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.

Major Comment:

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.

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.

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.

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 labelled 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.

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.

Key Issues:

And now for the key issues (bulleted, purple text) and my responses (normal)

- Does the research question make sense in light of the theory or applications? Is it clearly defined? Where the proposal includes hypotheses, are the hypotheses capable of answering the research question?

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)

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.

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…

- Is the protocol sufficiently detailed to enable replication by an expert in the field, and to close off sources of undisclosed procedural or analytic flexibility?

Given the detail 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).

- Is there an exact mapping between the theory, hypotheses, sampling plan (e.g. power analysis, where applicable), preregistered statistical tests, and possible interpretations given different outcomes?
- For proposals that test hypotheses, have the authors explained precisely which outcomes will confirm or disconfirm their predictions?

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 grey 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.

- Is the sample size sufficient to provide informative results?
- Where the proposal involves statistical hypothesis testing, does the sampling plan for each hypothesis propose a realistic and well justified estimate of the effect size?

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.

- 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.

- Have the authors minimised all discussion of post hoc exploratory analyses, apart from those that must be explained to justify specific design features? Maintaining this clear distinction at Stage 1 can prevent exploratory analyses at Stage 2 being inadvertently presented as pre-planned.

Yes.

- 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.

- 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.

- When proposing positive controls or other data quality checks that rely on inferential testing, have the authors included a statistical sampling plan that is sufficient in terms of statistical power or evidential strength?

N/A

- Does the proposed research fall within established ethical norms for its field? Regardless of whether the study has received ethical approval, have the authors adequately considered any ethical risks of the research?

It is my impression that the authors have considered the ethical risks to the research.

Minor Comments:

Page 8 “moral debits” I think should be “moral debts”

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…”

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

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?

“For example, 5-8 minutes survey would be paid 1USD per participant” was a confusing sentence to read, but I understood what was meant.

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”

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.

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., bonferoni correction for every test other than the ANOVA for H1 – the primary replication).

Signed,

Ethan A. Meyers.