**Review for PCI Registered Reports #185**

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).

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.

(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”.

(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.

(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 preregister this analysis.

(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.

Corey Cusimano