e.g., cheating on a spouse). As in Berman & Small, the structure of each vignette mirrored the current paper's opening scenario, with participants being introduced to two agents who encounter the same situation. In each vignette, one of the agents was described as experiencing a temptation, whereas the other was described as having no such temptation. Ultimately, the two agents always made the same decision.

After reading each vignette, participants made three judgments on 9-point Likert-like scales that were averaged together into a global virtuosity measure (i.e., virtuosity, honorability, and respectability). These judgments were made on relative bipolar scales, meaning that the midpoint of each measure represented a judgment that the non-tempted and tempted agents were similarly virtuous, whereas any deflection from the midpoint represented a judgment that one of the agents was more virtuous than the other. Importantly, as was the case in Berman & Small (2018), each vignette specified that the agents had no relation to one another (i.e., lived in different cities, worked at different places, etc.) to ensure that participants understood that the agents' decisions were not influenced by one another.

Statistical Power.

Each final analyzable dataset (Study 1a $N = 166$, Study 1b $N = 168$) yielded more than 80% power to detect $d = 0.20$ for one-tailed one-sample t-tests, as well as more than 80% power to detect $dz = 0.20$ for one-tailed paired-samples t-tests (Faul et al., 2007)[4].

Hypotheses.

Per our pre-registrations (https://aspredicted.org/M94_2Q4 and https://aspredicted.org/S5X_DZL) and Berman & Small (2018), we had one simple hypothesis:

> In Moral vignettes, people will judge agents who overcome temptation as *less* virtuous than agents who are never tempted in the first place.

To prepare data for hypothesis-relevant analyses, we followed methods from Berman & Small (2018), averaging across each participant's multiple Moral vignettes. That is, to get a participant-level value for global virtuosity, we averaged across a participant's three judgments.

Results

We note here (and as can be verified in our pre-registrations) that we use one-tailed tests to obtain p-values for typical group-level tests throughout this paper, as we always had directional hypotheses. However, we also report two-tailed confidence intervals of each effect size estimate for readers who are interested in bidirectional uncertainty.

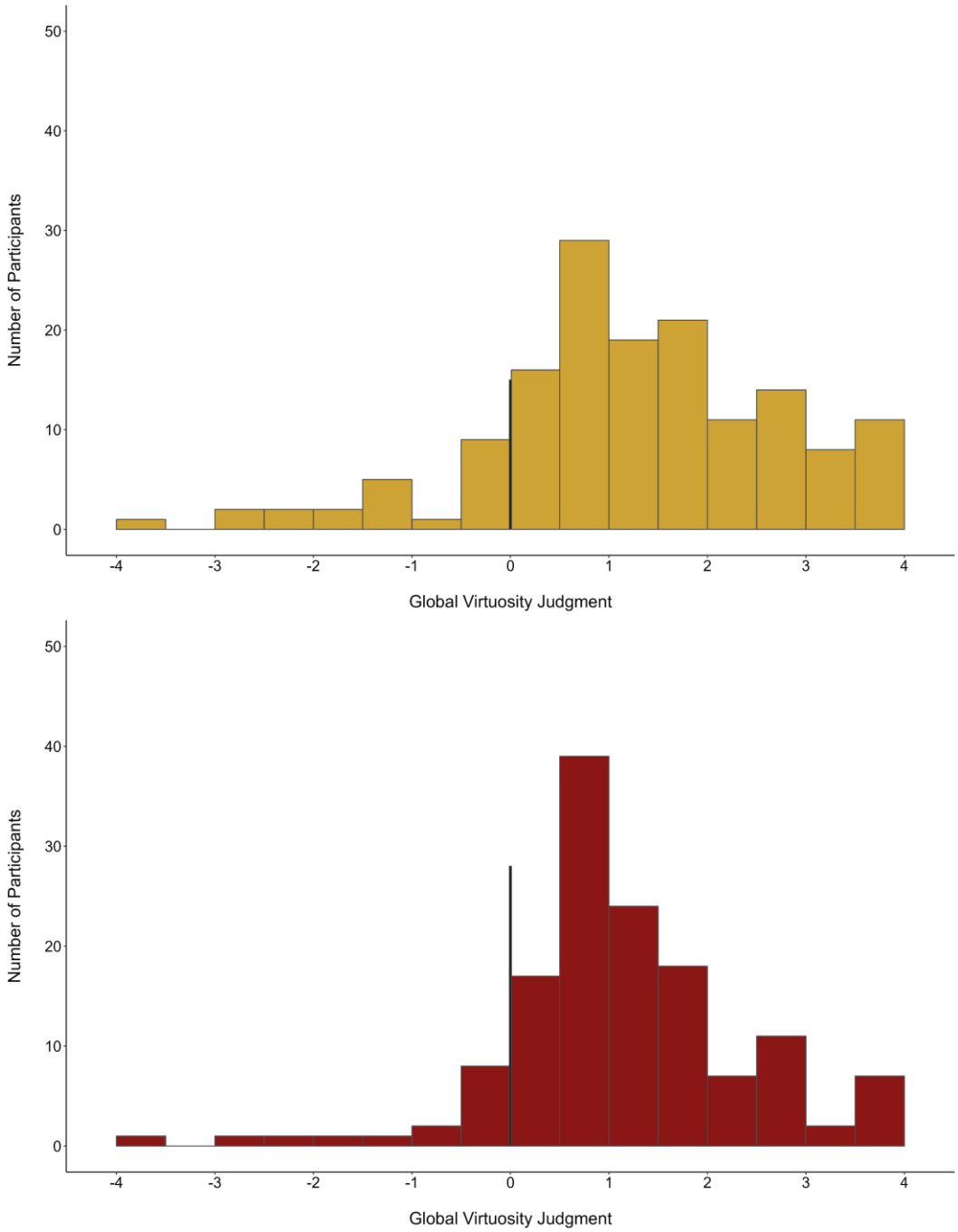

Figure 1. Histograms for global virtuosity judgments in moral contexts (top = Study 1a, bottom = Study 1b). Negative x-axis values correspond to judgments that tempted agents are judged as more virtuous than non-tempted agents, whereas positive x-axis values correspond to judgments that non-tempted agents are judged as more virtuous than tempted agents. Exact zeros (i.e., tempted and non-tempted agents are equivalent) are communicated by the thin black line centered at zero.

Typical Group-Level Tests.

Moral Vignettes. Non-tempted agents were judged as significantly more virtuous than agents who overcame temptation (Study 1a: M_{Diff} = 1.18, SD_{Diff} = 1.47), $t(165) = 10.28$, $p < .001$, $d = 0.80$ [95% CIs = 0.48, 1.12]; Study 1b: M_{Diff} = 0.96, SD_{Diff} = 1.20), $t(167) = 10.34$, $p < .001$, $d = 0.80$ [0.48, 1.11]).

Descriptive Pervasiveness.

Moral Vignettes. Most participants judged non-tempted agents as more virtuous than agents who overcame temptation (Study 1a = 78%; Study 1b = 74%)

Frequentist Prevalence Tests.

Moral Vignettes. In Study 1a, a majority of participants (78%) judged non-tempted agents as more virtuous than agents who overcame temptation, $p < .001$. Similarly, in Study 1b, a majority of participants (74%) judged non-tempted agents as more virtuous than agents who overcame temptation agents, $p < .001$. In both studies, we can reject the majority null and we therefore can decide to act as if most people in the population would make this judgment.

Bayesian Prevalence Estimation.

Moral Vignettes. In Study 1a, the most likely population prevalence value is estimated as 77% [96% Highest Posterior Density Intervals, or HPDIs = 69% - 83%]. Similarly, in Study 1b, the most likely population prevalence value is estimated as 73% [96% HPDIs = 65% - 80%]. Using the posterior distribution, we get the following probabilities for the majority null: Study 1a = 1.00; Study 1b = 1.00. Therefore, we also get similarly large probabilities for the global null: Study 1a = 1.00; Study 1b = 1.00.

Study 2

Results of Study 1 suggest that most people judged agents who overcame immoral temptations as less virtuous than non-tempted agents. This replicates the results of Berman & Small (2018), and rules out differences in level of analysis as a source of discrepancy between Berman & Small (2018) and Starmans & Bloom (2016). However, these findings are still at odds with results from Starmans & Bloom (2016, i.e., overcoming immoral temptation is more virtuous than never being immorally tempted). Study 2 therefore served the purpose of further resolving this discrepancy.

An additional explanation for the discrepancy is that different stimuli lead to different psychological effects (see Yarkoni, 2020). To address this, we conducted a pre-experiment study for the purposes of creating a new stimulus set. Because of our consistent replications of Berman & Small's (2018) effect (see Study 1), we asked participants to generate scenarios that would stack the deck in favor of Starmans & Bloom's (2016) findings. Specifically, participants were asked to generate scenarios in which they believed someone overcoming an immoral temptation would be more virtuous than someone who never had the immoral temptation in the first place. We used these scenarios to create 10 stimulus bases for Study 2. Therefore, in these new stimuli, if it is still the case that most people judge overcoming an immoral temptation as less virtuous than never being tempted, this provides strong evidence in favor of Berman & Small's claim that was replicated in Study 1, while providing equally strong evidence against Starmans & Bloom's claim[5].

Similar to the above stimulus sampling issue, different measures can also lead to different psychological effects. In Berman & Small (2018), participants assessed the relative "virtuosity" of two different agents on a 9-point Likert scale. Starmans & Bloom (2016), on the other hand, had participants assess which of the two agents was "more good" on a two-point binary scale. To

address this, rather than choosing our own new measure, or choosing one of previous research's measures, we randomly assign participants to one of prior research's measures. Specifically, in Study 2, half of our participants respond to the new stimuli with Berman & Small's (2018) continuous virtuosity measure, whereas the other half of participants respond using Starmans & Bloom's (2016) binary goodness measure. If the same psychological effect emerges for both measures (i.e., most people judge overcoming an immoral temptation as worse than non-temptation), this allows an inference of generalization across measures while simultaneously ruling out measurement differences as an explanation for previous research's discrepancies.

Method

Participants.

One U.S. sample (N = 361) was recruited via CloudResearch's approved participants list and compensated through Amazon's Mechanical Turk. Importantly, participants who completed Study 1, or our stimulus generation study, were unable to participate in Study 2. As stated in the pre-registration, participants were excluded if they failed a pre-task attention check that was disguised as a task-relevant stimulus, resulting in a final N = 350 (Gender: 191 males, 165 females, 2 non-binary, 1 preferred not to disclose; Age: M = 40.16, SD = 11.43).

Design and Procedure.

Study 2 used a similar design and procedure as Study 1, with three important differences. First, Study 2 used a 2 (Temptation: Non-Tempted vs Tempted) x 2 (Measure Type: Continuous Virtuosity vs Binary Goodness) design in which Temptation was manipulated within-subjects, whereas Measure Type was manipulated between-subjects. Second, unlike Study 1, participants in Study 2 were presented with *only* Moral vignettes (10 in total). These vignettes were created from stimulus bases generated by an independent set of participants who were instructed to come

up with situations in which an agent who overcomes an immoral temptation is more virtuous than an agent who is never tempted in the first place. Therefore, vignettes included in Study 2 were deemed moral not by an (additional) independent set of participants' ratings of moral relevance (as was the case in Berman & Small, 2018), but instead by instructing the stimulus-generating participants to create situations that they believed contained immoral temptations. We opted for this strategy to account for the possibility that different people will moralize different situations and therefore not all people will agree on the moral relevance of any one situation. As was the case in Study 1, the structure of each vignette mirrored the current paper's opening scenario, with participants being introduced to two agents encountering the same situation (one tempted and one not) who ultimately behave identically.

After reading each vignette, rather than making three judgments about global virtuosity, participants instead made single judgments. Because Study 1's virtuosity, honorability, and respectability judgments were highly correlated, we opted to use only the virtuosity measure by itself, and therefore half of participants made these judgments on a single 9-point Likert-like scale. These judgments were again made on a relative bipolar scale. In line with Starmans & Bloom's (2016) methods, the other half of participants made binary judgments indicating which of the two agents was "more good."

Statistical Power.

Each final analyzable dataset (Binary Goodness $N = 177$, Continuous Virtuosity $N = 177$) yielded more than 80% power to detect $d = 0.20$ for one-tailed one-sample t-tests (Faul et al., 2007).

Hypotheses.

Per our pre-registration (https://aspredicted.org/KZF_1G9), we had one simple hypothesis:

People will judge agents who overcome immoral temptation as *less* virtuous than agents who are never tempted in the first place.

To prepare data for hypothesis-relevant analyses, we again averaged across each participant's multiple vignettes. That is, to get a participant-level value for virtuosity or goodness, we averaged across a participant's 10 judgments.

Results

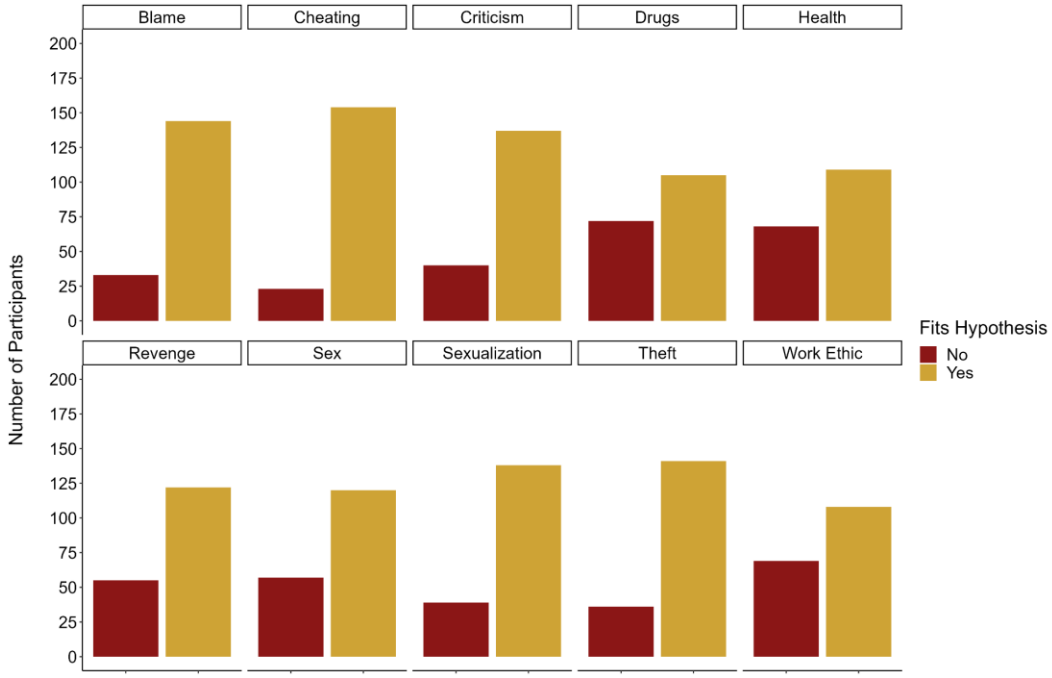

Figure 2. Participants whose responses matched hypothesized pattern for binary judgments.

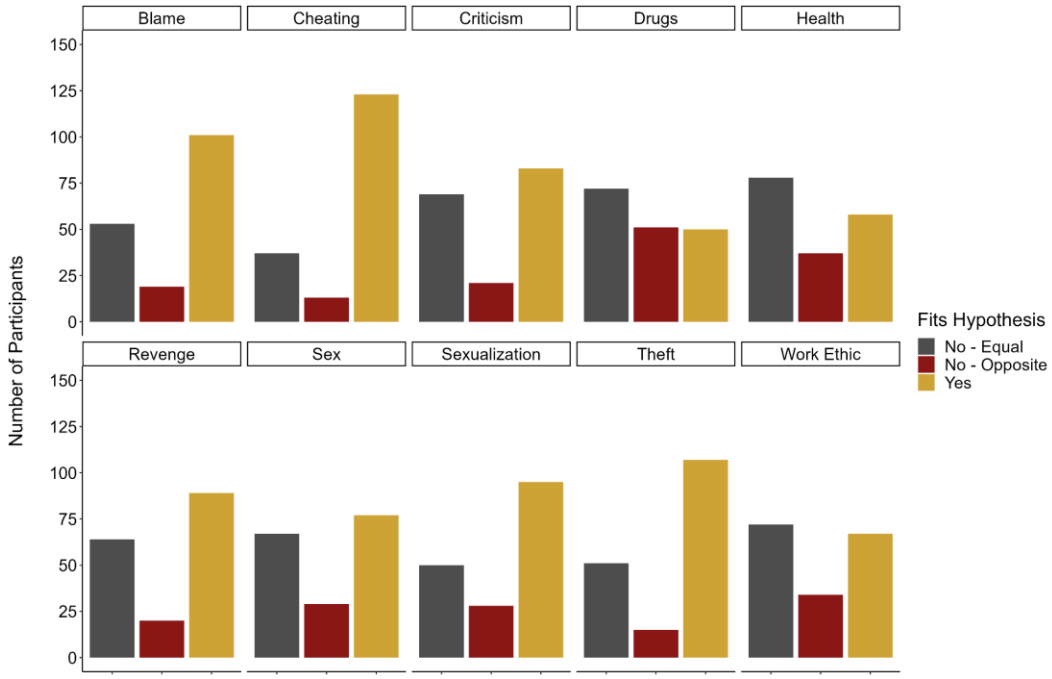

Figure 3. Participants whose responses matched hypothesized pattern for continuous judgments. Although some stimuli show non-majority effects, it was never the case that there was a majority effect in the opposite direction. That is, each stimulus' modal pattern was either the predicted pattern or no difference between non-tempted and tempted agents (i.e., "equal").

Typical Group-Level Tests.

Binary Goodness. For binary judgments, the null value tested against was 0.50, as this is the value that would indicate that any one participant chose the non-tempted and tempted agent at similar rates across vignettes. Therefore, a value larger than 0.50 indicates that participants chose the non-tempted agent more often than the tempted agent. Non-tempted agents were judged as significantly more good than agents who overcame temptation (M_{Diff} = 0.72, SD_{Diff} = 0.28), $t(176) = 10.47$, $p < .001$, $d = 0.79$ [0.48, 1.09].

Continuous Virtuosity. For continuous judgments, the null value tested against was 0, as is typical for one-sample t-tests (as we did in Study 1). Non-tempted agents were again judged as significantly more virtuous than agents who overcame temptation (M_{Diff} = 0.92, SD_{Diff} = 1.22), $t(172) = 9.94$, $p < .001$, $d = 0.76$ [0.44, 1.07].

Descriptive Pervasiveness.

Binary Goodness. When making binary judgments, 72% of participants judged non-tempted agents as more virtuous than agents who overcame temptation.

Continuous Virtuosity. When making continuous judgments, 75% of participants judged non-tempted agents as more virtuous than agents who overcame temptation.

Frequentist Prevalence Tests.

Binary Goodness. When making binary judgments, a majority of participants (72%) judged non-tempted agents as more virtuous than agents who overcame temptation, $p < .001$.

Continuous Virtuosity. When making continuous judgments, a majority of participants (75%) judged non-tempted agents as more virtuous than agents who overcame temptation, $p < .001$.

Bayesian Prevalence Estimation.

Binary Goodness. For binary judgments, the most likely population prevalence value is estimated as 71% [96% HPDIs = 63% - 78%]. Using the posterior distribution, we get the following probability for the majority null: $\Pr(\gamma > 0.50) = 1.00$. Therefore, we also get a similarly large probability for the global null: $\Pr(\gamma > 0) = 1.00$.

Continuous Virtuosity. For continuous judgments, the most likely population prevalence value is estimated as 74% [96% HPDIs = 66% - 80%]. Using the posterior distribution, we get the following probability for the majority null: $\Pr(\gamma > 0.50) = 1.00$. Therefore, we also get a similarly large probability for the global null: $\Pr(\gamma > 0) = 1.00$.

Interim Discussion

Results of Study 2 suggest that, across new stimuli and two measures, most people indeed judge overcoming an immoral temptation as less virtuous than never being tempted. This

suggests that measurement sampling is an unlikely explanation of the discrepant results of previous research. Additionally, our results suggest that Starmans & Bloom's (2016) stimuli may be outliers in the stimulus sampling space, and that perhaps our observed effect is one that will be most typical across new stimuli. We consider this to be especially likely due to 1) our use of participant-generated stimuli that were intended to yield results consistent with those in Starmans & Bloom (2016), and 2) not a single one of these stimuli showed an effect consistent with theirs (see Figures 2-3). Moreover, another potential discrepancy can be ruled out given our use of stimuli that vary in content. Specifically, Starmans & Bloom's stimuli might be considered low-stakes (e.g., breaking a promise to clean up toys), whereas Berman & Small's stimuli might be considered higher-stakes (e.g., cheating on a spouse), which could explain the opposing effects. However, our Study 2's stimuli varied in severity (e.g., cheating on a spouse vs criticizing someone's outfit), and, across these stimuli, we still found the effect predicted by Berman & Small (2018).

Study 3

Study 3 had the goal of determining whether an often overlooked factor in moral psychology, namely social relationship information, affects how people tend to judge overcoming immoral temptations. Perhaps people are especially unlikely to "value the struggle" when it is their close others who must do so, as this suggests that their close others have an impulse to violate special obligations to close others (Marshall et al., 2022; Marshall et al., 2021; McManus et al., 2020; McManus et al., 2021). If this is true, then people may also be especially harsh in their moral judgments when others are overcoming temptations to harm close others, a question that Study 3 attempts to answer.

To test this, we again conducted a pre-experiment study for the purposes of creating a new stimulus set, asking participants to generate scenarios that would stack the deck in favor of Starmans & Bloom's (2016) findings. This time, however, we specifically asked participants to generate scenarios in which temptations could harm close others and strangers; we did this to ensure that if we found differences as a function of relationship context, the differences could not be attributed to fundamental differences in content between stimuli. For example, being tempted to cheat on a spouse, by its nature, can only affect close others (not strangers); therefore, we did not use this kind of stimulus in Study 3. Along with some of Study 2's stimuli, we used this new set of participant-generated scenarios to create 20 stimulus bases for Study 3 that could be manipulated to be about immoral temptations affecting strangers or close others (i.e., close friends and siblings). Measurement-wise, after demonstrating generalization across measures (Study 2), we opted to use only Berman & Small's continuous virtuosity measure for Study 3. We chose this measure because it allows participants to make a judgment that tempted and non-tempted agents are similarly virtuous, whereas Starmans & Bloom's binary goodness measure forces participants to choose one agent over the other.

Method

Participants.

One U.S. sample (N = 300) was recruited via CloudResearch's approved participants list and compensated through Amazon's Mechanical Turk. Importantly, participants who completed any of the previous studies were unable to participate in Study 3. As stated in the pre-registration, participants were excluded if they failed a pre-task attention check that was disguised as a task-relevant stimulus, resulting in a final N = 225 (Gender: 119 males, 104 females, 2 preferred not to disclose; Age: M = 40.12, SD = 11.24).

Design and Procedure.

Study 3 used a 2 (Temptation: Non-Tempted vs Tempted) x 2 (Relationship Context: Stranger vs Close Other) design in which both factors were manipulated within-subjects. Here, the Relationship Context factor varied whether giving in to the temptation would have negative consequences for the agents' close friends or relatives versus complete strangers (see Table 1 for an example stimulus with all experimental variants). Unlike previous studies, participants in Study 3 were presented with many more vignettes (20 in total). These vignettes were again created from stimulus bases generated by an independent set of participants who were instructed to come up with situations in which an agent who overcomes an immoral temptation is more virtuous than an agent who is never tempted in the first place. As was the case in previous studies, the structure of each vignette mirrored the current paper's opening scenario, with participants being introduced to two agents who encounter the same situation and ultimately behave identically. Importantly, participants did not see the same vignette across the levels of Relationship Context; specifically, half of participants saw 10 vignettes for the Stranger level and the other 10 vignettes for the Close Other level, whereas the other half of participants saw the opposite list. Which vignette was assigned to which list was accomplished by using a random number generator before creating the study's Qualtrics survey. After reading each vignette, all participants made single judgments of virtuosity on a single 9-point Likert-like scale. These judgments were again made on a relative bipolar scale.

Table 1. Example stimulus ("Lying") and its experimental variants for Study 3

Stimulus Base	Sara and Chloe are flying on different planes. They are both physicians. One of the passengers on each plane begins to pass out. The flight attendant asks if anybody on board has medical experience to aid the passenger in distress. Sara and Chloe each have the option to lie to avoid the commitment, or they can assist the passenger in need.
Non-Tempted Agent	**Sara** and the passenger are (strangers / close friends). Sara feels very conflicted about this decision. She wants to lie to avoid the commitment of helping the passenger, and she is tempted to do so. However, even though she is tempted, she decides to assist the passenger.
Tempted Agent	**Chloe** and the passenger are (strangers / close friends). Chloe does not feel conflicted about this decision. She does not want to lie to avoid the commitment of helping the passenger, and she is not tempted to do so. She decides to assist the passenger.

Note: All stimuli were presented with the stimulus base, a non-tempted agent, and a tempted agent. Terms inside of parentheses (e.g., strangers / close friends) varied across Stranger and Close Other vignettes (respectively). Therefore, participants always saw the same stimulus when comparing non-tempted to tempted agents, but they saw different stimuli for Stranger versus Close Other vignettes.

Statistical Power.

The final analyzable dataset (N = 225) yielded more than 90% power to detect d = 0.20 for one-tailed one-sample t-tests, as well as more than 90% power to detect dz = 0.20 for one-tailed paired-samples t-tests (Faul et al., 2007).

Hypotheses.

Per our pre-registration (https://aspredicted.org/3B4_X4N), we had two simple hypotheses and one complex hypothesis:

1) For Stranger vignettes, people will judge agents who overcome immoral temptation as less virtuous than agents who are never tempted in the first place.
2) Similarly, for Close Other vignettes, people will judge agents who overcome immoral temptation as less virtuous than agents who are never tempted in the first place.
3) In addition to the two above hypotheses, people's virtuosity judgments for Stranger vignettes will be less extreme than their judgments for Close Other vignettes, meaning that they will judge being tempted to harm close others as even less virtuous than being tempted to harm strangers (i.e., the interaction hypothesis).

To prepare data for hypothesis-relevant analyses, we again averaged across each participant's multiple vignettes of the same kind. That is, to get a participant-level value for virtuosity in Stranger vignettes, we averaged across a participant's 10 Stranger judgments. Similarly, to get a participant-level value for virtuosity in Close Other vignettes, we averaged across a participant's 10 Close Other judgments.

Results

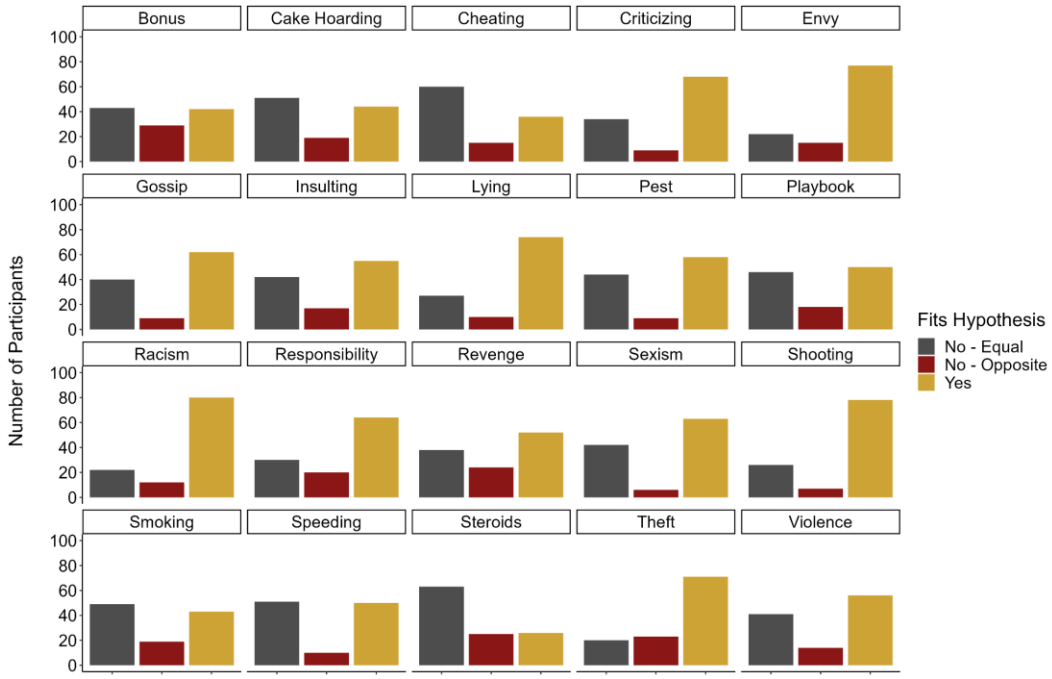

Figure 4. Participants whose responses matched the hypothesized pattern for Stranger vignettes. Although some stimuli show non-majority effects, it was never the case that there was a majority effect in the opposite direction. That is, each stimulus' modal pattern was either the predicted pattern or no difference between non-tempted and tempted agents (i.e., "equal").

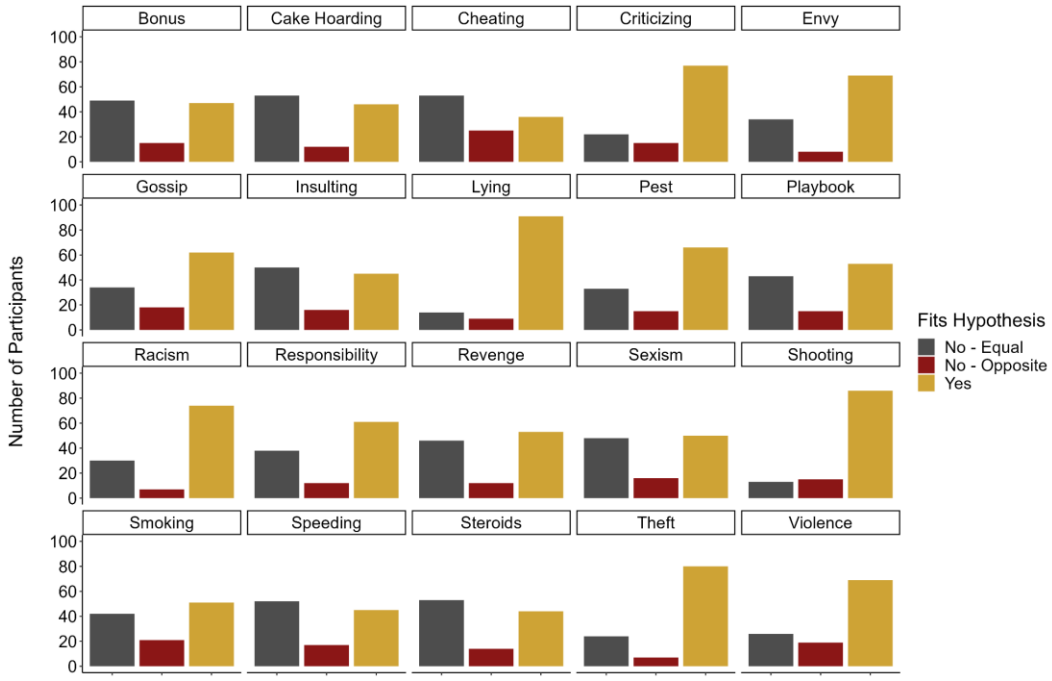

Figure 5. Participants whose responses matched the hypothesized pattern for Close Other vignettes. Although some stimuli show non-majority effects, it was never the case that there was

a majority effect in the opposite direction. That is, each stimulus' modal pattern was either the predicted pattern or no difference between non-tempted and tempted agents (i.e., "equal").

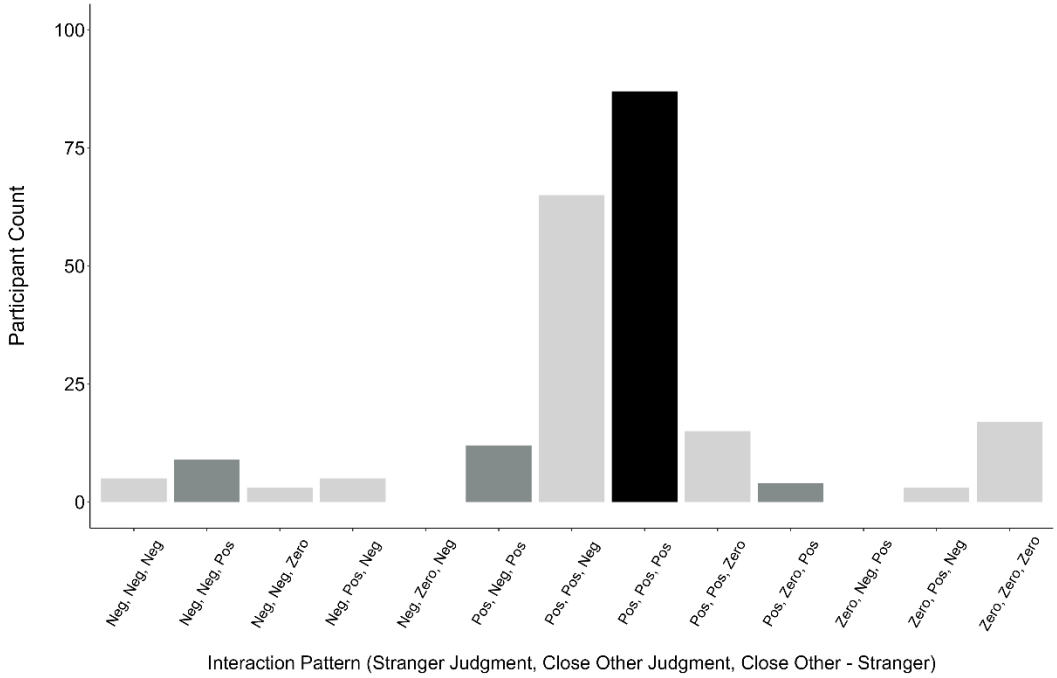

Figure 6. Possible and empirical interaction patterns in Study 3. Pattern descriptions (e.g., Pos, Neg, Pos) communicate the Stranger judgment, Close Other judgment, and Interaction (Close Other - Stranger), respectively. The black bar represents the hypothesized pattern, whereas dark grey bars represent patterns also yielding an interaction value consistent with the hypothesized pattern (i.e., contributed to the group-level interaction's emergence).

Typical Group-Level Tests.

Stranger Vignettes. Non-tempted agents were judged as significantly more virtuous than agents who overcame temptation (M_{Diff} = 0.98, SD_{Diff} = 1.20), $t(224) = 12.18$, $p < .001$, $d = 0.81$ [0.54, 1.09].

Close Other Vignettes. Non-tempted agents were judged as significantly more virtuous than agents who overcame temptation (M_{Diff} = 1.08, SD_{Diff} = 1.15), $t(224) = 13.98$, $p < .001$, $d = 0.93$ [0.66, 1.21].

Interaction. Virtuosity judgments were significantly different when comparing Stranger vignettes to Close Other vignettes (M_{Diff} = 0.10, SD_{Diff} = 0.65), $t(224) = 2.33$, $p = .010$, $d_z = 0.16$

[0.02, 0.29], d_{av} = 0.09 [0.01, 0.16], suggesting that non-tempted agents were judged more positively in Close Other vignettes. Considering our use of relative virtuosity scales, this confirms our interaction hypothesis: While non-tempted agents were judged as more virtuous than agents who overcame temptation across relationship contexts, being tempted to engage in a behavior that could harm close others was judged as even less virtuous than being tempted to engage in a behavior that could harm strangers.

Descriptive Pervasiveness.

Stranger Vignettes. 78% of participants judged non-tempted agents as more virtuous than agents who overcame temptation.

Close Other Vignettes. 81% of participants judged non-tempted agents as more virtuous than agents who overcame temptation.

Interaction. Only 39% of participants simultaneously judged non-tempted agents as more virtuous than tempted agents within both Stranger and Close Other vignettes while also having more extreme judgments for Close Other vignettes.

Randomization Test.

For Study 3's interaction, 1000 random shufflings led to 0(!) datasets yielding a descriptive pervasiveness percentage equal to or greater than the observed descriptive pervasiveness (c-value = 0), suggesting that the original 39% estimate is extremely *un*likely to have occurred via physical chance. This means that the observed descriptive pervasiveness percentage is indeed distinguishable from the possibility that participants were randomly selecting virtuosity values in each condition.

Frequentist Prevalence Tests.

Stranger Vignettes. A majority of participants (78%) judged non-tempted agents as more virtuous than agents who overcame temptation, $p < .001$.

Close Other Vignettes. A majority of participants (81%) judged non-tempted agents as more virtuous than agents who overcame temptation, $p < .001$.

Interaction. A minority (only 39%) of participants simultaneously judged non-tempted agents as more virtuous than tempted agents across Stranger and Close Other vignettes while also having more extreme judgments for Close Other vignettes, $p = .999$.

Bayesian Prevalence Estimation.

Stranger Vignettes. The most likely population prevalence value is estimated as 77% [96% HPDIs = 70% - 82%]. Using the posterior distribution, we get the following probability for the majority null: $\Pr(\gamma > 0.50) = 1.00$. Therefore, we also get a similarly large probability for the global null: $\Pr(\gamma > 0) = 1.00$.

Close Other Vignettes. The most likely population prevalence value is estimated as 80% [96% HPDIs = 74% - 86%]. Using the posterior distribution, we get the following probability for the majority null: $\Pr(\gamma > 0.50) = 1.00$. Therefore, we also get a similarly large probability for the global null: $\Pr(\gamma > 0) = 1.00$.

Interaction. The most likely population prevalence value for the predicted interaction pattern is estimated as 35% [96% HPDIs = 29% - 43%]. Using the posterior distribution, we get the following probability for the majority null: $\Pr(\gamma > 0.50) < .001$. However, we get a much larger probability for the global null: $\Pr(\gamma > 0) = 1.00$.

General Discussion

The current research aimed to replicate and extend, both methodologically and theoretically, previous research on how perceived temptation shapes third-party moral character

judgment. Across studies, people were asked to evaluate two agents simultaneously, one who was tempted but ultimately overcame it, and one who was never tempted in the first place. Specifically, Study 1 attempted to replicate recent research showing that agents who overcome immoral temptations are judged as less virtuous than non-tempted agents. Study 2 had the primary purpose of solving discrepancies between two prior publications that found opposite-signed effects (Berman & Small, 2018; Starmans & Bloom, 2016). Study 3 examined whether social relationship information moderated effects found in Studies 1 and 2, testing whether people's judgments change as a function of agents being tempted to harm strangers versus close others (i.e., close friends or siblings). See Table 2 for a summary of analyses across studies.

Table 2. Analysis summary across studies

		Analysis				
		Group Level	**Descriptive Pervasiveness**	**Randomization Test**	**Frequentist Prevalence**	**Bayesian Prevalence**
Study	**Effect**					
Study 1a						
	Moral (NT > T)	✓	✓	-	✓	✓
Study 1b						
	Moral (NT > T)	✓	✓	-	✓	✓
Study 2						
	Binary (NT > T)	✓	✓	-	✓	✓
	Continuous (NT > T)	✓	✓	-	✓	✓
Study 3						
	Stranger (NT > T)	✓	✓	-	✓	✓
	Close Other (NT > T)	✓	✓	-	✓	✓
	Interaction (NT > T in S) and (NT > T in CO) and (CO > S)	✓	☐	✓	☐	☐

Note: A green check communicates that the pre-registered effect occurred, whereas a red X communicates that the pre-registered effect did not occur. Pre-registered effects are specified in parentheses with the following labels: NT = Non-Tempted agent, T = Tempted agent, S = Stranger, CO = Closer Other. The "Group-Level" column shows whether predicted effects occurred using typical group-level tests (e.g., t-tests). The "Descriptive Pervasiveness," "Frequentist Prevalence," and "Bayesian Prevalence" columns show whether predicted effects occurred in a majority of the sample or can be expected to occur in the majority of the population. The "Randomization Test" column shows whether the pre-registered interaction effect occurred in a percentage of the sample that was unlikely to occur via repeated random shuffling of the data. Due to the nature of our experimental designs (i.e., using relative scales to measure differential goodness or virtuosity), randomization tests could only be conducted for our interaction predictions; that is, shuffling of the simple effect data on their own (i.e., a single column of judgments) would continually result in the same descriptive pervasiveness percentage.

Study 1 successfully replicated Berman & Small's (2018) finding across group-level and person-level analyses: Agents who overcame immoral temptations were judged as less virtuous than agents who never experienced temptation. Indeed, 74-78% of people made this judgment, suggesting that it is likely to be a general psychological regularity (see Figure 1). In Study 2, we attempted to resolve a discrepancy between two prior papers (Berman & Small, 2018; Starmans & Bloom, 2016). Berman & Small (2018), and our Study 1, found that agents who overcome immoral temptations are judged as less virtuous than agents who never experience temptation in the first place, whereas Starmans & Bloom (2016) found the opposite pattern. In Study 2, we considered two potential explanations for this discrepancy: stimulus and measurement sampling. Across a new set of participant-generated stimuli, and across two measures, people again judged agents who overcame immoral temptations as less virtuous than non-tempted agents. Indeed, 72-75% of people made this judgment, again suggesting it is a general psychological regularity; we also note the incredible precision and consistency in percentages when compared to Study 1's percentage range of 74-78%. When investigating stimulus-level variability, across measures, there was never a stimulus for which most people made judgments in the opposite direction (and there was never a stimulus in which the opposite pattern was even the modal pattern; see Figures 2-3). All together, Study 2's results are inconsistent with the pattern of results obtained in Starmans & Bloom (2016), suggesting that their stimuli may be outliers in the stimulus sampling space.

In Study 3, we attempted to extend the findings of Studies 1-2 by relying on a recently burgeoning subset of moral psychology research: the impact of close relationships (e.g., Everett et al., 2016; Law et al., 2021; Marshall et al., 2022; Marshall et al., 2021; McManus et al., 2020; McManus et al., 2021). We hypothesized that people's judgments should be moderated when

considering whether an agent's giving in to a temptation would negatively affect a stranger versus a close other (i.e., close friend or sibling). When investigating within-relationship judgments, we again found, across group-level and person-level analyses, that people judge overcoming immoral temptations as less virtuous than never being tempted. Indeed, 78% of people made this judgment for situations involving strangers, and 81% of people made this judgment for situations involving close others. As was true in Study 2, when investigating stimulus-level variability in Study 3, there was never a stimulus in which most people made judgments in the opposite direction (and there again was never a stimulus in which the opposite pattern was even the modal pattern; see Figures 4-5). Moreover, Study 3 used stimuli that varied more in content than Study 2, allowing a stronger test of the possibility that differences in stimulus severity might explain discrepancies between Starmans & Bloom and Berman & Small's effects. For example, Study 3 used extreme scenarios such as being tempted to shoot someone (likely killing them) and mundane scenarios such as being tempted to take multiple slices of cake at a wedding (such that other guests might not get cake). Across these stimuli, Starmans & Bloom's (2016) pattern never emerged as even the modal pattern.

When considering Study 3's interaction hypothesis (i.e., in addition to judging agents who overcame temptation as less virtuous within each relationship context, judgments would be more extreme for temptations to harm close others), group-level analyses showed the predicted effect. However, person-level analyses showed that only 39% of people indeed made such judgments (see Figure 6). Importantly, though, a randomization test suggested that this percentage was distinguishable from random responding, providing evidence in favor of the robustness of the pattern. While this means that it may not be a general psychological regularity in the majority sense (capturing only 39% of participants' responses), it is a psychological

regularity in another important sense. That is, if we were to randomly sample one additional person from the population and have them engage in Study 3's judgment task, the pattern we should predict is this pattern, as its occurrence is distinguishable from random responding and it has the highest observed probability of all possible patterns. We note that this may not be as compelling as finding a majority effect, but it is an empirical reality.

Overall, our results reveal the importance of perceived temptation on moral judgment. Theoretically, our studies provided tentative explanations for why previous research showed conflicting results, showing that, across stimuli and various ways of measuring moral judgment, people tend to judge agents who overcome immoral temptation as less virtuous than agents who never experience temptation in the first place. We also extended the literature on temptation, close relationships, and moral judgment by showing that the most common psychological experience is believing that being tempted to harm close others is even less virtuous than being tempted to harm strangers. These results add to a growing literature aimed at resituating moral psychology in everyday relational contexts (see Bloom, 2011). Methodologically, our findings corroborate recent calls to adopt analytic procedures that enable inferences about person-level psychology (e.g., Richters, 2021; McManus et al., *in press*; Speelman & McGann, 2020). We hope the current research is one of the first of many to adopt these approaches and therefore resituate moral psychology not just in everyday relational contexts, but also in moral psychologists' intended object of study: individual persons.

Situating the Current Research within the Broader Moral Psychology Literature

The current studies connect to work broadly assessing how motive inferences shape moral judgment. For example, donors who give anonymously in a double-blind way versus not, or those who behave prosocially in private versus in public, are judged as more charitable and

more virtuous (De Freitas et al., 2019, Kraft-Todd et al., *in press*), effects likely and empirically driven by differences in perceived reputational benefits. Additional research has documented how stated motives affect judgments of prosocial actors, with those motivated by empathy being judged more positively than those motivated by not wanting to feel distress (Erlandsson et al., 2020), and those motivated by warm-glow emotions being judged more positively than those who are motivated by material rewards (Barasch et al. 2014). Extensive work shows similar effects on the relation between inferred motives, stated motives, and moral judgment (see Berman & Silver, 2022, for an overview). All of this research suggests that the inferred or stated reasons for why agents make prosocial or selfless decisions have strong effects on assessments of their underlying character. In relating these findings to the current work, it is likely that participants in our paradigm were making inferences about what kind of person would be tempted to engage in certain immoral behaviors (e.g., cheat on a spouse) and therefore their likelihood of actually engaging in that behavior in the future.

More relevant to the current paper's studies, high effort prosocial actors are judged more positively than low effort prosocial actors (Berry & Lucas, 2022; Bigman & Tamir, 2016), as effort is a signal of how important it is to behave morally. However, if we consider our studies as manipulating mental effort, our findings are inconsistent with Berry & Lucas (2022) and Bigman & Tamir (2016). Our findings are most consistent with those of Critcher et al. (2012), where slow and deliberative (i.e., high mental effort) prosocial behavior results in worse character judgments when compared to quick (i.e., low mental effort) prosocial behavior. We suggest that this is due to the inferred motives of agents who experience temptation. As argued in Critcher et al. (2012), the presence of an immoral temptation communicates that the agent has multiple competing motives (i.e., to do both good and bad), or at least has a predisposition to consider

immoral behavior. As hypothesized earlier, the existence of these competing motives or predispositions might lead to the inference that the agent has a tainted moral character and therefore might actually behave immorally in the future. The absence of an immoral temptation, on the other hand, communicates that the agent does not have competing motives and may not have a predisposition for immoral behavior.

Limitations and Future Directions

The current research has several important limitations. First, as with much moral judgment research, our methodology, and therefore our inferences, relied on giving participants explicit access to mental state information that is rarely available outside of a lab setting. That is, participants were able to know that agents were internally conflicted versus unconflicted about engaging in prosocial behavior. In the real world, this information must typically be inferred via behavior (e.g., decision speed, see Critcher et al., 2012). Therefore, the current research might be most analogous to instances in which moral judgments are made after others have revealed their thoughts either directly through conversation or indirectly through gossip. In general, future research should consider the robustness of the observed effects across more ecologically valid manipulations of temptation. For example, consider two men, John and Tony, out at two separate bars. They both are in committed relationships, and both happen to run into ex-girlfriends from college. When the bars close, the ex-girlfriends ask the men to go home with them. When asked, John decides to take a taxi back to his ex-girlfriend's apartment. However, when he gets out of the taxi, he paces around and ultimately gets back in the taxi and goes home. Tony, on the other hand, declines the offer and calls a taxi to go home. Here, temptation is manipulated by varying how each agent behaves in a multi-step plan before ultimately making the same decision to go home.

Second, we measured only third-person judgments of temptation. However, as has been shown in other moral judgment research (e.g., Hirschfeld-Kroen et al., 2021), there may be interesting and important first- versus third-person dissociations. Specifically, because people often have stronger priors about themselves compared to others, perhaps they generate situational attributions for their own experiences of temptation and therefore do not discount their own virtuosity when this occurs. On the other hand, people may be more willing to generate dispositional attributions when they infer that others have experienced temptation, therefore discounting their virtuosity. This possibility is consistent with classic social psychological theorizing (e.g., Reeder & Brewer, 1979) and recent research on motivated versus rational belief maintenance from a third-party perspective (Kim, Park, & Young, 2020; Kim et al., 2021). Future research can determine whether similar processes occur when comparing first- to third-person attributions in general, and in the realm of temptation specifically.

Third, while our work expanded on others' research on the judgmental consequences of perceiving temptation (Berman & Small, 2018; Starmans & Bloom, 2016; Zhao & Kushnir, 2022), it cannot provide inferences about downstream behavioral consequences. As prior research suggests that trustworthiness is the most valued trait in others (Cottrell, Neuberg, & Li, 2007), understanding how inferred temptation affects perceived trustworthiness and actual trusting behavior is important. A modified two-stage economic exchange could accomplish this. In phase one, participants could first play the role of "receiver" in two one-shot dictator games. One confederate allocator could initially distribute resources in a selfish manner, but before the allocation is ultimately made, they could redistribute resources fairly. The other confederate allocator could simply allocate resources fairly from the outset. In phase two, the participants could play as "trustors" in a trust game with each of the confederates, revealing whether

differential trust occurs as a function of confederates' previous revealing (or not) of their temptation. Results of this study might suggest that social decision-making is affected by inferred temptations to behave selfishly.

Finally, recent calls have been made for researchers to communicate constraints on the generality of their findings (see Simons, Shoda, & Lindsay, 2017; Yarkoni, 2020). Importantly, the current studies surveyed only U.S. participants using various crowdsourcing platforms. Although research suggests that some moral judgment effects are consistent across most cultures (e.g., Barrett et al., 2016), it is unknown how temptation would be judged outside of the U.S. specifically or outside of WEIRD cultures generally (Henrich et al., 2010a, 2010b), though developmental evidence may provide a clue: young children judge agents who overcome immoral temptation as less virtuous than non-tempted agents (stimulus issues notwithstanding, Starmans & Bloom, 2016; Zhao & Kushnir, 2022). Future research is needed to determine the universality or uniqueness of the effects reported here. Relatedly, the methodology used throughout this paper (i.e., examining person-level responses) makes clear another sampling issue, even within cultures. Specifically, who are the people who show these effects or not, and more generally, who are the people being sampled? While robustness checks of our primary effect (i.e., that overcoming immoral temptation is judged as less virtuous than never being tempted) suggest that it occurred across all levels of all demographic factors collected (see the "Robustness Plots" in each RMarkdown on our OSF page: https://osf.io/bym4t/?view_only=ab62a6698bb84e34a53892e39576623f), the proportions of certain levels of some demographic factors were worrisome if our goal–or psychology's goal more generally–is to make universal claims. Our samples were overwhelmingly White, liberal, and educated, not atypical of psychology research conducted online or on college campuses.

Therefore, although our statistical methodology allowed sample-to-population inferences about the prevalence of our effects, it is an important and open question whether these prevalence estimates would systematically differ as demographic diversity of samples increase.

Conclusion

By expanding the theoretical and methodological reach of previous research, the current work investigated how people judge agents' temptations and whether these judgments are affected by agents' relationships to potential targets of their temptations. Specifically, most people (72-81%) judge agents who overcome temptations as less virtuous than agents who never experience temptation (see Studies 1-3). Moreover, when considering who would be harmed if an agent were to give in to their temptation, the most common pattern of judgment (from 39% of people) suggested that, in addition to judging agents who overcome temptations as less virtuous than agents who never experience temptation, agents who overcome temptations to harm close others (i.e., best friends or siblings) are judged as even less virtuous than agents who overcome temptations to harm strangers (see Study 3). Together, the methods and results of the current paper add to two exciting and growing literatures: the importance of resituating moral psychology into everyday relational contexts, and the importance of the individual person (rather than the population mean) in the study of psychology broadly. The continued interplay between these fields may end or rejuvenate classic debates in moral psychology, a prospect that, no matter the outcome, we are eager to witness.

Footnotes

1. These analyses attempt to control the false-positive rate at the person-level (i.e., through using separate typical group-level tests on each person's data). Therefore, because we only had low-trial data (i.e., no more than 10 trials per person per condition), we could not conduct these tests on each person's data, so we must assume person-level patterns are false-positive-controlled at the nominal .05 level. See Footnote 4 for more info.

2. We did not directly replicate Starmans & Bloom (2016) because we reasoned that Starmans & Bloom's stimuli were problematic for answering the question of whether tempted agents are more or less virtuous than non-tempted agents (in a way that differs from another critique of their stimuli as not being about inner conflict but rather being about costliness of actions, see Zhao & Kushnir, 2022). Specifically, Starmans & Bloom's stimuli confound non-temptation with active dislike of the activity that could constitute temptation, disallowing a clean comparison between tempted and non-tempted agents. For example, in Study 1 of Starmans & Bloom, two children promised their moms they would clean up their toys. In both cases, the children's friends are playing outside. The tempted child really wants to break the promise by going outside and playing with her friends, but she ultimately keeps her promise and cleans up her toys even though it was very hard to make this decision. The non-tempted child is described as not wanting to play with her friends and disliking playing outside, making it an easy decision to keep her promise and clean up her toys. That is, Starmans and Blooms' non-tempted agents are not just lacking the desire to break their promise, they also actively dislike the activity that could give rise to a temptation in the first place.

3. Although we collected these data, they ultimately were not relevant to the primary aim of the paper. We included them to test the robustness of Berman & Small's claimed interaction pattern (i.e., that overcoming temptation would be judged as more virtuous than non-temptation in non-moral contexts, but overcoming temptation would be judged as less virtuous than non-temptation in moral contexts). We failed to consistently replicate the non-moral context effect and therefore failed to consistently replicate the claimed interaction effect. These analyses are reported in the SOM.

4. We pre-registered sample sizes based on a person-level analysis that we no longer believe is appropriate. For context, while conducting the current research, two of the four authors had a methods and statistics paper under review (McManus et al., *in press*) that, over multiple revisions and resubmissions, ultimately led to a reformulation of what ought to be considered appropriate for person-level analyses. Therefore, some of our pre-registrations' original justifications for sample sizes are no longer relevant.

5. Although such results would provide strong evidence against Starmans & Bloom's (2016) results being typical across a variety of stimuli, it would not rule out the possibility that their results can occur for a minority of stimuli. Indeed, in their general discussion, they clearly communicate that their results provide evidence that their claimed effects "can" occur. That is, interpretation of their data seems to be more aligned with their providing an existence proof than a general regularity.

References

Allefeld, C., Görgen, K., & Haynes, J. D. (2016). Valid population inference for information based imaging: From the second-level t-test to prevalence inference. *Neuroimage, 141*, 378-392.

Aristotle. (trans. 1985). *Nicomachean ethics* (T. Irwin, Trans.). Indianapolis, IN: Hackett.

Barasch A., Levine, E.E., Berman, J.Z., Small, D.A. (2014). Selfish or selfless? On the signal value of emotion in altruistic behavior. *Journal of Personality and Social Psychology, 107*, 393–413.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., et al. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences, 113*(17), 4688–4693.

Berman, J.Z., & Silver, I. (2022). Prosocial behavior and reputation: When does doing good lead to looking good? *Current Opinion in Psychology, 43*, 102-107.

Berman, J. Z., & Small, D. A. (2018). Discipline and desire: On the relative importance of willpower and purity in signaling virtue. *Journal of Experimental Social Psychology, 76*, 220-230.

Berry, Z., & Lucas, B. J. (2022). How much is enough? The relationship between prosocial effort and moral character judgments. *Personality and Social Psychology Bulletin.*

Bigman, Y, & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General, 145*, 1654–1669

Bloom, P. (2011). Family, community, trolley problems, and the crisis in moral psychology. *The Yale Review, 99*(2), 26 - 43.

Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology, 92*, 208–231.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How Quick Decisions Illuminate Moral Character. *Social Psychological and Personality Science, 4*(3), 308–315.

De Freitas, J., DeScioli, P., Thomas, K.A., Pinker, S. (2018). Maimonides' ladder: States of mutual knowledge and the perception of charitability. *Journal of Experimental Psychology: General, 148,* 158–173.

Donhauser, P. W., Florin, E., & Baillet, S. (2018). Imaging of neural oscillations with embedded inferential and group prevalence statistics. *PLoS Computational Biology, 14*(2), e1005990.

Erlandsson, A., Wingren, M., & Andersson, P.A. (2020). Type and amount of help as predictors for impression of helpers. *PloS One, 15*, 1–23.

Everett, J.A.C., Pizarro, D., & Crockett, M. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772 - 787.

Faul, F., Erdfelder, E., Lang, A., & 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.

Grice, J.W. (2021). Drawing inferences from randomization tests. *Personality and Individual Differences, 179*, 110963.

Grice, J.W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science, 3*(4), 443-455.

Halfon, Mark, 1989. *Integrity: A Philosophical Inquiry*, Philadelphia: Temple University Press.

Henrich, J., Heine, S.J., & Norenzayan, A. (2010a). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61-135.

Henrich, J. Heine, S.J., & Norenzayan, A. (2010b). Most people are not WEIRD. *Nature, 466*, 29-29.

Hirschfeld-Kroen, J., Jiang, K., Wasserman, E., Anzellotti, S., & Young, L. (2021). When my wrongs are worse than yours: Behavioral and neural asymmetries in first-person and third-person perspectives of accidental harms. *Journal of Experimental Social Psychology, 94*, 104102.

Ince, R, A.A., Kay, J.W., & Schyns, P.G. (2021). Within-participant statistics for cognitive science. *Trends in Cognitive Sciences, 26*(8), 626-630.

Ince, R, A.A., Paton, A.T., Kay, J.W., & Schyns, P.G. (2021). Bayesian inference of population prevalence. *eLife, 10*, e62461.

Kant, I. (1998). *The groundwork for the metaphysics of morals* (M. J. Gregor, Trans.). Cambridge, England: Cambridge University Press. (Original work published 1785)

Kim, M., Mende-Siedlecki, P., Anzelotti, S., & Young, L. (2020). Theory of mind following the violating of strong and weak prior beliefs. *Cerebral Cortex, 31*(2), 884-898.

Kim, M., Park, B., & Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences, 24*(2), 101-111.

Kraft-Todd, G., Kleiman-Weiner, M., & Young, L. (in press). Virtue discounting: Observability reduces moral actors' perceived virtue. *Open Mind*.

Law, K.F., Campbell, D., & Gaesser, B. (2021). Biased benevolence: The perceived morality of effective altruism across social distance. *Personality and Social Psychological Bulletin, 48*(3), 426-444.

Marshall, J., Gollwitzer, A., Mermin-Bunnell, N., Shinomiya, M., Retelsdorf, J., & Bloom, P. (2022). How development and culture shape intuitions about prosocial obligations. *Journal of Experimental Psychology: General*.

Marshall, J., Wynn, K., & Bloom, P. (2020). Do children and adults take social relationship into account when evaluating other peoples' actions? *Child Development*, 91, 1395–1835.

McManus, R.M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science, 31*(3), 227-242.

McManus, R.M., Mason, J.E., Young, L. (2021). Re-examining the role of family relationships in structuring perceived helping obligations, and their impact on moral evaluation. *Journal of Experimental Social Psychology, 96*, 104182.

McManus, R.M., Young, L., & Sweetman, J. (*in press*). Psychology is a property of persons, not averages or distributions: Confronting the group-to-person generalizability problem in experimental psychology. *Advanced in Methods and Practices in Psychological Science.*

Moore, S., Speelman, C. P., & McGann, M. (2023). Pervasiveness of effects in sample-based experimental psychology: A re-examination of replication data from nine famous psychology experiments. *New Ideas in Psychology, 68,* 100978.

Reeder, G.D., & Brewer, M.B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86*(1), 61–79.

Richters, J.E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology, 43*(6), 366-405.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128.

Speelman, C.P., & McGann, M. (2020). Statements about the pervasiveness of behavior require data about the pervasiveness of behavior. *Frontiers in Psychology*, *11*, 1-16.

Starmans, C., & Bloom, P. (2016). When the spirit is willing, but the flesh is weak: Developmental differences in judgments about inner moral conflict. *Psychological Science, 27*(11), 1498 - 1506.

Taylor, Gabriele, 1981. 'Integrity,' *Proceedings of the Aristotelian Society* (Supplementary Volume) 55: 143–159.

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 45, E1.

Zhao, X., & Kushnir, T. (2022). When it's not easy to do the right thing: Developmental changes in understanding cost drive evaluations of moral praiseworthiness. *Developmental science*, e13257.