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 with our reply underneath in normal script.

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

A track-changes manuscript is provided with the file: PCIRR-RNR2-Monin & Miller 2001-manuscript-v2-G-trackchanges.docx (https://osf.io/yzvkf)

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
Methods	Ed: Addressed points raised by R3 that were previously missed R1: 1. Reframed hypotheses to match the planned analyses 2. Explained the decision to exclude some participants from analyses; proposed to conduct analyses before and after the exclusion; made sure to revisit this point in Discussion 3. Clarified that all tests will by default be two-tailed 4. Changed the analysis plan to resolve incorrect interpretation of the previous model R2: 1. Removed the numeric values from the labels of scale points 2. Labeled the scale options; switched from 5-point to 4-point scales 3. Added another exploratory measure based on R2's suggestion 4. Provided clearer description of the comprehension checks 5. Changed "gender/ethnicity preference" to "hiring preference" 6. Changed "typing in the last name" to indicate selection to "typing in the full name" R3: 1. Made more explicit that we did not plan our sample size for the extension hypotheses 2. Explained in more detail the decision to remove participants who favor minority candidates 3. Corrected typos
Discussion	R2: Registered to discuss a few points raised by R2

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

[We note that we are not familiar with the titles and ranks of the reviewers, and looking for that information proves tricky. To try and err on the side of caution, we refer to all reviewers with the rank Dr./Prof. We apologize for any possible misalignments and are happy to amend that in future correspondence.]

Reply to Editor: Prof. Chris Chambers

> **I now have received re-reviews from three of the reviewers who evaluated your improved in the previous round. All of the reviewers judge the manuscript to be substantially improved and we now are much closer to Stage 1 IPA. You will find some remaining points to address, principally in clarifying materials and rationale and further strengthening the study design.**

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit.

> **In responding, please also address points (c) and (d) in the original review by Marek Vranka, which you appear to have overlooked in your previous response. I have extracted and pasted below the relevant points below.**

Thank you for informing us about the overlooked points. We have now addressed them below.

> **c) As mentioned above, for testing H3 and H4, participants who favor minority candidate will be excluded. I believe it makes sense, but the decision is not discussed or explained in any detail whatsoever. Is there not a risk of bias? For example, if those with high reputational concerns in condition without credentials will favor the minority candidates, they will get excluded. Only those with low values will remain and there will likely be no association with the DV. In the condition with credentials, even those with high concerns remain and thus there will be the expected negative association. The exclusion would thus lead to the interaction, but in the opposite direction than predicted.**

We decided to exclude those participants who favor minority candidates because including them can raise issues regarding what the primary dependent measure measures. We explained this decision in detail in the first paragraph in section “confirmatory analyses.” We pasted the relevant text below:

> “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 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 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 also agree that there is potential for a bias—though we are not entirely sure how one can reasonably interpret it—so we will do all analyses with and without those participants who favor minority candidates.

> **d) This is really a minor point, but since the authors ask participants to select the applicant by writing his or her name, it is likely that there will be mistakes / typos. It could be mentioned how this will be handled. (Alternatively – and I am not 100% sure whether it is possible in Qualtrics, but I guess it should be – on the page with all profiles, the authors could upload each profile as a clickable image and ask participants to click on the selected applicant – instead of circling it – and then write his or her name. See e.g., https://community.qualtrics.com/XMcommunity/discussion/1596/make-pictures-as-answers-clickable)**

Our current setup is such that the input must be an exact match; otherwise participants cannot proceed. We decided not to implement the clickable image suggestion. Given that the candidates' names are sufficiently different from each other, we believe there is a very minor chance for error selection.

> **There is a typo in the last sentence before the begininng of Results section: "... on this measure as we (sic) the ones we ran on prejudiced preferences".**

Thank you for catching the typo. We corrected it.

Reply to Reviewer #1: Dr./Prof. 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 authors have addressed most of my concerns. However, I still have a few comments (comments #2, #3 and #4 I included in my previous review have not been addressed by the authors):**

Thank you for your valuable suggestions in the previous review round and for your effort reviewing our revision. We are deeply sorry for missing some of the points. We somehow failed to transfer all the comments from the decision letter/review files to our response template. We are grateful for your patience in bringing them to our attention once again.

> **1) H4 talks about correlations, but the analysis uses multiple linear regression. Note that a difference in correlations is not the same thing as a difference in slopes in regression (Rohrer & Arslan, 2021).**

Thank you for raising this important point. We have revised the hypotheses to better match the analyses we plan to conduct. See below:

> H_3: Trait reputational concern negatively predicts preferences for males/Whites in those who have no moral credentials.
>
> H_4: Non-sexist/non-racist moral credentials reduce the negative predictive power of trait reputational concern for preferences for males/Whites (as hypothesized in H_3).

> **2) If preference for males/Whites is influenced by the manipulation of credentials, excluding all participants with some preference for sex/race in analysis of H3 and H4 may result in the problem of conditioning on a post-treatment variable (see Montgomery et al., 2018). It seems that Reviewer 1 made the same comment (#4) and the authors disagreed with him, but still they removed the mention of exclusion of "participants who favor females/Blacks in the sexist/racist scenarios". I am now not sure whether the participants will not exclude these participants, or whether the change was that they will exclude them from tests of all hypotheses and thus do not mention the exclusion specifically when describing the test of H3 and H4. In the latter case, I believe that the exclusion possibly introduces a bias and that it is better to analyze the data including all the participants.**

Thank you very much. We agree it is important to be clear on this point. And you are correct in that we decided to exclude these participants from all analyses (rather than just those concerning H_3 and H_4) and as such mentioned the decision elsewhere (the relevant text is pasted below). We believe there is a strong point for excluding those participants who favor candidates from disadvantaged groups (see also our response to the Editor above). We are aware of the issue of conditioning on post-treatment variables, and we acknowledge the potential for a bias. Therefore, we will do all analyses with and without those participants. Evaluating replication outcomes will be based on results before this exclusion (as the original did not exclude those participants). We also "registered" that we will discuss this point in our Discussion.

Relevant text (first paragraph in section '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 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 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."

> **3) I am not entirely sure, but I believe that using non-centered variables and their interactions in the test of H3 makes the main effect non-interpretable (see Dalal & Zickar, 2012).**

Thank you for pointing this out. You are correct. Our interpretation of the model we planned to build was wrong because we ignored the dummy-coded scenario variable. We have changed our analysis plan so that the categorical variables are no longer dummy-coded:

> "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."

> **4) The hypotheses are directional, will the statistical tests be one-tailed?**

The tests will be two-tailed despite directional hypotheses since (1) the original tests were two-tailed, and we attempt to replicate the original findings with the same tests; (2) two-tailed tests are more conservative with respect to concluding statistical significance. We now specified this ("the tests will be two-tailed unless noted otherwise") in the revised manuscript to ensure that it is determined from the outset.

5) It is not clear from the description of the pretest ("The data of these 30 participants were not analyzed separately [...]") whether the participants from the pretest would be included in the analysis from the main study.

These participants will be included in the analyses for the main study. The relevant text is now:

> "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."

References:

Dalal, D. K., & Zickar, M. J. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. Organizational Research Methods, 15(3), 339-362.

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.

Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. Advances in Methods and Practices in Psychological Science, 4(2), 25152459211007368.

Reply to Reviewer #2: Dr./Prof. Ethan Meyers

> **I think the authors did a superb job responding to my review. Unless explicitly mentioned below, the authors can assume that each of my previous concerns has been remedied, including most of them. Below, I note my outstanding concerns as well as a couple of new ones introduced in this revision.**

Thank you for your thorough evaluation and comments in the previous round and for the constructive suggestions below.

> **Major. First off, thanks to the authors for correcting many of my misunderstandings regarding the "anti" vs. "non"-sexism distinction, as well as what is and what is not required for moral licensing to occur (e.g., no clearly moral initial act is required).**
> **In their response, the authors targeted one of my weaknesses. As an empiricist at heart, their response of "let's test to see if it matters" is something I'm likely to be in favor of by default, and this is no exception. I support the authors' proposed solutions, namely to include the following questions: (1) whether participants viewed a task judgment as sexist/racist, (2) how morally good the decision to hire the best candidate was, and (3) whether participants believed the preferences to be prejudiced or not. I have strong preferences as to exactly how these questions should be presented and worded (resulting in only minor tweaks to the authors' intended method), but if the authors disagree with my preferences, then I wouldn't necessarily expect to see any sort of rebuttal/reply as to why. I'll detail these below, but first, I'd like to note that I find these questions especially helpful.**
>
> **One concern I had while reading the revised work is that the way the authors discuss how moral credentials work is somewhat at odds with how it is tested. (Before I go any further, it's possible that this concern too is a result of my misunderstanding something, but I don't think it is). For example, the authors write:**
> **"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. Importantly, the credentials license morally dubious behaviors by altering how people interpret them" (page 8)**
> **And**

"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" (page 9)

The authors claim that people's interpretations of morally dubious behaviors can change when they have moral credentials, as opposed to when they do not.

Yet, this was not tested in the original experiment – people's willingness to express an aggregate sex or racial preference for a specific occupation is not the same as altering the interpretation of a behavior. Thus, I am not convinced the original experiment assessed moral credentials if it lacked a test of whether interpretations of the morally dubious behavior were different across the licensing conditions. Indeed, in their response to my review, the authors acknowledge that the original work "may not be inherently appropriate for testing moral licensing, or specifically, the moral credential effect."

However, the additional questions proposed by the authors lend themselves to explicitly testing whether people's interpretations differ across credential conditions. In fact, my judgment is that these questions cut to the heart of the matter. By asking about the morality of the act, the authors can observe whether people's moral judgments of the act do indeed change. I find this to be informative regardless of whether H1 replicates or not.

As for how the questions should be presented, I advocate for the following presentation format:
1) Providing a label for each option (see below), especially a midpoint when relevant
2) Including a short definition of prejudice when participants are going to judge prejudice
3) Including these judgments as early as possible in the design

I think labelling all points of the scale is especially important for these questions. Participants should be given the option to state they do not have enough information to make a judgment as the moral credential effect implies they do have enough information (how else could one's moral impression of an act change?).

> **I also believe everyone should be on the same page for what is meant by "prejudiced". I support providing a definition for "racism" and "sexism" as well when relevant but view this as a bit less important. Finally, I believe that these questions should be presented as soon as possible in the design (before reputational concern measure). This is because I view them as essential to the main question of the study. Furthermore, I doubt responding to these questions would affect participant's responding to later measures (e.g., reputational concern). However, if the authors are concerned about potential order effects, then they could counterbalance the presentation order or randomize it. Frankly my judgment is that the order is most likely not too important and so I would be okay with the authors ignoring my suggestion.**

We agree with you that the original studies did not really test whether the moral credential effect is indeed due to altered interpretations of morally dubious behaviors, though it is likely because the theoretical framework of moral licensing (i.e., the moral credits vs. credentials distinction) only comes later. So while we agree on the importance of examining whether moral credentials work by altering how people interpret behaviors, because we aim at replication, we would rather not make it the central focus of this proposal.

We suspect that this setup is of the kind where order effects are most likely to kick in, since the exploratory questions clearly imply a reputational mechanism. Although this is, of course, something that we can test empirically by randomizing the order of the reputational concern scale and the exploratory questions), we would rather not do it here given that we already may not have good power for the extension hypotheses; having to discard half of data in the case of an order effect could be detrimental to the evidential value of whatever we find with the extension.

We have, however, "registered" that we will come back to this point in the Discussion. We find the point valuable, and we share your opinion that future investigations should provide more direct tests of the underlying mechanisms of the moral credential effect.

> **I recommend using the following response options for each question:**
> **Sexism/Racism:**
> **1 = very unlikely to be sexist/racist**
> **2 = somewhat unlikely to be sexist/racist**
> **3 = not enough information to judge**
> **4 = somewhat likely to be sexist/racist**
> **5 = very likely to be sexist/racist**
> **(The authors could also organize this as a 1-4 scale with 5 as the neutral "not enough information to judge".)**

We followed your suggestion to label each option, but we decided not to include "not enough information to judge" or anything similar as the middle option or a separate option in the end.

First, such an option in the middle does not match with the underlying numeric value (say, "very unlikely" represents a perceived probability of 5% of being sexist and very likely, 95%, then the middle option should in theory represent 50%, and the corresponding label would be something like "equally likely to be sexist or non-sexist." This does not make much sense, and we do not think "not enough information to judge" is an accurate label, either (Does indicating an inability to make a judgment mean believing the target is equally likely to be sexist and non-sexist? Not necessarily, we believe).

Second, we believe it is better not to let participants explicitly consider if they have enough information or not (our bet is that if you give them such an option, the majority of them will go for it since indeed, there is not much information; but we want to force a judgment here). Below is the revised setup (we also switched from 5-point to a 4-point scale to avoid the middle neutral option that is difficult to frame).

Before change:

> "… 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 to be sexist/racist, 5 = very unlikely to be sexist/racist; only the endpoints are labeled …)"

After change:

> "… 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 …)"

Morally Good:
1 = strongly disagree
2 = disagree
3 = neither agree nor disagree
4 = agree
5 = strongly agree

I would also recommend the authors include one more question that is essentially a foil to this question:
"Selecting anyone but [candidate's name] for the position is a morally bad decision" with the same response options as above. I think this would be helpful as it's possible that people might not differ in the extent to which they'd endorse the moral goodness of hiring the best candidate (i.e., a ceiling effect) but they may view hiring the best in the non-racism and non-sexism conditions respectively as morally worse than not hiring the best candidate in the no-credential condition.

Great suggestion. We made the changes.

Prejudice
Include the following in the instructions: "by 'prejudice' we mean, "an adverse opinion or leaning formed without just grounds or before sufficient knowledge"" (from the Merriam-wester dictionary: https://www.merriam-webster.com/dictionary/prejudice)

1 = not at all prejudiced
2 = disagree
3 = neither agree nor disagree
4 = agree
5 = strongly agree

We carefully thought about this suggestion and decided that we would not implement it. We believe that providing a standard definition would change the question from how participants themselves perceive the preferences to whether participants consider the preferences to fit the standard definition of prejudice. Both are important questions, but in the context of this study we think the former matters more.

> **Minor: On page 18, the authors describe the comprehension checks in a way that suggests that participants can pass the check even if they did not comprehend the vignette. It is not clear whether failing the check would lead to exclusion from the study. To address this issue, the authors could consider pre-determining a reasonable filter (e.g., participants who fail more than two attempts are filtered out) to increase the odds of excluding participants who did not comprehend the vignette.**

Failing to pass the comprehension checks at the first attempt will not result in exclusion. Rather, participants will be alerted that they made mistakes, asked to reread the text, and attempt the questions again until they finally answer all questions correctly. We believe this procedure is sufficiently capable of ensuring good comprehension. We have revised our explanation of the procedure so that it should be clearer now.

> "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."

> **On page 16 and throughout the manuscript, the authors use the terms "gender preference" and "non-sexist credential". To promote consistency, it would be better to use either "sex preference and sex credential" or "gender preference and gender credential". Given that the authors use "male" and "female" throughout the manuscript and are specifically studying non-sexism, I suggest they replace the word "gender" with "sex" wherever relevant.**

While we agree that this change would promote consistency, we decided to keep using "gender" as gender and sex do carry different meanings in academic writings.

> **On page 19, the authors include numbers as part of the scale, but they plan to deviate from the initial method by tweaking the values from negative to positive. One suggestion is to remove the numbers altogether since each scale point has a clear text label associated with it. Removing the numbers would eliminate any concern about the pluses/minuses being associated with ethnicities/sexes. Alternatively, the authors could counterbalance the scale order to test whether this could have affected the original results. Otherwise, I suggest the authors not proceed with their subtle sign change unless they can provide reasonable justification as to why their concern (that some people might be bothered by minuses being associated with Blacks/females) wouldn't have also applied to the participants of the original design.**

That would be a nicer solution to the issue. We followed your suggestion and removed the numeric labels. We also noted this as a deviation from the original study.

> **On page 19 and throughout the manuscript, I suggest the authors avoid using the term "gender/ethnicity preference" as it could imply that someone generally prefers Whites or Blacks or Men or Women, even though the authors do not mean it this way. One alternative is to use the language of "hiring preference," where the preference is to hire the best candidate, and any differences in hiring preference across the sexism or racism conditions reflect some expression of sexism or racism.**

Good point. We have revised accordingly.

> **On page 21, the authors deviated from circling and writing down the full name of the candidate to typing just the last name, citing the lack of a straightforward way to implement circling on Qualtrics. One suggestion is to consider using radio buttons underneath the profiles or some variant of that. Additionally, it is unclear whether the inability to circle has any bearing on whether the participant must enter the full name vs. just the last name.**

Indeed, the inability to circle has nothing to do with changing from writing down the full vs. only the last name. We revised the text to make that clearer. Now we also do not see a strong point to limit the input to the last name, so we will ask participants to enter the full name, including the middle initials.

> "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 that on Qualtrics."

Reply to Reviewer #3: Dr./Prof. Marek Vranka

> **Thank you for considering my comments and providing thoughtful responses. Overall, I feel that my concerns have been adequately addressed.**

Thank you for reviewing our manuscript. We appreciate your time and effort.

> **However, there are two remaining points: While I understand and accept your focus on replicating the original design without aiming for well-powered extensions, I believe it is important to discuss this aspect in the paper, particularly in the discussion section as a limitation. As a Registered Report (RR), potential readers might assume that all tests are well-powered, and non-significant findings for the extensions could inadvertently hinder further research in this area. I am confident that you are aware of these limitations, and I simply wish to ensure that they are clearly communicated in the manuscript. That being said, I find your rationale for the decision sound and do not have any issues with it.**

We understand this concern. We added the sentence below in our sample size justification section:

> "Therefore, any results in favor or disfavor of those extension hypotheses should be considered exploratory only and would require further confirmatory investigation."

We also made sure to revisit and be explicit about this point in the Discussion. In case an extension result becomes worth mentioning in the abstract, we will also make a note.

> **It appears that you may have overlooked two of my suggestions and a postscript note. Your response concludes with point b) from my suggestions, but points c) and d) seem to have been missed.**
>
> **Additionally, you have not corrected the typo mentioned in the postscript, which indicates that you likely overlooked these items as well.**

We are deeply sorry for failing to address those points. We somehow missed them when we transferred the reviews from the decision letter to our response template. We have now addressed them in the Response to Editor section. The typo has also been corrected.