Reply to PCIRR decision letter #185:
Monin and Miller (2001) replication and extension

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response as well as a tally of all the changes that were made in the manuscript. For an easier overview of all the changes made, we also provide a summary of changes.

Please note that the editor's and reviewers' comments are in bold while our answers are underneath in normal script.

A track-changes comparison of the previous submission and the revised submission can be found on: https://draftable.com/compare/yMFKXgSwREhe

A track-changes manuscript is provided with the file:
"PCIRR-RNR-Monin & Miller 2001-manuscript-v6-G-trackchanges.docx"

Summary of changes

Below we provide a table with a summary of the main changes to the manuscript and our response to the editor and reviewers:

Section	**Actions taken in the current manuscript**
General	Ed: We added exploratory questions to empirically test R4's concerns. R4: We neutralized our wording and avoided "prejudice."
Introduction	R1: We removed content on "blatant vs. ambiguous" transgressions as this is not relevant to the current investigation. R3: We included more details about the design of the original study and Study 1 in the target paper. R4: 1. We expanded the introduction to include more details about previous replications. 2. We rephrased the extension introduction so that it will not read as if we will test a mediation instead of a moderation.

Section	**Actions taken in the current manuscript**
Methods	R1: 1. We noted the deviation of scale point labels and justified it. 2. We clarified why the general attitude measures were included (they were used in the original study but were not reported. 3. We justified analyzing data using only a subset of participants (i.e., those indicating preferences for females/Blacks should be excluded). R2: 1. We specified what is included in the funneling section. 2. We elaborated what will happen if participants fail comprehension checks. 3. We specified what we meant by "U.S. participants." R3: 1. We elaborated which results will support our hypotheses. 2. We clarified how we will compensate participants. 3. We further justified why we do not aim at the smallest meta-analytic estimate in the literature and elaborated what we meant by an effect too small to be of interest. R4: 1. We added exploratory questions to examine validity concerns. 2. We made explicit that we do not plan our sample based on extension hypotheses. 3. We clarified how we will compensate participants. 4. We decided to follow the original and present the candidates' profiles all at once. 5. We applied more stringent corrections for error rate inflation.
Reporting	R3: We specified based on which test we will evaluate replication outcomes.

Note. Ed = Editor, R1/R2/R3/R4 = Reviewer 1/2/3/4

Response to Editor: Prof. Chris Chambers

> **I have now received four very helpful and high-quality reviews of your submission. Broadly, the reviewers are enthusiastic about this proposal (with one notable exception that I will expand upon below) while also raising a range of specific concerns. Some of the main issues include considering (and possibly increasing) the statistical power of the extension hypotheses, making clear the consequences of failed comprehension checks, reconsidering the (over)stringency of the exclusion criteria to ensure a close replication, addressing risk of bias and potentially invalid conclusions in the proposed application of some exclusion criteria, considering differences in participant populations as a potential deviation (while also justifying other specific deviations), and considering an in-person study alongside the online study.**

Many thanks for obtaining the reviews and inviting a resubmission. Below we respond to the reviewers' comments. We benefited greatly from their constructive feedback. We hope that through addressing their comments, we have made our manuscript stronger to meet the threshold for recommendation at PCI-RR.

> **Of the four reviewers, three judge the replication to be of sufficient value to meet Stage 1 criteria 1A and 1B. However, one reviewer (Meyers) presents a strong critique of the scientific validity of the original study, to such an extent that he is "highly doubtful that this work would provide anything valuable". This is a very interesting point that I feel I need to address specifically in this decision letter. You may be aware that at a small number of journals, there is a modified version of Registered Reports called Accountable Replications in which the journal commits to publishing methodological close replication attempts regardless of the validity of the original design (e.g., see policy at Royal Society Open Science). PCI RR, however, does not automatically consider a RR proposal to be valid just because it replicates a previous published study, which is to say that major flaws in the original work can be grounds for rejecting a Stage 1 submission.**

> **In this particular case, I see several possible ways forward: (1) to rebut the reviewer's argument if you believe the original methods are in fact sound (and to provide a summarised version of this rebuttal in the revised manuscript as well as the response to reviewers); (2) to add an additional study that corrects the identified flaw; (3) to keep the current study the same but acknowledge the flaw in the Introduction and (eventually) the Stage 2 Discussion, providing a strong justification in the Introduction for why the study remains important to replicate regardless. I will leave you to consider these possibilities (and potentially add any that may not have occurred to me). In one way or another, the matter must be addressed head on.**

We are grateful for your clear suggestions on how to move forward. We did a mixture of what you suggested.

First, we indicated that we do not fully agree with the reviewer and provide the reasons as to why in this response letter below. You indicated that some reviewers saw merit and value in such a replication, and it is likely that there are others who would share that view.

Second, we added a few exploratory questions in our study in the hope that we can obtain some preliminary evidence addressing some of the issues the reviewer raised.

We tried to refrain from opening a debate on the point of whether the original design has merit or not, and instead to try to accommodate their suggestions and use that to maximize insights we can gain from the replication study.

Response to Reviewer #1: Prof./Dr. Corey Cusimano

> **The authors propose a replication and extension of Monin & Miller (2001) Study 2. I think that this is a valuable project and that the team that proposed the project is ideally suited for carrying it out. In particular, I thought that the literature review was thorough and that the motivating rationale for the replication was convincing. Moreover, I think that the proposed sample size is sufficient. And I think that the deviations from the original protocol are (for the most part, see below) inconsequential and well-motivated (e.g., using higher resolution photos of candidates).**

Thank you for your constructive and valuable feedback.

> **I had a few concerns that I hope the authors can address:**
>
> **(1) One deviation from the original study is the move from lab study to online participants. There are two obvious issues that this raises. First, there is greater psychological distance between the participant and experimenter, then lowering the demand for socially desirable responding and reputation management. This may lead to a failure to replicate the original study if, as previous work has shown, moral licensing effects are weaker in online studies (Rotella et al., 2022). The authors discuss this possibility in their rationale for including an individual difference measure of reputational concern as part of their extension. But what could we conclude in the absence of a replication and an absence of a correlation with the individual difference measure? I think "not much" as there are many reasons (including unreliable measurement and ceiling/floor effects on the effect) why you'd fail to find a difference despite a true effect.**
>
> **Second, online participants have much more experience doing experimental studies, and very likely, doing studies related to race/prejudice. This familiarity may affect how they respond. This is an obvious and well-trodden worry in the literature (but that doesn't mean it isn't relevant to the current investigation). One potential extension would be to include a protocol that measures awareness of the experiment's purpose. If participants by-and-large seem unaware of the study's goal, then I think such a result would obviate worries about the context of online studies substantively deviating from in-person studies. (Note that I do not feel strongly about this suggestion in particular. I am just trying to get the**

> **authors to think more about what a failure to replicate the results would mean in light of the shift to an online context.)**
>
> **With these two worries in mind, I think that the authors should consider doing an in-person replication. Such a replication would not have to include the extensions nor be as highly powered to be meaningful. Indeed, just 2.5x the original study would be informative (Simonsohn, 2015). I still think that the current project is worthwhile, but an in-person replication would be more valuable in the specific event that the authors fail to replicate the original study.**

We respond to the above comments together, as these responses are related.

First, generally speaking, we believe that failed replication attempts provide equally valuable information as successful ones do. In the current case, a failure to replicate the original result with an online study is not less valuable than a success in replicating it without setting change; both outcomes are valuable data points for future meta-analyses. Also, if this replication fails to support the original finding, it further highlights the need of replicating the original experiment (and probably many other moral licensing experiments) in person rather than online. This is an important insight to draw.

Second, we agree that an in-person replication would be desirable. But we do not believe it is necessary for this proposal (we therefore appreciate that you suggested rather than demanded one). The community will surely appreciate a well powered and executed in-person study of moral licensing. By comparing such a study with an online equivalent, we can examine Rotella et al.'s (2022) claim that observation/reputational concern is the main driver of the licensing effect.

These studies, however, need not be done by the same researchers. For one thing, as a small team of insufficiently funded ECRs, we cannot commit to conducting both studies in one RR. For another, we believe the replicability of any phenomenon is better addressed incrementally: we contribute data from a large online sample; based on our results, future replicators can better decide if they should start directly with an in-person replication study. The literature would advance faster if replication efforts are crowdsourced.

We thank you for the suggestion to assess participants' awareness of the study's purpose. We have one question asking just this towards the end. We are not sure, however, whether it will help us distinguish naive and experienced participants. After all, experience with prejudice/race studies might not play a substantial role. In Ebersole et al.'s (2016) large-scale replication of Study 1 in Monin and Miller (2001; used the same sexist scenario as Study 2 and in our proposal), the effect sizes were very close for MTurk participants and for those from university participant pools (*d* = 0.17 vs. 0.15).

> **(2) The authors report the primary DV in the following way:**
> **"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; we used only positive numbers for the scale points to avoid any potential bias, though preferences for females/Blacks were coded as negative values)."**
>
> **The original study used negative numbers in their scale labels. Thus, this change to showing participants only positive numbers represents a (subtle) deviation from the original design. This deviation is not reported (see page 20).**
>
> **I do not think the authors should deviate from the original! I don't know what bias could be caused by having negative numbers, and including "0" as a midpoint is more intuitive than "4".**

Thank you for catching this. Indeed, we did not note this deviation, which we now do.

We decided to deviate because we thought representing preferences for Black people as minus (and preferences for White people as a plus) can potentially bother some to the extent of upsetting them. This can be easily addressed by showing only positive values. Since this is subtle and addresses what seems like an increasingly sensitive topic in the US, we would rather keep it. In the planned discussion section to be completed in Stage 2 we added a planned discussion of this deviations as a limitation.

We are willing to revert to the original setup if given clear editorial guidelines.

> **(3) The authors include a measure of explicit prejudice but do not justify their doing so. On page 19, the authors indicate that the last thing participants will do is indicate whether they agree with the following statement: "Women are just as able as men to do any kind of job" or "Blacks are just as able as Whites to do any kind of job" (7-point scale: −3 = disagree strongly, −2 = disagree, −1 = disagree slightly, 0 = neither agree nor disagree, 1 = agree slightly, 2 = agree, 3 = agree strongly). Why are they including this? These measures do not show up in the original study nor in their planned analyses. This is another planned extension? I recommend the authors remove this measure or say much more to justify it.**

These measures were included in the original study, but were not reported in the published article. We noticed this when we went through the original materials we received from the original authors. These materials have been posted to OSF.

We made this point clearer in our revised manuscript:

> "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). These measures were not reported in the original published article but were included in the study materials. We included them here to have a faithful replication. We call this dependent measure gender/ethnicity attitude henceforth."

> **(4) Removing participants from the reputational concern moderation analysis is unjustified. On page 24 the authors write, "participants who favor females/Blacks in the sexist/racist scenarios will be excluded from this analysis."**
>
> **How does this aid interpretation? The impact of credentialing on prejudice**

> **may (in theory) occur at any point on the scale of the DV. Likewise, the impact on reputation on the main DV may occur at any point of the scale. So I do not see how removing a portion of participants aids interpretation. At best, the only downside is that it reduces power. But this is still a terrible downside (especially if, for instance, the authors find a null effect that they now have to interpret) and there are potentially other downsides too. The authors should not pre-register this analysis.**

We respectfully disagree with you on this point. It is *necessary* to remove those participants who favored women/Blacks in these scenarios, even for the replication part. This is because this study assumes that stronger preferences for males/Whites are more morally problematic/more prejudicial (but can be licensed with credentials), but does not assume that stronger preferences for females/Blacks are less so, compared with a neutral preference. But this is what's assumed by feeding all data from the gender/ethnicity preference measure to an ANOVA.

To illustrate, a slight preference for females is coded as minus one, whereas a neutral preference is coded as zero. It is an open question whether each participant perceives the former to be more moral than the latter. However, we would impose that perception on all participants by including those who prefer females in the analysis and interpreting a higher mean of the credential group over the no-credential group as evidence for the credential effect. We do not believe that perception is universal, or even shared by the majority.

We have further justified our decision in the revised manuscript:

> "We conducted confirmatory analyses both with and without those participants who indicated a preference for females/Blacks in the respective scenarios (whenever the gender/ethnicity preference variable 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/Whites in the respective scenarios can be perceived as 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/Blacks are less problematic, or more moral, compared with a neutral preference and preferences for males/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 they appear along the entire scale. As such, removing those participants is necessary. We will report results without those participants in the main manuscript (and with them, in the supplemental materials, if the results differ substantially)."

On a side note, it is unclear whether removing those participants will make a big difference. It would seem that there are not many people who indicate a preference for females or for Blacks, after all (e.g., only around 4.7% of the 3,134 participants in the Many Labs 3 replication of Monin and Miller's [2001] Study 1 expressed a preference for females).

> **(5) The authors discuss the issue of 'blatant' versus 'ambiguous' transgressions. But the conclusion of this discussion confused me. Do they think that their study is addressing this confusion in the literature because they are testing 'blatant vs ambiguous' transgressions? Or do they think that, given the confusion, there should be more work understanding how strong the effect is in ambiguous situations? I just didn't get it, and thought they could clarify this section.**

Thank you, appreciated!

We agree that introducing the "blatant vs. ambiguous" distinction could be confusing, as this replication study does not involve blatant transgressions. We removed relevant sections.

Response to Reviewer #2: Prof./Dr. Marek Vranka

> **It was my pleasure to review this Stage 1 RR. Overall, I find all materials well prepared and authors' writing easy to follow. The rationale behind the replication study is clearly stated and its importance is well-documented by citing the popularity of the original findings, low power of existing studies, meta-analytical evidence suggesting overestimation of the effect size, and mixed results of already conducted direct and conceptual replications. The target study of this replication is well-chosen as it was not previously replicated and it also allows for an elegant extension, namely to explore the effect of "mixed-credentials". The authors propose an additional extension, that is to explore the effect of the "reputational concern" trait. Neither extension interferes with the design of the replicated study, so it can still be considered a very close replication. Ethical concerns are adequately considered.**

Thank you for the positive opening note and your feedback.

> **H1 derives directly from the original study, H2 is related to the first extension and H3 and H4 relates to the second extension. All hypotheses are clearly stated and mapped to sensible, clearly defined research questions. Preregistered statistical tests are suitable for testing of the hypotheses and interpretation of the results is unambiguous, as the authors explicitly describe the results that would support their predictions.**
>
> **In the analysis, there is a possible issue, because the authors plan to first test hypotheses H1 and H2 separately, using two 2x2 ANOVAs, but then also test them jointly using one 3x2 ANOVA. This could in theory lead to a situation in which the results of these two approaches are mismatched and it is not clear what would be the conclusion in such a case.**

Good point, we appreciate the note.

We would evaluate whether the replication is successful based on the outcome of the analysis that follows the original's.

We noted this in the revised manuscript:

> "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."

> **The sample size calculation is reasonable and even though one would ideally prefer larger sample size, the available funding is understandably not unlimited. For the direct replication, the study has sufficient power to detect an effect d = 0.25, which is roughly half of the effect size reported in the original study.**
>
> **The first extension is likely looking for a smaller effect (the difference between the effects of matched and mismatched credentials) and thus the test could be underpowered. Similarly for the second extension, when only a part of the sample (those not favoring the minority candidate) will be used and an interaction effect is tested (for H4). The RR can thus be strengthened by adding more detailed discussion of the statistical power for the extensions of the original study.**

Our main aim is the replication with added exploratory extensions. We decide not to consider power for the exploratory extension hypotheses in this study, and our initial findings regarding the extension can be used for power analyses in future research examining the phenomenon.

> **The methodological part is written in a great detail, providing enough information for any possible future replications. In this regard, only a few minor points:**
>
> **a) In the "Participants" section (p. 16), it is not clear that only participants from and currently in the U.S. will take part in the study, as evident from the exclusion criterion (p. S12). Moreover, the exclusion criterion itself ("Participants who are not from or currently in the U.S.") could be rewritten to make clear whether only one or both conditions must be satisfied.**

Thank you for the suggestion. We added the following to the "Participants" subsection:

> "In our recruitment, we indicated that we are looking for participants that were born and currently living in the U.S."

> **b) It is not clear what will happen when participants fail the comprehension checks (p. 19). Will the survey end or will they be able to examine the scenario and attempt to answer correctly for a second time?**

The check questions are on the same page as the scenario. Participants have to complete those correctly in order to be able to proceed to answering the dependent variables. Meaning, that who fail to answer correctly will stay on the page, and may again attempt the questions as many times as they would like until they finally pass the checks. We have now clarified this in the revised manuscript.

> **c) In the description of the funneling section at the end of the questionnaire, the item about previous encounters with the materials used in the study is omitted (p. 20). It could be mentioned for the sake of completion (but see point b) in the section below.**

Thank you for the suggestion. We made our description more comprehensive.

> **Some additional questions and suggestions:**
>
> **a) It is possible that measuring the trait reputational concern after the experimental manipulation may bias the analysis of H3 and H4? (As participants may answer differently depending on whether they have matched, mismatched or no credentials in the previous study; see e.g., Montgomery et al., 2018). I understand that it is not feasible to measure it before the experiment in order to keep the design close to the original study. Maybe it will be possible to contact participants again and have them fill the questionnaire (this could be easy for example with Prolific, but probably also with MTurk).**

We appreciate your suggestion. We understand the potential issues of controlling for a post-treatment variable (as discussed in Montgomery et al., 2018). Nonetheless, we decide to keep the current design and accept the potential bias as a limitation. We aim primarily at replication, and with the extensions, we only hope to provide some preliminary evidence for future follow-up confirmatory studies. Therefore, we would prefer to try and simplify the design, avoid unneeded complexities with multiple wave data collections, and keep our investigation focused on the replication repeating the target's design as closely as possible.

Also, we note that though having follow-up surveys partly addresses the issue of conditioning on post-treatment variables, it creates other new issues, such as drop-outs and selection bias (on those who choose to re-enroll in the study). It is hard to anticipate the impact of either and determine which has a bigger influence.

To address this point, we added a planned discussion of this limitation in the Discussion section.

> **b) One of the proposed exclusion criteria (no. 4, p. S12) is "Participants who indicate that they have seen or done similar surveys". I am worried that the question asking about this is too vague and can be answered yes by anyone who has ever taken part in a study in which they selected a job applicant. As far as I know, this criterion was not used in the original study**

> **and there are no strong theoretical reasons for why previous participation in a similar study should matter. I recommend either making the question specifically about study with this scenario, or drop the criterion completely.**

Thank you for your suggestion.

We reworded the question to be “have seen or completed surveys involving similar scenarios.” In any case, we would analyze data before and after exclusion and share our data. Those who are interested in the results without the application of this exclusion criterion will have access to them regardless of whether we drop this criterion or not.

Response to Reviewer #3: Prof./Dr. Štěpán Bahník

[Disclosure: This reviewer and the corresponding author have previously discussed a potential collaboration, which at the end did not mature beyond the initial discussions. We felt it important to include this here, especially since for a period of time there was mention of the reviewer being listed as a member of a team coordinated by the corresponding author.]

> **The registered report aims to replicate Study 2 from Monin and Miller (2001), which is an influential paper in the moral licensing literature. The proposed study also aims to extend the original study using additional two conditions which will examine domain-specificity of moral licensing. I do not follow the literature on moral licensing closely, but in my opinion the manuscript describes the literature well and the overview of the existing studies provides sufficient justification for the need of the proposed replication.**

Thank you for the positive and constructive feedback.

> **Given that the planned study largely uses methods of the original study, I do not have many comments regarding the methods. Most of my comments are rather minor and should be fairly straightforward to address:**
>
> **1) The paper mentions an existing replication of Study 1 from Monin and Miller (2001), but does not go into details of the original study and the replication. It might be good to briefly describe the original Study 1 and how it differs from Study 2 the registered report aims to replicate (pp. 12-13).**

Thank you for the suggestion. We provided more details in the updated manuscript. In particular:

> “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 1,300 citations as of December 2022 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 replicated; for those that were replicated, there are 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. 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 that study 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 their corresponding controls.”

> **2) I do not think that resources of an average lab are relevant to the justification of the sample size. I also do not believe that limited potential impact can be directly concluded from a small effect size. A small effect can have a large impact, for example, if the phenomenon occurs frequently or if the effect might be larger in the real-world (p. 15).**

We agree that a small effect can have a large impact if it happens frequently in real life. However, effects have their contexts. Here, we are interested in an effect in an experiment, induced with a simple decision and examined with a text-based scenario. If moral licensing is to be studied this way (as it has been in many other papers), it needs to be large enough so that the average lab can afford to study it with enough power. Otherwise, it becomes prohibitive and less relevant.

We believe it makes sense to justify a target effect size citing reasonable resource constraints. We revised our manuscript to clarify our points (section Sample size planning):

> “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.”

3) The method section is said to be written in past tense, but the beginning of the participants subsection is in future tense (p. 16).

Thank you for catching that. We fixed this discrepancy.

4) The federal wage is written to be “7.25USD/hour, per minute”. The “per minute” is probably a mistake. It should also be mentioned that it is “U.S.” federal wage.

We revised and clarified it.

> **5) The difference in participant populations is not mentioned in deviations. The original study was conducted with undergraduate students more than 20 year ago, so it is possible that racist and sexist attitudes might differ between the original and replication samples.**

Thank you for pointing this out.

We now address this deviation in the revised manuscript. It is possible that the racist and sexist attitudes are different between the original and the replication sample (see, e.g., Eagly et al., 2020, 10.1037/amp0000494). It is, however, hard to say what impact this would have on the replication outcome. For instance, does the decrease in expressed prejudice reflect a genuine reduction in prejudice or a stronger pressure to be politically correct?

We added a planned discussion of this point in Stage 2.

> **6) It is not clear from H1 (and also H3 and H4) whether it relates to all credentials or only to the same-domain credentials. The hypotheses seem to include all credentials, but the analysis later uses only the same-domain credentials.**

We aim primarily at a replication. As such, testing of H1 will only include the same-domain conditions. We do not plan to include the mismatched-domain conditions in testing H3 and H4 as part of our confirmatory analyses. This is because we do not know clearly how large the credential effect would be in these conditions. Certainly, these conditions could be included for exploration if there turn out to be appreciable credential effects. We have made further clarifications about this point in our revised manuscript.

> **7) H4 is phrased in such a way that it seems to specifically predict that H3 holds and the relationship between reputational concern and prejudice is just smaller in size with moral credentials. However, the analysis just looks at whether the interaction is positive, which is also consistent, for example, with no relationship of reputational concern and prejudice in the condition without credentials and a positive relationship between the two variables in the condition with credentials.**

Thank you for the point. We further specified what our hypotheses would predict, as below (in Hypotheses section):

> H3: Trait reputational concern would be negatively correlated with preferences for males/Whites in those who have no moral credentials.
>
> H4: Non-sexist/non-racist moral credentials would attenuate the negative correlation between trait reputational concern and preferences for males/Whites (H3).

> 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). 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.

> **8) On p. 23 it is mentioned that there is no prediction with regards to the difference between different-domain credentials and no credentials. It is okay to not have a prediction, but it seems important to note that this tests whether cross-domain licensing also works. This cannot be determined by the test of H2 alone, because if the effect predicted by H2 holds, it is still possible both that cross-domain licensing has no effect and that it has an effect, and it is just smaller.**

Thank you for raising this. We further noted this point in our planned analysis:

> “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).”

> **9) Why are H1 and H2 tested once separately and once together? I am not sure whether the two analyses are mathematically equivalent, but if they are, then they are redundant, and if they are not, then it should be mentioned which is the focal test of the hypothesis.**

H1 and H2 can be tested together, but because we aim at replication, we decide to first follow the original’s analysis, which tests *only* H1. This is the focal test for the *replication*. But since we have H2, and we can test the two hypotheses with one ANOVA, we also planned this broader analysis.

The results of these two tests will likely differ a bit, but they will not be much different. We prefer to include both analyses, though some of them could be put into the supplementary (which can be determined at Stage 2 after the results come in).

In the revised manuscript, we now mention that we will evaluate replication outcomes based on the same analysis as in the original study:

> “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.”

Response to Reviewer #4: Prof./Dr. Ethan Meyers

> **A big hello to everyone who is reading this. I have not reviewed a registered report before. Instead of trying to implement my typical format of reviewing, I decided to follow the guidelines provided by the PCI RR. Aside from an initial major point that I raise below, my review will comprise of my answers to the "Key issues to consider at Stage 1". I recognize that this will make responding to my letter a bit more difficult. But I trust that the authors will be able to figure out a reasonable solution.**

Thank you for your comprehensive and constructive review.

> **I think the work only suffers from one major flaw. Study 2 from Monin and Miller (2001), is not a conceptually strong (or methodologically strong) study. If one aim of the present work is to further our understanding of moral licensing effects, then I think the authors ought to select an experiment that more clearly assesses moral licensing or fix the issues that I raise below. Even though the authors want to specifically offer a replication of a seminal study of moral licensing, there remain plenty of other suitable choices. As provided by the authors, the definition of moral licensing is "when 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." Based on this there are two required elements to moral licensing, (1) a clear initial moral act and, (2) a subsequent immoral act. In my view, the decision vignettes lack in both elements.**
>
> **The proposed experiment features a hiring task and then a job description with an opinion question. The hiring task, supposedly the source of the initial moral act, asks participants who they should hire out of 20 possible candidates. In each condition there are 19 white and male applicants and one star applicant who is either a black male, a white female, or a white male. The idea is choosing to hire the best candidate in each condition might afford a moral license to those who hired the black male or the white female but not to the person hiring the white male. This is because it might be seen as being "anti-sexist" for hiring the female or "anti-racist" for hiring the black male. These moral credentials seem vaguely possible, but they don't make much sense. If your goal was to hire the best person for the job and the person also happened to be black, in what world are you being anti-racist? Hiring a worse candidate because they are white and not black would be an expression of racism. Hiring the best candidate who happens to be black because they are the best candidate is not an**

> **expression of anti-racism. Hiring a worse candidate because they are black and not white might be an expression of anti-racism as it is argued (I would argue it is still just racism) and thus might be the only scenario in which I see an "anti-racist" credential being possible. To summarize, it is not clear to me what moral credential the average person might obtain from the anti-racist condition.**

Thank you for these in-depth comments on the materials.

We appreciate the foundational critique of the literature and the methods employed, and yet we think it necessary to focus the evaluation of the replication on our ability to closely reproduce and test the replicability of the findings, rather than the methods of the target. Many of the points raised here go far beyond the scope of the replication. We aim at establishing if the findings reported using their methods (which were for the most part taken by a vast literature as support for that phenomenon with such an interpretation) can be supported again, and to what extent. We do hope that regardless of our findings, our revisiting and reporting on this phenomenon would reignite a discussion about the methods with suggested ways of improving and advancing both theory and methodology on this phenomenon.

Per this specific point, though we cannot speak on behalf of the original authors, the materials may not be inherently inappropriate for testing moral licensing, or specifically, the moral credential effect.

First, it is not clear that obtaining moral credentials requires a "clear initial moral act." In fact, the extent to which the initial act is clearly moral (or the *ambiguity* of this act) has been proposed and studied as a moderator of the licensing effect (see Mullen & Monin, 2016, Table 2). There are many studies in this literature that do not involve a "clearly initial moral act." For instance, in Effron et al. (2009; *J. Exp. Soc. Psychol.*), an initial expression of endorsement of Barack Obama made participants favor Whites over Blacks. We do not think endorsing Obama is a clear moral act (though some may have a different idea, and that is why ambiguity matters). Overall, researchers consider how clearly moral the initial act is as a moderator, but do not take a clearly moral act as a prerequisite, for moral licensing.

Second, certainly, hiring a non-White because the person is the best candidate for the job is not an expression of anti-racism (and if we stated anything like that, it was our mistake; we have carefully revised the manuscript to try and ensure we are not making claims that suggest that, please help us to identify possible oversights, if you see any). People would not see this hirer as anti-racist just because of this hiring decision. We agree with you on this point. However, whether this decision expresses anti-racism is not what matters for the credential effect to take place. What matters, based on our understanding of the literature, is how the hirer themself and others will interpret this decision when the hirer intends to take a questionable move at a later point of time. When the hirer wants to say that they prefer a White for a job with a hostile working environment towards Blacks (as in this study), they fear that this preference would appear racist. Nonetheless, believing they have already demonstrated their non-racism with the decision to hire a Black (and hoping others would also interpret the decision as showing such) may make them feel more comfortable in expressing that preference.

Also note that the hiring decision is not aimed at establishing a credential of "anti-racism," but rather of "non-racism.", and though it may seem subtle there are major differences. This may have led to the (mis)understanding that the decision is not capable of providing the credentials in question (again, if we said anything like that in the last submission, it was a mistake and should be corrected, please alert us to possible oversights). We rechecked the manuscript to try and ensure that we are clear on this point.

(Another minor point: in the initial hiring scenario participants review 5, but not 20, applicants.)

> **My thoughts and arguments are identical for the anti-sexism condition – hiring a woman who is the best candidate is not anti-sexism, it is hiring the best candidate. Hiring a worse man instead of a better woman is sexism, as is hiring a worse woman instead of a better man (sometimes argued to be "anti-sexist"). So, it is not clear to me what moral credential the average person might obtain from the anti-sexist condition.**

Our reply to this comment would be similar as above. Also, what is established is a "non-sexist" credential rather than an "anti-sexist" credential.

> **To further complicate matters, it also isn't clear to me how "hiring the best person" might not count as a moral credential. A person might value meritocracy highly. In this case, the moral act is to hire the absolute best person regardless of what they look like. In this sense, it shouldn't matter to me whether the person I hired is a female or black; if I hired the best, I've done the right thing. Under this view, couldn't each condition be granted a moral license? Unless this can be ruled out, I don't think a license vs. no license comparison (the primary test of the replication) would be meaningful or possible.**

We see the point regarding "hiring the best person" possibly providing a "meritocratic" credential. Yet, it is a credential that participants in all conditions obtain. We believe we are not comparing license vs. no license, where no license means the complete absence of any license. Instead, the "no license" here would mean "no non-sexist license," which only those in the non-sexist credential condition obtain.

We do, however, recognize that we might not entirely follow the full scope of this comment and therefore possibly miss part of the point made. We would appreciate it if you could further elaborate on this point, preferably with potential constructive ways of amending or addressing it in a revision.

> **So, participants complete the hiring task, possibly picking up some sort of vague credential, maybe not, or maybe they are feeling extra moral for having hired the best person. They are then presented with another hiring scenario. This one is a bit different, however. Instead of hiring a candidate, participants are provided lots of details about a job (e.g., that it requires exuding confidence) and are told that they have already hired someone. They are then asked the key question: whether the job is particularly suited to one ethnicity or gender (scenario dependent). Unlike the first part of the task where participants had to actively select the individual candidate they wished to hire, in this portion of the task they are expressing a group level preference. It's already a stretch to label answering this question a "behavior", but to claim it is an expression of prejudice makes little sense to me. Allow me to elaborate. Consider the question asked in the anti-racist condition:**
>
> **"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?"**
>
> **It isn't clear to me how a belief that any single ethnicity might be better suited for this role would necessarily be prejudiced. First and foremost, what is not being asked for here is some sort of individual-level evaluation. There is no set of applicants one must choose between, weighing all sorts of factors. Instead, it's simply asking whether the person expects any sort of differences between any two possible ethnicities. The comparisons are endless! Moreover, it seems completely plausible that a person might think that ethnicity imperfectly tracks culture. Cultures differ and thus produce people of different values, some of which will be more or less suited for any occupation. Answering something akin to "it seems possible" or "it might" or "it could" despite holding this perspective will be labeled by the authors as prejudiced. This makes little sense to me. Especially if we consider that prejudice means an opinion not based on reason or actual experience, it's not clear to me how one could even begin to infer prejudice without making an earnest attempt to seriously understand the reasoning or experiences of each individual participant.**
>
> **The anti-sexism case is similar. The participant is told that one needs to exude confidence in the job (among other things) and then are asked whether they think if broadly speaking, one gender might be more suited**

> **for this role. If I hold the belief that men are a bit more confident (and especially overconfident) than women, why wouldn't I think that on a group level men might be more suited for this job? Believing this would certainly never exclude you from hiring a woman. Indeed, I could simultaneously endorse the truth that men tend to be a bit more confident AND hire only women for the position, because, at the level of the individual, where hiring decisions actually take place, I am selecting for the best candidate as they appear. Despite group averages, individual women can be extremely confident and likewise, individual men can be extremely unconfident.**

[First, a quick clarification. You noted:

> "instead of hiring a candidate, participants are provided lots of details about a job (e.g., that it requires exuding confidence) and are told that they have already hired someone."

This is not the case. They are not told that they have already hired someone.]

Thank you for raising all these concerns. We try and remain neutral on the target's design and whether these are crucial concerns or not, yet we would like to try and address potential concerns, if possible.

The starting point of how we understood the scenarios was that even though, as you said, preferring males in general for a job that requires exuding confidence does not mean one is prejudiced, others could see this preference as such.

We suggest that we take these concerns (and ones you raised above about the first hiring scenario) to empirical tests. For this purpose, we included several exploratory questions at the end asking how participants perceive the different behavioral options and preferences in the scenarios. This is not a perfect test, yet we hope that this way we can at least get a sense of whether, and to what extent, these concerns materialize:

> "**Exploratory questions**. One reviewer for this Registered Report at Stage 1 raised concerns over the appropriateness of the manipulation in providing 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 over gender or ethnicity. The reviewer also questioned whether participants—with or without credentials—would find it prejudicial to prefer males/Whites in the gender/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 study design. 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 likely sexist/racist, 5 = very unlikely sexist/racist; only the endpoints are labeled; 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, 5 = strongly agree; only the endpoints are labeled). Therefore, there were three evaluations for the decision to hire each candidate.
>
> 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, and with only texts but not numbers) 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 make what was manipulated obvious. Also, to keep the replication part intact, we had to place all these exploratory questions to the very end, though it might be better 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."

We also agree with you that it is probably inappropriate to call the preferences outright as "prejudiced." That is a very good point, and much appreciated. We therefore worked to neutralize the wording throughout the manuscript. We now call the DVs gender or ethnicity preferences.

> **To summarize: Moral licensing first requires a moral act and then requires an immoral one. In the scenarios to be presented, the original moral act isn't clear, and neither is the immoral one.**
>
> **I hope these arguments illustrate the conceptual problems with Study 2 of Monin and Miller (2001). To be perfectly clear on my position: I don't think any researcher should consider Study 2 to be evidence of anything.**

> **Meta analyses that have included this study should promptly exclude it, and the rest of the included studies should have similar criteria applied and evaluated for.**
>
> **In my view it is not a requirement of the authors to "fix" the problems of the original work. However, without doing so, I would be highly doubtful that this work would provide anything valuable. This would also apply to potential extensions on this flawed work. I think the authors have two options. One is to attempt to resolve these confounds and present an even stronger test than was originally offered. The other is to simply pick a different experiment to replicate.**

In response, we clarified that the moral credential effect requires neither an initial act that is clearly moral nor a subsequent act that is clearly immoral. Clarity should be treated as a moderator rather than a prerequisite. We share with your opinion that it is not up to replicators to fix problems with the original work. Although we are not fully convinced of all the issues that you brought up, we are very happy to modify the design further so that there is some chance to address the concerns empirically and provide insights.

> **The researchers do a good job in setting up the importance for replicating moral licensing work. Despite several replication attempts already undertaken, the number of recent meta-analyses on the subject suggest that there is a debate in the literature large enough to warrant further replication attempts. Overall, I felt convinced that replicating moral licensing work would be worthwhile if it could be used to help determine whether the effect is real or not.**
>
> **I found the proposed extensions both clear and unclear and raise a potential concern about each of them. First, the reputational concern hypothesis is very sensible. The authors did a great job in pointing out the role that reputational concern could play (and consulted the work of Rotella et al.) in this effect and of highlighting the importance of studying it further. However, I'm afraid they might have done "too good a job" at this. I can't help but wonder if manipulating reputational exposure rather than measuring reputational concern would be a much better test of this hypothesis.**
>
> **According to the work of Rotella et al., studies with explicit observation produced larger effects than those with only some or no observation. I would think the most natural extension of this finding is to then manipulate whether participants are being observed (or think they are being observed) or not. The correlation that the authors are proposing would indeed help to**

> **answer this question. But we tend to think of experiments as stronger tests of mechanisms than observational studies. In my opinion that certainly applies here. I also don't think running a study of this variety would suffer from many practical (especially resource) constraints. For instance, the authors are already planning to recruit 350 participants to assess the correlation anyways. These participants could be used for the experiment instead. This would also allow you to run fewer participants in the replication experiment as the moral licensing task would remain across both experiments (so you could combine samples for the strongest estimate of the effect).**

We agree that an experimental study manipulating reputational concern would be valuable. Yet we think it best to take an incremental approach, and address each step separately, focusing this investigation on replicability with minor exploratory extensions. Given our focus on the replication, and the limited resources available to us, we prioritize the proposed study with an online sample. Also, experimentally manipulating observation requires a large number of offline participants that we do not have access to. We can only make use of what we have, so we decide that a well powered replication with online samples would be the best we can contribute to the literature.

> **Second, I am not completely convinced by the rationale of the domain extension. While I understand the "racist" and "sexist" domain-specific moral credentials idea, I wonder how the rationale could not also apply to a "hiring decision" domain. That is, the first judgment takes place in a hiring situation and so does the second. Could it be the case that each decision in this experiment takes place in a "hiring domain"? I think it's plausible. Because it is plausible the authors risk finding no effect of domain based on their definition despite there being a real effect of domain but at level other than what the authors were considering. To remedy this problem, the domains should be made further apart to minimize any alternative explanations for the observed effects. Without this, I am not sure the domain test is convincing enough, which might call into question the value of this extension as proposed.**

If our understanding is correct, you suggest that if we find no support for our hypothesis regarding the domain extension (i.e., we find that a non-sexist credential is as effective in licensing potentially racist behaviors as a non-racist credential), this could be because participants actually considered themselves to have done something good in the "hiring domain." With a "good-hirer credential," they feel more comfortable stating a hiring preference that may otherwise be perceived as problematic, regardless of whether this preference has to do with ethnicity or gender.

This is a valid point. But we do not see it as a problem to remedy. How far apart the domains need to be to produce a domain effect on the size of moral licensing is an empirical question, one that should be solved by aggregating and meta-analyzing a host of studies examining it. In our reply to Reviewer #1, we suggested that a null effect does not imply that the effect is non-existent, and a single study is not likely to yield conclusive evidence (regarding whether domain plays a role or not). Since there is not much evidence at present, we do feel that the domain extension would be contributive.

> **I also wonder the extent to which the domain hypothesis is already tested in the original experiment/is observable in the planned analyses of the replication attempt. Given that the test is a 2 Credential (Yes/No) x 2 Scenario (sexism/racism), wouldn't a significant 2-way interaction imply an effect of domain? Similarly, wouldn't a null interaction imply that if there is an effect of domain that these two domains aren't far enough apart? I could be wrong in my reasoning here.**

The "domain hypothesis" here is that a moral credential from the same domain as the morally questionable behavior to be licensed has a larger licensing effect than a credential from a different domain. For instance, a non-sexist credential is more effective in licensing a potentially sexist behavior than a non-racist credential.

This hypothesis was not/could not be tested in the original study, because participants there always obtained moral credentials in the same domain as the potentially problematic behaviors. There, those assigned to read the sexist scenario could only obtain a non-sexist credential. A null interaction in the original study meant that a *non-sexist* credential is as effective in licensing potentially *sexist* behaviors as a *non-racist* credential is in licensing potential *racist* behaviors. It says nothing about, for instance, whether a non-racist credential is as effective as a non-sexist credential in licensing potentially sexist behaviors (we predict that it would be less effective). This is what we plan to test with the domain extension here.

> **Given the details provided by the authors, I believe that I could run their proposed experiment and analyses right now if they would be willing to fund it ;) (and I wouldn't necessarily call myself an expert).**
>
> **For the replication hypothesis, the authors state their evaluation of a non-replication will follow LeBel et al. (2019)'s criteria. So, aside from any gray area this contains, yes. In terms of the rest of their hypotheses the authors do not seem to have indicated how they will interpret different results. For the most part the interpretation seems to be clear when the results are in line with their prediction. However, it is not clear to me (as I raised above**

> **in the domain case) how they might handle null results here. Some further detail would be appreciated.**

We believe that there are multiple possible interpretations (small effect size, weak manipulation, sample characteristics) for null results, and we do not intend to (and the study is not able to) test if any of the interpretations makes better sense than the rest. This is not our goal here. Therefore, we are not entirely sure if we should also specify how we *should* (rather than how we *can*) interpret a null result.

> **The sample size is sufficient for the replication hypothesis according to the power analysis conducted by the authors. I think the power analysis proposed (*d* = .25, power = 90%, alpha = .05) is reasonable. However, I wonder if it might be the absolute best decision to power to *d* = .18, the smallest value in the estimated range of the uncorrected moral license effect size according to the meta-analyses cited in the intro. But I understand the potential practical (i.e., resource) constraints facing the authors so I leave it to them to decide.**

> **The extensions are less clear. As stated above, it seems like the authors calculated power for the replication attempt and are essentially assuming/hoping that the effects of the other tests will be at least as large. I don’t think this is unreasonable, but it should probably be made explicit. Otherwise, I would request that the authors justify their smallest effect size of interest for each of the extensions and ensure that they are sufficiently powered to test those hypotheses.**

We decided to do power analyses based on *d* = 0.25 rather than *d* = 0.18 because of resource constraints. Since our primary aim here is replication, and extensions are meant to be exploratory additions to the replication, we do not plan our samples for them. We made this more explicit in the revised manuscript.

> ***Have the authors avoided the common pitfall of relying on conventional null hypothesis significance testing to conclude evidence of absence from null results? Where the authors intend to interpret a negative result as evidence that an effect is absent, have authors proposed an inferential method that is capable of drawing such a conclusion, such as Bayesian hypothesis testing or frequentist equivalence testing?***

> **It is not clear. I believe the authors need to first establish what a null result means conceptually before being concerned about its statistical evaluation.**

We do not intend to draw inferences about the absence of an effect. We are aware that NHST is not capable of achieving this. We can conduct Bayesian analyses as exploration, but we do not feel the need to commit to them at Stage 1.

> ***Have the authors clearly distinguished work that has already been done (e.g. preliminary studies and data analyses) from work yet to be done?***
>
> **The authors do a good job at pointing out work already done. However, I think they should spend a bit more time explaining some of the replication attempts (especially Blanken et al.) that have been undertaken for moral licensing. Right now, I think the intro gives the impression that "some work has been done" without providing any further understanding about how, when, or why the work was conducted or what it found.**

We have expanded our introduction and made the review more informative. We also included more details about previous replications, such as Blanken et al.

> ***Have the authors prespecified positive controls, manipulation checks or other data quality checks? If not, have they justified why such tests are either infeasible or unnecessary? Is the design sufficiently well controlled in all other respects?***
>
> **I am not certain what the comprehension questions are (I could not find them in the supplement). Otherwise, their exclusion criteria are prespecified and, in my evaluation, reasonable.**

The comprehension questions were with the survey files shared on OSF, and we made revisions to spell out the comprehension questions more clearly. We added these to the supplementary materials.

The comprehension questions for the gender preference scenario (options in parentheses and right answer bolded) were:

1. Which is NOT among the duties of this job that you are hiring for? (visiting building sites/preparing financial statements/**negotiating contracts**)
2. An ideal candidate for this position would be one who is: (**technical, aggressive, and confident**/careful, intelligent, and thoughtful/caring, empathetic, and cheerful)

For the ethnicity preference scenario:

1. In your town, what are the attitudes like towards ethnicities other than Whites? (welcoming/indifferent/**unfavorable**)
2. Why did the African-American patrolman quitted your unit? (because of low pay/**because of the hostile working conditions**/because he wanted a different job)

Minor Comments: Page 8 “moral debits” I think should be “moral debts.”

While “moral debts” also makes sense to us, the literature used “debits” (e.g., Miller & Effron, 2010; West & Zhong, 2015). So we decided to follow with that.

References:

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

West, C., & Zhong, C.-B. (2015). Moral cleansing. *Current Opinion in Psychology*, *6*, 221–225. https://doi.org/10.1016/j.copsyc.2015.09.022

Page 11, paragraph 2 first sentence “has” should be “Have”.

Page 11, paragraph 2 in parentheses “…might not help when the person engage…” should be “engages.”

Page 11 near the bottom “…concluded no support for idea…”

We corrected these errors. Thank you for pointing them out.

Page 13, bottom paragraph, the authors claim they are testing moderation but write it as if they are testing mediation.

We revised the paragraph. It now reads:

> “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.”

Participants. The first paragraph has “etc.” after both sets of options. I’m not too sure what was meant by this. What are the other options you are planning to employ?

We moved these to the supplemental materials and included more details about them there.

> **"For example, a 5 - 8 minute survey would be paid 1 USD per participant" was a confusing sentence to read, but I understood what was meant.**

We revised the paragraph to be as follows:

> "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 adjusted payment if needed. The data of these 30 participants were not analyzed, and they would be paid a bonus if the payment was adjusted upwards."

> **Design and Procedure. Paragraph 1 final sentence – "… who did not " should be "who would not" and "… pay attention but only" should be "… pay attention and only".**

Thank you for catching this. We revised accordingly.

> **Deviations. I get the rationale behind wanting to deviate from the original design and present the profiles individually at first. But I cannot shake a concern that it's possible it has a completely unexpected effect (perhaps in the opposite direction) which may harm the replication component of the work. If Monin & Miller were able to find a genuine result with their method, I lean toward replicating it as closely as possible (with fixes to the confounds I raised) meaning removing this addition. The rest of the deviations seem sensible.**

We decided to follow your suggestion and have removed the individual presentations. Participants now only see the profiles together and only once.

> **Confirmatory Analyses. These seem reasonable. My only concern is that the authors are planning to run many uncorrected tests, often using the same variables across these unprotected tests. They plan a Tukey post-hoc comparison only after they plan to compute three ANOVAs and four linear contrasts uncorrected. For the purposes of a replication, I wonder whether it is best to be as strict as possible with Type 1 error rate inflation. If the authors do not want to correct for some of these additional tests as part of their confirmatory analyses, then so be it, but I would strongly suggest that they at least include footnotes that specify where the results diverge if a more stringent correction is applied to all but a single ANOVA (e.g., Bonferroni correction for every test other than the ANOVA for H1 – the primary replication).**

Thanks for the advice. The replication test (the ANOVA for H_1) will be the same as in the original. For the 3-by-2 ANOVA, we will apply Tukey correction to the four planned contrasts (instead of using no correction as in the last submission) and for the exploration on whether a mismatched credential can still license (vs. the control), we will apply Bonferroni correction.