**Licensing via credentials: Replication of Monin and Miller (2001) with extensions investigating the domain-specificity of moral credentials and the association between the credential effect and trait reputational concern**

Qinyu Xiao

Department of Occupational, Economic, and Social Psychology, Faculty of Psychology, University of Vienna, Austria

[qinyu.xiao@univie.ac.at](mailto:qinyu.xiao@univie.ac.at)

ORCID: 0000-0002-9824-9247

\*Lok Ching Li, \*Ying Lam Au, \*See Ngueh Tan, \*Wing Tung Chung

Department of Psychology, University of Hong Kong, Hong Kong S.A.R.

[u3547924@connect.hku.hk](mailto:u3547924@connect.hku.hk) / [lilokching@gmail.com](mailto:lilokching@gmail.com)

[u3546649@connect.hku.hk](mailto:u3546649@connect.hku.hk) / [auyinglam0803@gmail.com](mailto:auyinglam0803@gmail.com)

[u3547916@connect.hku.hk](mailto:u3547916@connect.hku.hk) / [pristinatan@gmail.com](mailto:pristinatan@gmail.com)

[u3548801@connect.hku.hk](mailto:u3548801@connect.hku.hk) / [phebechung@gmail.com](mailto:phebechung@gmail.com)

^Gilad Feldman

Department of Psychology, University of Hong Kong, Hong Kong S.A.R.

[gfeldman@hku.hk](mailto:gfeldman@hku.hk) / [giladfel@gmail.com](mailto:giladfel@gmail.com)

\* Contributed equally, joint second author

^ Corresponding author

**Author bios**

Qinyu Xiao is a Ph.D. student at the Department of Occupational, Economic and Social Psychology, University of Vienna.

Lok Ching Li, Ying Lam Au, See Ngueh Tan, and Wing Tung Chung were students at the Department of Psychology, University of Hong Kong during the academic year 2020-21.

Gilad Feldman is an assistant professor at the Department of Psychology, University of Hong Kong.

**Declaration of conflict of interest**

The authors declare no potential conflicts of interests with respect to the authorship and/or publication of this article.

**Financial disclosure/funding**

The project is supported by the University of Hong Kong Teaching Development Grant.

**Authorship declaration**

Lok Ching Li, Ying Lam Au, See Ngueh Tan, and Wing Tung Chung analyzed the target article, produced the study materials, and wrote initial statistical analysis plans. Qinyu Xiao supervised the team, finalized the materials and analysis scripts, and prepared the manuscript for submission. Gilad Feldman guided the project, collected data, conducted the pre-registration, and edited drafts for submission.

**Rights**

CC-BY or equivalent license is applied to the AAM arising from this submission ([clarification](https://www.researchgate.net/deref/https%3A%2F%2Fbit.ly%2Frrs-primer)).

# Contributor Roles Taxonomy

Based on [CRediT](https://www.casrai.org/credit.html), the authors’ respective contributions are identified in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Role** | **Qinyu Xiao** | **Lok Ching Li, Ying Lam Au, See Ngueh Tan, & Wing Tung Chung** | **Gilad Feldman** |
| Conceptualization |  | X | X |
| Pre-registration (writing) | X | X |  |
| Data curation | X |  | X |
| Formal analysis | X | X |  |
| Funding acquisition |  |  | X |
| Investigation | X | X |  |
| Pre-registration peer review | X | X |  |
| Pre-registration verification | X |  | X |
| Data analysis peer review | X | X |  |
| Data analysis verification | X |  |  |
| Methodology | X | X |  |
| Project administration | X |  | X |
| Resources |  |  | X |
| Supervision | X |  | X |
| Validation | X |  |  |
| Visualization | X |  |  |
| Writing – original draft | X | X |  |
| Writing – review and editing | X |  | X |

# Stage 1 Snapshot (Revised)

**Provisional title.**

Licensing via credentials: Replication of Monin and Miller (2001) with extensions investigating the domain-specificity of moral credentials and the association between the credential effect and trait reputational concern

**Authors and affiliations.**

Qinyu Xiao, \*Lok Ching Li, \*Ying Lam Au, \*Wing Tung Chung, \*See Ngueh Tan, ^Gilad Feldman

[Department of Psychology, University of Hong Kong, Hong Kong S.A.R., China; \*contributed equally, shared second author, ^corresponding author]

**Field and keywords.**

Field: social psychology; Keywords: moral licensing, moral credential, morality, replication, reputation

**Research question(s) and/or theory.**

We aim to replicate and extend Monin and Miller (2001) Study 2 and will test their theory and hypotheses with added extensions. Research questions:

* + - 1. Replication: Do previous moral behaviors (which presumably give one the “credential” of being moral) lead to subsequent morally questionable behaviors?
      2. Extension 1: Is the moral credential effect stronger when the previous moral behavior and the subsequent morally questionable behavior are in the same (vs. different) domain?
      3. Extension 2: Are individual differences in trait-level reputational concern associated with the size of the credential effect?

**Hypotheses (where applicable).**

H1: “Moral credential effect”: Participants who acted in a way that can presumably establish them as moral are more likely to engage in a potentially morally problematic act subsequently.

H2: The moral credential effect (H1) is more prominent when the previous acts and the subsequent acts are in the same domain.

H3: People higher on reputational concern exhibit a larger credential effect.

**Study design and methods.**

In this replication, we use a 3 (moral credential type) by 2 (scenario) between-subjects design. Participants will be given a chance to obtain a non-sexist credential (i.e., a chance to act in a way that shows “I am not a sexist.”) or a non-racist credential, or no such chance (control). They will then be presented with a scenario that allures them to indicate a conceivably sexist or racist preference. It is predicted that those with a credential (vs. control) will tend to indicate a more conceivably sexist/racist preference in these scenarios.

We added “cross-over” conditions to examine whether the effects of credentials are domain-specific (such that, e.g., a non-sexist credential does not make people more likely to express conceivably racist preferences). We will also measure participants’ trait reputational concern in the end.

We aim to recruit US participants from MTurk. We will employ best practices, tools, and survey design for ensuring comprehension, attentiveness, and high-quality data collection in labor markets. We aim to determine sample size with a power analysis (at least 80%, 0.05) on a reasonable estimate of effect size from the literature.

**Key analyses that will test the hypotheses and/or answer the research question(s).**

The key dependent variable is participant’s gender/ethnicity preferences (which might be considered sexist/racist in the scenarios used here). A two by three factorial ANOVA will be conducted on this variable, with planned contrasts comparing the credential conditions against the control. We will use multiple linear regression to examine the relationship between trait reputational concern and licensing.

**Conclusions that will be drawn given different results.**

We will evaluate the replicability of our findings against the original based on LeBel et al.’s (2019) criteria. Our extensions will offer new insights and directions for the literature of moral licensing.

**Key references.**

1. Monin and Miller (2001). <https://doi.org/10.1037/0022-3514.81.1.33>

2. LeBel et al. (2019). <https://doi.org/10.15626/MP.2018.843>

# Study Design Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Research question** | **Hypothesis** | **Analysis plan** | **Sampling plan** | **Rationale for the tests** | **Interpretation given different outcomes** | **Theory that could be shown wrong by the outcomes** |
| Do previous moral behaviors that give one moral credentials make people more likely to engage in morally questionable behaviors later? | Moral credentials make people more likely to engage in subsequent morally questionable acts. | ANOVA | Amazon Mechanical Turk via CloudResearch (with .90 power to detect a *d* = 0.25 credential effect) | We used the same test as in our replication target (Study 2 in Monin & Miller, 2001), albeit with a minor tweak to test our extension hypothesis. | There could be multiple reasons behind a non-replication. Our evaluation of the replication outcomes will follow LeBel et al.’s (2019) criteria. | The moral credential model of moral licensing |
| Do moral credentials work better in licensing immoral behaviors in the same domain than in a different domain? | Moral credentials work better in licensing immoral behaviors in the same domain than in a different domain. | N/A | Ambiguous moral transgressions (in the study: expression of conceivably prejudiced preference) are better licensed by credentials in the same domain than in a different domain (Effron & Monin, 2010). |
| Is trait reputational concern negatively associated with the expression of conceivably prejudiced preferences? | Trait reputational concern is negatively associated with the expression of conceivably prejudiced preferences. | Multiple linear regression | We want to examine whether and under what conditions (particularly, with vs. without credentials) do reputational concern predicts expression of conceivably prejudiced preferences. | N/A | N/A |
| Do moral credentials moderate the relationship between reputational concern and the expression of conceivably prejudiced preferences? | Moral credentials attenuate the negative association between reputational concern and the expression of conceivably prejudiced preferences. | N/A | N/A |

# Abstract

**IMPORTANT: This is a Registered Report Stage 1 before data collection. Written in past tense as a template to simulate what the final manuscript will look like. No pre-registration or data collection have been conducted.**

The moral credential effect is the phenomenon where an initial behavior that presumably establishes one as moral “licenses” the person to subsequently engage in morally questionable behaviors. In line with this effect, Monin and Miller (2001, Study 2) found that participants initially given an opportunity to hire a job candidate from disadvantaged groups (vs. those without such an opportunity) subsequently indicated preferences that were more likely to be perceived as prejudiced. We conducted a direct replication of this study with U.S. participants from Amazon Mechanical Turk using CloudResearch (*n* = []). We found [week/medium/strong] support for the original findings. With our extensions to the replication study, we found []. All materials, data, and analysis scripts are shared at <https://osf.io/phym3>.

*Keywords*: moral licensing, moral credentials, morality, replication, reputation

Licensing via credentials: Replication of Monin and Miller (2001) with extensions investigating the domain-specificity of moral credentials and the association between the credential effect and trait reputational concern

**IMPORTANT: This is a Registered Report Stage 1 before data collection. Written in past tense as a template to simulate what the final manuscript will look like. No pre-registration or data collection have been conducted.**

Moral licensing is the phenomenon that moral acts “liberate individuals to engage in behaviors that are immoral, unethical, or otherwise problematic, behaviors that they would otherwise avoid for fear of feeling or appearing immoral” (Merritt et al., 2010). Imagine that a manager of a small cement manufacturing company is seeking to hire a new representative to travel to building sites to solicit new clients and negotiate contracts. Since the market is highly competitive and technical, the representative is expected to be aggressive during bargaining and show confidence when demonstrating skills. Knowing this, would the manager say that the job is better suited for a male, a female, or that it would be equally suited for both genders? The personal characteristics that this job demands might make the manager feel that the job is, in general, better suited for males. Yet the concern that this preference might appear sexist, however, might make them refrain from expressing this preference and instead say that gender is not a factor to consider. People sometimes suppress their views that they worry might be considered prejudiced (Crandall & Eshleman, 2003). Nonetheless, if they could establish that they are not prejudiced in advance, they would feel more comfortable—or “licensed”—to express a conceivably prejudiced preference. Indeed, Monin and Miller (2001) found that when participants had a chance to disagree with blatantly sexist statements (vs. those who had no such chance), they were more likely to indicate a preference for males in the scenario described above.

The moral licensing literature has proliferated in the past two decades, with hundreds of articles published on relevant topics (Rotella et al., 2022; Rotella & Barclay, 2020). While the sheer amount of supporting evidence may suggest that the phenomenon is robust, recent studies, however, point to a considerable publication bias (Blanken et al., 2015; Kuper & Bott, 2019; Rotella et al., 2022; Simbrunner & Schlegelmilch, 2017). There are several recently published unsuccessful replications (Blanken et al., 2014; Giurge et al., 2021; Rotella & Barclay, 2020; Urban et al., 2019, 2020) and insufficient support for some of the theorized moderators (see, e.g., Blanken et al., 2015). The moral licensing literature thus would benefit from more pre-registered and highly powered direct replications. Ideally, these replications would be published as Registered Reports, as this emerging publication format effectively reduces publication bias (Scheel et al., 2021).

This Registered Report is a direct replication and extension of Study 2 in Monin and Miller (2001), which is the pioneering work on the moral credential effect, a subcategory of moral licensing. In the following, we present a brief review of the phenomenon and outline our motivations for conducting this replication. We conclude with an overview of our replication study.

## Moral licensing: Credentials and credits

The idea behind the moral credential effect is that a certain behavioral history (e.g., having minority friends on social media) can help people establish “credentials” that they possess certain positive characteristics (e.g., being anti- or non-racist). As a result, subsequently morally questionable behaviors (e.g., making conceivably prejudiced comments against ethnical minorities) are less attributed to genuine prejudice (but more to, for instance, situational factors) and may appear less wrong (Bradley-Geist et al., 2010; Thai et al., 2016). Importantly, the credentials license morally dubious behaviors by altering how people interpret them (Merritt et al., 2010; Miller & Effron, 2010). To illustrate, consider Dutton’s (1971) observation that restaurants with dress code regulations were more likely to turn down Black couples who violated those regulations if they had previously turned down a White couple for the same reason. Presumably, turning down a White couple who did not comply with dress code regulations provided the front desk with a moral credential that later helped them be confident that their decision would not be considered prejudiced (but fair) as they decided to turn down a Black couple (Miller & Effron, 2010). Because moral credentials license by altering interpretations of behaviors, in theory, they work best when the behaviors are morally ambiguous, which, due to their ambiguity, afford multiple interpretations (Effron & Monin, 2010; Miller & Effron, 2010; Mullen & Monin, 2016). We study such ambiguous behaviors in this research.

Less of a focus here are *moral credits*, with which one can also be licensed. Moral credits are compared to bank deposits, and the idea is that one accrues these credits by doing good and uses them to balance out subsequent transgressions, which are correspondingly conceptualized as *moral debits* (Miller & Effron, 2010, p. 125). So long as one has enough “savings” in their account, one would feel more comfortable with spending them to engage in immoral behaviors, and others would condone these behaviors to some extent. When credits are used up, however, one goes morally “bankrupt” and may receive harsher blame and heavier punishment for the same misdeed (consider the Boy Who Cried Wolf). Unlike moral credentials, moral credits do not change the interpretation of misdeeds. The misdeeds licensed by moral credits would *not* be judged less wrong or harmful, or simply not immoral. An analogy may be drawn between moral credits and the “carbon offsets” one purchases before engaging in an environmentally harmful action. Although the purchase makes one’s environmental impact neutral on paper, it has no influence on the harmfulness of that very environmentally unfriendly action (Miller & Effron, 2010).

## Need for replication

Moral licensing has received empirical support from both experiments (e.g., Conway & Peetz, 2012; Monin & Miller, 2001; Sachdeva et al., 2009) and field studies (e.g., Hofmann et al., 2014; Lacasse, 2019; Meijers et al., 2015) and across a wide variety of contexts, such as hiring (Effron et al., 2009; Monin & Miller, 2001), environmental conservation (Geng et al., 2016; Lalot et al., 2018), charitable giving (Conway & Peetz, 2012; Meijers et al., 2015), and volunteering (Conway & Peetz, 2012). Researchers have also proposed and tested many extensions of the effect. For instance, when people anticipate doing something morally dubious, they seem to strategically establish moral credentials in advance by demonstrating, if not exaggerating, their good morals (Merritt et al., 2012). There is also evidence that people can be morally licensed not only by their own good behaviors, but also by those of their ingroup members, a phenomenon called vicarious moral licensing (Kouchaki, 2011).

Despite the proliferating literature, there is, however, also evidence that the moral licensing effect is weaker than what the large number of relevant studies may imply. Multiple meta-analyses have revealed evidence of publication bias in this literature (Blanken et al., 2015; Kuper & Bott, 2019; Rotella et al., 2022; Simbrunner & Schlegelmilch, 2017). The effect size estimates ranged from 0.18 to 0.32 in Cohen’s *d* prior to bias correction, indicating a small-to-medium effect. But they dropped to *d* = 0.18, 95% CI [0.06, 0.29] (Kuper & Bott, 2019) and *d* = −0.003, 95% CI [−0.07, 0.06] (Rotella et al., 2022) when bias was corrected with three-parameter selection models (Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995), and even to *d* = −0.02, 95% CI [−0.12, 0.08] (Rotella et al., 2022) or *d* = −0.05, 95% CI [−0.26, 0.16] (Kuper & Bott, 2019) with PET-PEESE (Stanley & Doucouliagos, 2014). Some of those meta-analytic estimates imply a tiny effect in the opposite direction of moral licensing, or what is called a “moral consistency” effect (Mullen & Monin, 2016). That is, previous moral behaviors drive people to continue doing good. Perhaps unsurprisingly, most studies included in these meta-analyses did not have sufficient power to detect even the most optimistic effect size estimate; the average *n* was 130.6 (Rotella et al., 2022).[[1]](#footnote-2)

The lack of power in original studies might explain the null findings in some follow-up, high-powered conceptual and direct replications (Blanken et al., 2014; Giurge et al., 2021; Urban et al., 2019). Contrary to moral licensing and results from Sachdeva et al. (2009), Blanken et al. (2014) found no evidence that writing positively about oneself makes participants donate more to charities than writing neutrally. Contrary to Mazar and Zhong (2010), with both conceptual and very close replications, Urban et al. (2019) reported that participants were not more likely to cheat after consuming green products—essentially act pro-environmentally. Effron et al. (2009) reported that participants who had the chance to endorse Barack Obama—the first African American U.S. president—favored a White job applicant subsequently. Giurge et al. (2021), however, failed to find evidence that endorsing a female Democrat against male candidates would make Democrat participants favor males over females for a stereotypically masculine job. Concurring with the authors of existing meta-analyses on moral licensing (Blanken et al., 2015; Kuper & Bott, 2019; Rotella et al., 2022), we believe that this literature would benefit from more highly powered direct replications of previous studies to obtain more accurate effect size estimates and potentially also verify conclusions about the moderators of the effect.

## The replication target: Study 2 in Monin and Miller (2001)

We chose to replicate Study 2 in Monin and Miller (2001) for two reasons. First, the article pioneered the study of moral licensing/credentials and has been highly impactful, with over 1,400 citations as of May 2023 per Google Scholar data. The high impact of the article makes the findings especially important to revisit and reassess (Coles et al., 2018; Isager, 2018). Second, despite its impact, not all studies in the article have been subjected to a replication; for those that were, there were notable differences between the replication and the original results. A previous large-scale multi-site collaboration attempted to replicate Study 1 in the article (Ebersole et al., 2016). In the original study, participants first had to indicate whether they found right or wrong five statements that were either blatantly sexist (e.g., “most women are better off at home taking care of the children.”) in one condition or less so (e.g., “some women are better off at home taking care of the children.”) in the other.[[2]](#footnote-3) According to Monin and Miller (2001), because participants in the former condition would disagree with more statements, they “would presumably feel that they had stronger credentials as non-sexists and be correspondingly more willing to voice a politically incorrect preference” (p. 35). The results of the original study partially aligned with this prediction: male participants who read the blatantly sexist statements subsequently indicated stronger preferences for males for a job that requires male-typical characteristics (when confronted with the scenario described at the beginning of this article) than their counterparts who read the other version (*d* = 0.87); for female participants, the difference was negligible (*d* = 0.10). In contrast, the replication found similar moral credential effects across genders, but the effect size was much smaller (*d* = 0.14). This finding motivates examination of the replicability of the other findings in the original article. To our knowledge, there are no published pre-registered direct replications of Study 2 therein. Therefore, we chose Study 2 as our replication target.

Study 2 used similar dependent measures as Study 1: participants were either assigned to read the scenario mentioned above that asked preference between males and females for a job that demands male-typical characteristics, or a similar scenario that asked preferences between White and Black ethnicities for a position in a working environment that was described to be hostile to Black people. The study, however, used a different manipulation. It manipulated moral credentials with a recruitment task that required participants’ active choice. Participants were first to select one applicant from a total of five for a starting position at a large consulting firm. Crucially, one of the five applicants was made outstanding (i.e., the applicant had the best grade and graduated from the most prestigious college); this outstanding applicant was a White female in the *non-sexist* credential condition, a Black male in the *non-racist* credential condition, or a White male in the no-credential (or control) condition. The other applicants were all White males regardless of condition. It was reasoned that selecting the outstanding applicant who happened to be female/Black would give participants a non-sexist/non-racist credential (despite that the choice could have nothing to do with the applicants’ gender or ethnicity). And in line with the moral credential effect, in the original study, those in the non-sexist/non-racist credential conditions expressed stronger preferences for males/Whites in the subsequent scenario than the corresponding controls.

## Extensions: Domain specificity and reputational concern

We added two extensions to our replication. First, we tweaked the original study design and tested the idea that *ambiguous* moral transgressions are better licensed by moral credentials in the same domain than those in a different domain (Effron & Monin, 2010). For instance, a person who somehow proved that they are not sexist would be less likely to be blamed for behaviors that *might* be considered sexist (and hence ambiguous), like preferring males for a job that demands male-typical characteristics. This is because the non-sexist moral credential will lead people to attribute conceivably sexist behaviors to factors other than sexism. If this person only proved to be non-racist and had only a non-racist credential, they can still be accused of sexism. In other words, non-racist moral credentials are less effective in licensing conceivably sexist behaviors. To examine this idea, we included another two between-subjects conditions to our study (see Methods for details). Because these were between-subjects, the replication part was intact.

Second, we tested whether individual differences in reputational concern moderate the moral credential effect, which can be larger in those who are dispositionally more concerned about their reputations. A recent meta-analysis on the moral licensing literature revealed that studies with explicit observation (by experimenters or other participants) found larger effects than those with only some or no observation (e.g., online studies; Rotella et al., 2022). This suggests that licensing may result partly—if not mainly (Rotella et al., 2022)—from one’s feeling that they have established a good reputation with the previous moral acts (Effron, 2014; Miller & Effron, 2010). It follows that if people do not care about their reputation at all, having moral credentials or not will not matter, and moral credentials will have little to no effect. Few studies directly examined the relationship between moral licensing/credentials and (trait-level) reputational concern (but see Study 3 in Monin & Miller, 2001). We therefore added this extension.

## Overview of study

We replicated Study 2 in Monin and Miller (2001) with a U.S. sample recruited from the Amazon Mechanical Turk on CloudResearch. We included two major extensions to the replication. This research was submitted as a Registered Report (Chambers & Tzavella, 2022; Nosek & Lakens, 2014; Scheel et al., 2021; Wiseman et al., 2019). We made our data, study materials, and analysis scripts openly available at <https://osf.io/phym3>. We reported the results after exclusion (see the supplemental materials for our exclusion criteria) in the main text and full sample results in the supplemental materials. We confirm that all measures, manipulations, exclusions, and power analyses conducted for this investigation have been reported.

# Method

**IMPORTANT: Method and results sections were written using a randomized dataset produced by Qualtrics to simulate what these sections will look like after data collection. These will be updated following the data collection. This is written in past tense, yet no pre-registration or data collection have been conducted.**

## Sample size planning

Considering the results of existing replications (Ebersole et al., 2016) and meta-analyses (Blanken et al., 2015; Kuper & Bott, 2019; Rotella et al., 2022; Simbrunner & Schlegelmilch, 2017), we aimed to detect a moral licensing effect of *d* = 0.25. Detecting an effect of this size with a two-tailed independent-samples *t*-test at 90% power and .05 alpha requires 338 participants, 169 for each sample. The original study had four between-subjects conditions—two experimental and two control conditions—and for them we decided to recruit 700 participants. Because we had two additional conditions as our extension (see below for details), we aimed for 1,050 participants in total for this investigation. Note that since we aimed primarily at replication, we did *not* plan our sample size to ensure that we had sufficient power for the extension hypotheses. Therefore, any results in favor or disfavor of those extension hypotheses should be considered exploratory only and would require further confirmatory investigation.

Our justification for this planned sample size was primarily based on the maximum resources available to us for this project, and what we perceived to be reasonable resource constraints for typical labs (Lakens, 2022). The planned sample size was smaller than what would be ideally required to detect conservative meta-analytic effect size estimates, but still larger than typical sample sizes in the moral licensing literature. We believe requiring more participants beyond our planned sample size just for reliably detecting the moral licensing effect signals that the way we study the effect is not optimal and cost-efficient. And instead of using bigger samples, priority should be given to establishing alternative methods that yield robust effects at a cost that average research teams would find affordable.

## Participants

We recruited participants from Amazon Mechanical Turk (MTurk) via CloudResearch (Buhrmester et al., 2011; Litman et al., 2017; Paolacci & Chandler, 2014; Thomas & Clifford, 2017). Based on our experience conducting replications of social psychology and judgment and decision-making experiments using MTurk (CORE Team, 2023), and to ensure high data quality, we employed a series of CloudResearch options (see the supplemental materials for details). We compensated our participants based on the U.S. federal minimum wage of $7.25 per hour. We first pre-tested our study with 30 participants—paying each $1.00 based on an estimated duration of eight minutes—to ensure that we had an accurate estimate of completion time, and we adjusted payment if needed. The data of these 30 participants were not analyzed separately (but were included in the final sample for analysis), and they would be paid a bonus if the payment was adjusted upwards. In our recruitment, we indicated that we are looking for participants that were born and currently living in the U.S.

In total, [] participants completed a Qualtrics survey for $[], and [] were excluded based on the pre-registered criteria detailed in the supplemental materials. The sample after exclusion had [] participants (*M*age = [], *SD* = [], [] participants did not disclose their ages; [] ([]%) males, [] ([]%) females, [] ([]%) indicated their gender as others, and [] ([]%) preferred not to disclose this information).

## Design and procedure

The study had a two (*scenario*: gender preference or ethnicity preference) by three (*credential type*: non-sexist credential, non-racist credential, or no credential) between-subjects factorial design. The replication part consists of four of these conditions (gender preference/non-sexist credential, gender preference/no credential, ethnicity preference/non-racist credential, and ethnicity preference/no credential), and the remaining two conditions (gender preference/non-racist credential and ethnicity preference/non-sexist credential) were included as extensions (which will be referred to as “mismatched-credential” conditions below). Participants provided consent in the beginning. Then, they were asked two simple confirmation questions to ensure that they were willing and able to take part. If they did not answer yes to these questions, their sessions would be terminated, and they would be asked to return the task. These confirmation questions could help us exclude those participants who would not pay attention and only randomly click through the survey.

After that, participants went through two hiring scenarios. First, they were to select one applicant out of five for a starting position at a consulting firm. The profiles of the applicants presented to participants included the applicants’ photos, names, educational backgrounds, grades, and majors. Across all three credential-type conditions, one applicant was made the most appealing. The applicant had the highest GPA, was a Harvard graduate, and majored in economics. Crucially, this applicant was a White female in the non-sexist credential condition, a Black male in the non-racist credential condition, and a White male in the no-credential condition (names and photos were accordingly adjusted; see the supplemental materials for details). All other applicants were White males and had the same profiles across the conditions. The rationale behind this manipulation was that by selecting the female/Black star applicant, participants could obtain a non-sexist/non-racist moral credential. Participants saw the profiles together, with the star applicant presented at the same position (4th from top to bottom) across conditions. To make their choice, participants typed in the full name of the chosen applicant exactly as what was shown in the profiles. They could not proceed if their input was different.

After this hiring decision, participants completed three filler items about the building industry (if they were randomly assigned to the gender preference scenario condition) or the police force (if assigned to the ethnicity preference scenario condition). Participants then read either the gender preference scenario or the ethnicity preference scenario:

**The gender preference scenario**. Imagine that you are the manager of a small (45-person) cement manufacturing company based in New Jersey. Last year was a particularly good one, and after you invested in increasing the output capacity of your plant, you decide that it would be very fruitful if you could find clients in other states to increase your business. Because you cannot spend too much time away from the plant, you decide to appoint someone to go around to prospective clients and negotiate contracts. This is a highly specialized market, and the job will mostly consist in going from one building site to another, establishing contacts with foremen and building contractors. It is also a highly competitive market, so bargaining may at some points be harsh. Finally, it’s a very technical market, and a representative that did not exude confidence in their technical skills would not be taken seriously by potential clients. Realizing how useful such a help would be for you, you decide to give the person chosen one of the top-five salaries in your company. Do you feel that this job is better suited for one gender rather than the other?

**The ethnicity preference scenario**. Imagine that you are the police chief of a small town in a rural area of the U.S. Historically the population of the town has been exclusively White, and attitudes towards other ethnicities tend to be unfavorable. As much as you regret it, you know this is especially the case within your unit. You couldn’t help overhearing racist jokes coming from people you otherwise consider excellent officers. In fact, a couple of years ago an African American patrolman joined your unit, and within a year he quit, complaining about hostile working conditions. You are doing what you can to change attitudes, but your main objective is that the police force should do its job, and so far it has been rather effective so you do not want to provoke any major unrest within the ranks. The time has come to recruit a new officer. As a general rule, officers need to be responsible and trustworthy, show quick intelligence enabling them to make split-second decisions in crisis situations. Recent scandals have also highlighted the need for a high level of integrity, resistance to corruption, mild manners and a calm temper. You have just received applications from the new graduates of the local Police Academy. You wonder whether ethnicity should be a factor in your choice. Do you feel that this specific position (described above) is better suited for any one ethnicity?

As shown, the scenarios were constructed to imply a hostile working environment for females/Black people and a potentially justifiable preference for a male/White for the positions. The scenarios were presented first without the underscored part, and participants had to correctly answer two comprehension questions about the scenarios before they could proceed. If they answered any of the questions incorrectly, they would stay on the page and reattempt the questions. They could attempt as many times as they would like to, until they passed the checks.

After participants passed the comprehension checks, they were presented with the scenario again, this time with the underscored part. They then indicated whether they preferred a specific gender/ethnicity for the job described in the scenario on a 7-point scale (−3 = *Yes, much better for women/a Black*, −2 = *Yes, better for women/a Black*, −1 = *Yes, slightly better for women/a Black*, 0 = *No, I do not feel this way at all*, 1 = *Yes, slightly better for men/a White*, 2 = *Yes, better for men/a White*, 3 = *Yes, much better for men/a White*). Note that only text but no numeric labels were presented. This was a deviation from the original, which presented the numeric labels. We decided to deviate to address the possibility that some participants might be bothered or upset when seeing that preferences for female or Black people are represented as minuses, and preferences for male or White people as pluses. Since this was a subtle deviation, we did not expect that it would have a systematic influence. We henceforth call this dependent measure gender or ethnicity preference when referring to only one of them and hiring preference when referring to them together.

On a separate page, participants indicated their agreement with one of the following statements: “*Women are just as able as men to do any kind of job*” (if they were assigned to the gender preference scenario) or “*Blacks are just as able as Whites to do any kind of job*” (if assigned to the ethnicity preference scenario; 7-point scale: −3 = *disagree strongly*, −2 = *disagree*, −1 = *disagree slightly*, 0 = *neither agree nor disagree*, 1 = *agree slightly*, 2 = *agree*, 3 = *agree strongly*; again, only text labels were presented). These measures were not reported in the original published article but were included in the study materials. We included them to have a faithful replication. We call this dependent measure gender or ethnicity attitude henceforth.

**Reputational concern (extension)**. Following the hiring scenarios, participants completed a 7-item concern-for-reputation scale (de Cremer & Tyler, 2005) on a 5-point scale (1 = *not at all characteristic of me*; 5 = *extremely characteristic of me*). Sample items of the scale include “I am rarely concerned about my reputation” and “I do not consider what others say about me.” (Cronbach’s α = []; McDonald’s omega = []). The item scores were averaged to obtain an index of general, trait-level reputational concern.

**Exploratory questions**. One reviewer for this Registered Report at Stage 1 raised concerns over whether the manipulation can actually provide participants with moral credentials, suggesting that choosing the most outstanding candidate in the first hiring scenario does not necessarily imply anything about the decision-makers’ attitudes towards different genders or ethnicities. The reviewer also questioned whether participants—with or without credentials—would find it prejudicial to prefer males (or Whites) in the gender (or ethnicity) preference scenarios to begin with. As replicators, we had no clear answers to these questions. Nonetheless, addressing these concerns may prove fruitful and provide additional insights about the design of the original study. Therefore, we added a few exploratory questions towards the end of the survey and after the reputational concern scale.

Specifically, on one page, we presented participants with the same candidates’ profiles from the first hiring scenario again, and asked them to respond to the following items for each candidate: (1) “selecting [candidate’s last name] for the position means that the person who makes this decision is:” (1 = *very unlikely to be sexist/racist*, 2 = *somewhat unlikely to be sexist/racist*, 3 = *somewhat likely to be sexist/racist*, 4 = *very likely to be sexist/racist*; all participants evaluated both how sexist and racist the decisions were, separately and in random orders); (2) “selecting [candidate’s name] for the position is a morally good decision” (1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly agree*; only text labels were presented); (3) “selecting anyone but [candidate’s name] for the position is a morally bad decision” (1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly agree*; only text labels were presented). Therefore, there were four evaluations for each candidate, and 20 in total.

On another two pages, we asked questions about the gender and ethnicity preference scenarios, respectively. Specifically, we first presented the scenario, and asked participants to what extent people would consider different preferences prejudiced (1 = *not at all prejudiced*, 5 = *very prejudiced*; only the endpoints are labeled with text) for each of the preference options (e.g., “feeling that the job is much better suited for women”). The three pages (i.e., including the one that asked about the first hiring scenario) were presented in uniquely randomized orders. We asked participants about both gender and ethnicity preference scenarios because participants’ perceptions of general people’s attitudes in these scenarios could be influenced by whether they have expressed their own (Ross et al., 1977). Confronting them with the scenario that they did not encounter previously might reduce this influence. We did not do the same with the first hiring decision scenario (for example, giving participants also the profiles from the other credential condition and asked questions about them) because there were four candidates that remained the same across the conditions, and we did not want to reveal what was manipulated. Also, to keep the replication part intact, we placed all these exploratory questions to the very end, though it might be more desirable to have them directly after the corresponding scenarios. This was a limitation we had to accept, and we intended to gather only preliminary data on participants’ perception of the scenarios with these questions.

After these exploratory questions, participants completed a funneling section where they reported their guesses about the purpose of the study, how serious they were in filling in the survey, whether they had seen or completed surveys using similar scenarios. After reporting demographics, they were thanked and debriefed.

## Deviations

Our study had several deviations. (1) We conducted the study online whereas the original was conducted in a laboratory and in the paper-and-pencil format. (2) We conducted the study with U.S. residents on MTurk whereas the original was conducted with Princeton undergraduate students. (3) In the first hiring scenario, we asked participants to type in the full name of the applicant of their choice, whereas the original asked participants to circle the person’s profile and then write down the full name. We did not ask participants to “circle” because there was no straightforward way to implement the same on Qualtrics. (4) For the first scenario we did not use the original profile pictures but a different set of pictures from the Chicago Face Database (Ma et al., 2015) because the original pictures had low resolution (see the supplemental materials for how we selected those pictures). (5) We employed comprehension questions for the second hiring scenario to ensure that participants had a proper understanding of it; the original did not. (6) We did not present numbers in the scale point labels. We did not expect any of these deviations to systematically impact the replication outcomes.

## Hypotheses

We tested the following confirmatory hypotheses:

H1: Participants with non-sexist/non-racist moral credentials indicate stronger preferences for males/Whites than participants without moral credentials.

H2: Participants with non-sexist/non-racist moral credentials indicate stronger preferences for males/Whites than participants with non-racist/non-sexist moral credentials. In other words, moral credentials in the same domain as the behavior to be licensed produce a larger moral credential effect than credentials in a different domain.

H3: Trait reputational concern negatively predicts preferences for males/Whites in those who have no moral credentials.

H4: Non-sexist/non-racist moral credentials reduce the negative predictive power of trait reputational concern for preferences for males/Whites (as hypothesized in H3).

H1 describes the moral credential effect that Monin and Miller (2001) observed. We discussed the rationale behind H2 in the introduction: People should be more likely to condone a conceivably sexist act of a person who has proved to be a non-sexist, than the same act of one who only proved to be a non-racist. H3 describes the intuitively plausible idea that reputational concern prevents people from expressing their real attitudes or preferences on sensitive topics. H4 was motivated by the finding that higher observability, which is presumably associated with a higher reputational concern, was associated with a larger moral licensing effect (Rotella et al., 2022). In addition, it is most likely that moral credentials attenuate the negative association between reputational concern and expressed prejudice rather than reverse its direction. Hence our H4.

We note three additional points about our hypotheses. First, our H4 was based on an individual difference measure of reputational concern, whereas Rotella et al.’s (2022) meta-analysis only suggested an effect of contextual reputational concern. We thus caution our readers not to interpret our results as directly supportive or against Rotella et al.’s.

Second, we did not include any hypothesis concerning the gender or ethnicity attitude dependent measure. This measure was included in the original study, but the findings were not reported. Also, Ebersole et al. (2016) did not find evidence that moral credentials affected the expression of more general attitudes towards genders and ethnicities. We thus did not expect to find any substantial effects, either. We, however, included this measure to have a faithful replication. We will conduct similar exploratory analyses on this measure as those we conduct on hiring preferences.

Finally, given that we aimed at replication, the confirmatory testing of H1 did not include those “mismatched-credentials” conditions. Also, since H3 and H4 were largely exploratory, and we were uncertain how large the effects would be for these hypotheses, we also excluded “mismatched-credentials” conditions in our confirmatory analyses when testing these two hypotheses.

# Results

## Manipulation check

We will first examine whether most participants have chosen the star applicant, and whether the choice rate differs across different credential-type conditions (with a multiple-proportions test). Regardless of the results, we will follow the original article to conduct analyses both before and after removing those who do not choose the star applicant.

## Confirmatory analyses

We conducted confirmatory analyses both with and without those participants who indicated a preference for females/Blacks in the respective scenarios (whenever hiring preference was involved). By including them, we followed the original analyses. But we believe results are only internally valid without those participants. To illustrate, the study assumed that stronger preferences for males or Whites in the respective scenarios can be perceived to be more morally problematic (so that participants would be more likely to express them when they had credentials). It does not follow from this assumption that stronger preferences for females or Blacks are less problematic, or more moral, compared with a neutral preference and preferences for males or Whites. Nonetheless, that should be the case if we analyze our data the way the original did, which assumed a monotonic relationship between preferences (for one gender/ethnicity over the other) and how moral the preferences would appear along the entire scale. As such, removing those participants is necessary. We, however, will conduct analyses both with and without those participants, and we will report results without those participants in the main manuscript (and with them, in the supplemental materials, if the results differ substantially). We will evaluate the replication outcomes based on the results including these participants, as the original study did not exclude them.

Because we aim at direct replication, we will first perform the same analysis as in the original article. Also, all tests will be two-tailed unless noted otherwise. We will conduct a two-way (whether one has a moral credential [2: yes vs. no] × domain of credential/scenario [2: sexism vs. racism]) factorial ANOVA with hiring preference as the dependent variable. As such, this analysis does not include the two extension (or “mismatched-credentials”) conditions. Our H1 predicts a main effect of moral credentials, such that participants with a moral credential would indicate more preferences for males/Whites than those without credentials.

To test our hypotheses (specifically, H1 and H2) more parsimoniously, we will conduct a two-way factorial (credential type [3] × scenario [2]) ANOVA. The two hypotheses jointly predict an interaction between the two factors, such that: (1) for the sexist scenario, participants with a non-sexist credential will express stronger preferences for males than participants in the other two credential-type conditions; (2) for the racist scenario, participants with a non-racist credential will express stronger preferences for Whites than participants in the other two credential-type conditions (hence there are four planned contrasts). We have no prediction concerning whether the “mismatched-credential” conditions will differ from the no-credential control conditions, though differences are possible. If participants report more neutral preferences in the “mismatched-credential” conditions than in the no-credential conditions, this is evidence that credentials in a different domain can also have a licensing effect (a smaller one, compared with credentials in the same domain). We will use Tukey-corrected *p*-values to determine support for the four planned contrasts. Support for the two contrasts that compare a “mismatched-credential” condition with the no-credential condition will be determined using Bonferroni-corrected *p*-values (where a family comprises the three pairwise tests within one [sexist or racist] scenario condition). Also, the overall moral credential effect would be calculated based on the difference between the two matched-credential conditions combined and the two control conditions combined (if we found no effect of scenario type on the size of the credential effect). It is possible that one ANOVA suggests support for the moral credential effect, whereas the other suggests a failure to support the effect. We determine whether the replication is successful based on the 2 × 2 ANOVA, the analysis conducted in the original study. We would, however, evaluate replication outcomes primarily based on effect sizes rather than statistical significance, and effect sizes should not differ much between the two analyses.

To test H3 and H4, we will build a multiple linear regression model with hiring preferences as the outcome variable. To simplify the model and facilitate interpretation of results, we will only use data from those whose credential types matched their scenarios. Participants who have mismatched credentials (e.g., those assigned to the non-racist credential condition and the sexist scenario condition) will be excluded from this analysis. The predictors will include reputational concern (centered), whether one has a credential (effect-coded: 0.5 = *yes*, −0.5 = *no*), and the scenario one is presented with (effect-coded: 0.5 = *sexist*, −0.5 = *racist*), as well as their interactions.

H4 suggests that the coefficient for the interaction term of reputational concern and credential should be different from zero and positive. We do not expect an effect of scenario. As such, the coefficients for scenario and terms involving it should not be significantly different from zero. If these are observed, we build multiple linear regression models separately for those with credentials and those without. These models use reputational concern and scenario (effect-coded) to predict hiring preferences. Based on H3 and H4, we expect that (1) reputational concern negatively predicts hiring preferences in those without credentials and (2) reputational concern does not positively predict hiring preferences in those with credentials. Again, we do not expect that scenario will have any effect in either model.

Although we do not have hypotheses about the gender or ethnicity attitude measures, we will conduct the same analysis with this measure as the outcome and the same set of predictors as above to examine whether any association will emerge.

## Exploratory analyses

[To be added at Stage 2]

## Evaluating replication outcomes

Replication results will be compared with the original results with reference to LeBel et al.’s (2019) criteria for evaluating replication results.

# Discussion

[To be added at Stage 2]

## Limitations and future directions

[In Stage 2 we plan to discuss the following:

* We measured reputational concern after manipulation/random assignment to conditions. Controlling for post-treatment variables when estimating treatment effects can potentially create a bias (see Montgomery et al., 2018). This is a limitation that we decide to accept.
* Online studies can be inherently limited given evidence showing that moral licensing is more prominent under explicit observation.
* Racism and sexism can be different between the original and the replication sample (see, e.g., Eagly et al., 2020, 10.1037/amp0000494)
* Is it sensible to exclude participants who favor candidates from disadvantaged groups?
* There is a need for more direct testing of the “altering interpretation” account as the underlying mechanism for the moral credential effect.
* We will reiterate the limitation that we did not plan our sample size for the extension hypotheses.]

# References

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, *41*(4), 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*(3), 232–238. <https://doi.org/10.1027/1864-9335/a000189>

Bradley-Geist, J. C., King, E. B., Skorinko, J., Hebl, M. R., & McKenna, C. (2010). Moral credentialing by association: The importance of choice and relationship closeness. *Personality and Social Psychology Bulletin*, *36*(11), 1564–1575. <https://doi.org/10.1177/0146167210385920>

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *9*(1), 3–5. <https://doi.org/10.1177/1745691610393980>

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*, 29–42. https://doi.org/10.1038/s41562-021-01193-7

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. *Behavioral and Brain Sciences*, *41*, E124. <https://doi.org/10.1017/S0140525X18000596>

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*(7), 907–919. <https://doi.org/10.1177/0146167212442394>

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin*, *129*(3), 414–446. <https://doi.org/10.1037/0033-2909.129.3.414>

de Cremer, D., & Tyler, T. R. (2005). Am I respected or not?: Inclusion and reputation as issues in group membership. *Social Justice Research*, *18*, 121–153. <https://doi.org/10.1007/s11211-005-7366-3>

Dutton, D. G. (1971). Reactions of restaurateurs to blacks and whites violating restaurant dress requirements. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *3*(3), 298–302. <https://doi.org/10.1037/h0082272>

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. (2014). Making mountains of morality from molehills of virtue: Threat causes people to overestimate their moral credentials. *Personality and Social Psychology Bulletin*, *40*(8), 972–985. <https://doi.org/10.1177/0146167214533131>

Effron, D. A., Cameron, J. S., & Monin, B. (2009). Endorsing Obama licenses favoring Whites. *Journal of Experimental Social Psychology*, *45*(3), 590–593. <https://doi.org/10.1016/j.jesp.2009.02.001>

Effron, D. A., & Monin, B. (2010). Letting people off the hook: When do good deeds excuse transgressions? *Personality and Social Psychology Bulletin*, *36*(12), 1618–1634. <https://doi.org/10.1177/0146167210385922>

Geng, L., Cheng, X., Tang, Z., Zhou, K., & Ye, L. (2016). Can previous pro-environmental behaviours influence subsequent environmental behaviours? The licensing effect of pro-environmental behaviours. *Journal of Pacific Rim Psychology*, *10*, e9. <https://doi.org/10.1017/prp.2016.6>

Giurge, L. M., Lin, E. H.-L., & Effron, D. A. (2021). Moral credentials and the 2020 democratic presidential primary: No evidence that endorsing female candidates licenses people to favor men. *Journal of Experimental Social Psychology*, *95*, 104144. <https://doi.org/10.1016/j.jesp.2021.104144>

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

Isager, P. M. (2018). *What to replicate? Justifications of study choice from 85 replication studies*. <https://doi.org/10.5281/zenodo.1286715>

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*(1), 109–117. <https://doi.org/10.1214/ss/1177013012>

Kouchaki, M. (2011). Vicarious moral licensing: The influence of others’ past moral actions on moral behavior. *Journal of Personality and Social Psychology*, *101*(4), 702–715. <https://doi.org/10.1037/a0024552>

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>

Lacasse, K. (2019). Can’t hurt, might help: Examining the spillover effects from purposefully adopting a new pro-environmental behavior. *Environment and Behavior*, *51*(3), 259–287. <https://doi.org/10.1177/0013916517748164>

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), Article 33267. <https://doi.org/10.1525/collabra.33267>

Lalot, F., Falomir-Pichastor, J. M., & Quiamzade, A. (2018). Compensation and consistency effects in proenvironmental behaviour: The moderating role of majority and minority support for proenvironmental values. *Group Processes & Intergroup Relations*, *21*(3), 403–421. <https://doi.org/10.1177/1368430217733117>

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433–442. <https://doi.org/10.3758/s13428-016-0727-z>

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. <https://doi.org/10.3758/s13428-014-0532-5>

Mazar, N., & Zhong, C.-B. (2010). Do green products make us better people? *Psychological Science*, *21*(4), 494–498. <https://doi.org/10.1177/0956797610363538>

Meijers, M. H. C., Verlegh, P. W. J., Noordewier, M. K., & Smit, E. G. (2015). The dark side of donating: How donating may license environmentally unfriendly behavior. *Social Influence*, *10*(4), 250–263. <https://doi.org/10.1080/15534510.2015.1092468>

Merritt, A. C., Effron, D. A., Fein, S., Savitsky, K. K., Tuller, D. M., & Monin, B. (2012). The strategic pursuit of moral credentials. *Journal of Experimental Social Psychology*, *48*(3), 774–777. <https://doi.org/10.1016/j.jesp.2011.12.017>

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass*, *4*(5), 344–357. <https://doi.org/10.1111/j.1751-9004.2010.00263.x>

Miller, D. T., & Effron, D. A. (2010). Psychological license: When it is needed and how it functions. In M. P. Zanna & J. M. Olson (Eds.), *Advances in Experimental Social Psychology* (Vol. 43, pp. 115–155). Academic Press. <https://doi.org/10.1016/S0065-2601(10)43003-8>

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*(1), 33–43. <https://doi.org/10.1037/0022-3514.81.1.33>

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*(3), 760–775. <https://doi.org/10.1111/ajps.12357>

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>

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. <https://doi.org/10.1027/1864-9335/a000192>

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. <https://doi.org/10.1177/0963721414531598>

Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301. <https://doi.org/10.1016/0022-1031(77)90049-X>

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. (2022). *Observation and moral ambiguity matter: A meta-analysis on moral licensing* [Manuscript submitted for publication]. <https://osf.io/m4k6b/>

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science*, *20*(4), 523–528. <https://doi.org/10.1111/j.1467-9280.2009.02326.x>

Scheel, A. M., Schijen, M., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–12. <https://doi.org/10.1177/25152459211007467>

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>

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. <https://doi.org/10.1002/jrsm.1095>

Thai, M., Hornsey, M. J., & Barlow, F. K. (2016). Friends with moral credentials: Minority friendships reduce attributions of racism for majority group members who make conceivably racist comments. *Social Psychological and Personality Science*, *7*(3), 272–280. <https://doi.org/10.1177/1948550615624140>

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197. <https://doi.org/10.1016/j.chb.2017.08.038>

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>

Urban, J., Kohlová, M. B., & Bahník, Š. (2020). No evidence of within-domain moral licensing in the environmental domain. *Environment and Behavior*. <https://doi.org/10.1177/0013916520942604>

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*, 419–435. <https://doi.org/10.1007/BF02294384>

Wiseman, R., Watt, C., & Kornbrot, D. (2019). Registered reports: An early example and analysis. *PeerJ*, *7*, e6232. <https://doi.org/10.7717/peerj.6232>

1. This was calculated by dividing the total number of participants with the total number of effect sizes included in the Rotella et al’s (2022) meta-analysis. These effect sizes are typically derived out of a comparison between two independent groups. Therefore, the average sample size per group is 65.3. [↑](#footnote-ref-2)
2. The original study had an additional base-rate condition where participants did not read any statement. This base-rate condition was not included in Ebersole et al.’s (2016) replication. [↑](#footnote-ref-3)