Response letter (Stage 1, Round 1)

Dear Dr. Fillon (Recommender), Dr. Celniker (Reviewer 1), and Dr. Kouassi (Reviewer 2),

thank you for your considerate feedback and the time you invested in our project. We are very happy to have received such positive feedback and hope to address all open points appropriately. Below, we go through the reviews in chronological order of reception. From our point of view, we have followed all suggested changes and have further performed a second power analysis, as suggested by Dr. Kouassi.

We look forward to further developing our manuscript based on your helpful reviews and the upcoming steps in the process.

Best regards

Leopold Roth (corresponding author)

Dr. Fillon (Recommender)

Dear Dr. Fillon,

thank you very much for again handling our manuscript. We are very grateful for your engagement and that you organized these productive reviews. Hopefully, we have appropriately addressed all points and have developed our draft according to your expectations. We are looking forward to having our project managed by you as our recommender at PCI.

Dr. Celniker (Reviewer 1)

Dear Dr. Celniker,

we are very happy to see that you also decided to review this project of ours. It is great to have your experience in this literature, and we hope to have addressed your comments appropriately. From our view, they strengthened the manuscript's clarity and we hope that you will get this impression as well from reading the updated version.

Detailed Comments Dr. Celniker:

I think the authors are proposing valid and important studies. They are correct in noting gaps in the literature related to insufficient diversity in target gender and social contexts. These studies will address issues about the generalizability of effort moralization while also addressing questions about gender & contextual differences in person perception that are interesting in their own right.

Response: Thank you very much for this very positive comment. We are happy for this feedback.

I think the general rationale of the studies is sound. That said, I found the set-up of the competing hypotheses lacking in detail and clarity. I appreciated that the authors are setting *up the studies in a way that may support multiple patterns of results, but it was not clear to me what theoretical insights would be gleaned from one pattern being supported vs. another. Also, it was not clear to me how the authors would interpret the package of findings if one of Models A-C being is supported in Study 1 and Model D is supported in Study 2. To what extent are these models complementary or competing, and why would the change from work to care contexts alter the pattern of results? The justifications for the 4 different models were the weakest part of the proposal, in my opinion, and I think the authors could strengthen the proposal by further detailing the key competing predictions of these models and expressing what the theoretical implications are of one pattern of results being supported versus another. The names of the different models succinctly capture the pattern of results being hypothesized, but the theoretical justifications for the models were lacking. To be clear, I do not think the competing models are essential to the proposal - I think that noting the gaps in the literature and wanting to examine variation in effort moralization effects are sufficient for motivating the research. But if the studies are designed to adjudicate between different theoretical perspectives, I would like the authors to help me understand what those perspectives are and the implications of one or another being supported.*

Response: This was a very helpful point to raise, which we took very seriously. After careful consideration, we decided to follow your suggestion and to remove the explicit models from the introduction and the manuscript in general. We hope that the manuscript gained clarity by doing so and we are looking forward to your opinion.

The methodology is sound and closely based on prior work in the area. The power analyses seem sufficient and based on reasonable effect sizes of interest.

Response: Thank you very much. We are glad to receive your positive feedback.

The information provided about the methods and focal dependent variables is sufficient, I do not think there is room for undisclosed flexibility in analyses.

Response: Thank you for this positive evaluation of our methodological approach in the manuscript.

I believe the proposed studies are well-suited for tackling the research questions and that the authors considered issues of comparison conditions, etc. That said, while I like the studies they offer and think they are the right ones to tackle first, I'm wondering if another study may be included in the package to provide further information about effort moralization dynamics: having one male and one female character in the same vignette while crossing the effort & work context. This would be interesting and provide a stronger test of some of the theoretical models offered by the authors about the roles of gender & social context in effort moralization. This design might reveal nuances that aren't observable in the designs the authors offered; however, the authors' designs seem positioned to offer insights that the design I'm proposing can't afford. So my proposal makes the most sense as an additional Study 3 rather than a replacement for what the authors are proposing. Obviously, there are resource limitations that may preclude this possibility. But I'd like the authors to consider the tradeoffs of this design compared to what they proposed to see whether my design may afford them further traction in testing the competing hypotheses they offer.

Response: This is a very good suggestion and we would have been very happy to include the proposed study immediately. Yet, we are working with a small budget at the moment and hence can not afford an additional data collection, while preserving the power of the first two studies. We hope to realize this comparison in a future project, but sadly have no resources to accommodate the suggestion at this time.

I only have one other point of major feedback, which is related to issues I discussed in 1B - I would like to see the authors investigate potential moderators (e.g., endorsement of work norms, endorsement of traditional gender roles, etc). As far as I'm aware, nobody has identified consistent moderators of effort moralization. Testing for moderators would contribute more to the literature overall, even if those are treated as exploratory. This may also help the authors identify viable explanations for whichever pattern of results is supported. I think testing the influence of gender role endorsement would be particularly interesting, but there may be other potential moderators that are better suited for helping address the authors' key questions of interest. I ask the authors to consider whether a short scale (or multiple short scales) might be included in these studies to test for moderation effects.

Response: Thank you for this very productive suggestion. We agree that this could be a valuable contribution to the general understanding of the effect. To keep the survey time short, we decided to use the Gender Role Beliefs Scale – Short Version (Brown & Gladstone, 2012), which uses 10 items (e.g. "Women with children should not work outside the home if they don't have to financially."). We further considered the possibility of including further moderators and sadly are not financially equipped to further extend the running time of the survey while keeping the high level of power in both studies. Yet, we fully agree that further moderators should be tested, while our contribution in the current study will stay limited to the influence of gender norm endorsement.

pg. 6: "which has meaningful implications, especially for work and education environments." - This can probably just be simplified to "which may have important implications for work and education environments." I also don't know why education is featured here instead of care, given the emphasis on care going forward. Is this a typo?

Response: Thank you for pointing this aspect out to us. We adapted this section to solely focus on the relevance of cooperation selection at work. We hope the section hence gained clarity and precision.

pg. 6: "blind spots" in the section header seems a bit pejorative, it might be better to say "gaps" or something like that. Anecdotally, we were very aware of the gap in studying gender effects in our 2023 paper, but we decided to focus on male names to simplify our initial studies, knowing that these issues deserved their own paper. So I took "blind spots" a little more personally than you likely intended. It is certainly an important gap, but it's one that I think many of us have been aware of.

Response: Thank you very much for this very transparent comment. We did not mean to attack previous work with this framing but rather motivate readers to engage with these topics, which are still often deemed of less importance. We renamed the section and are happy to deliver complementary data to your prior work.

pg. 9 : "This provides insights that more accurately reflect the cooperative dynamics found in everyday life." - I think you can make the point about studying another important cooperative dynamic without making unwarranted (or at least unsupported) claims about this being a "more accurate" reflection of cooperative dynamics. It is common to choose cooperation partners, you don't need to deny this to make your point about forced cooperation also being a common dynamic. If you want to double down on this, then provide some citations about how partner choice is not a real or important cooperative dynamic; otherwise, I'd suggest walking this back a bit and simply note that assigned cooperation is also important.

Response: Thank you for providing us with your perspective on this matter. We have adjusted the wording in this section, so that it describes both processes as rather complementary to each other. We hope to have done this appropriately and that future readers can retrieve knowledge on both forms of cooperation initiation.

pg 11 - "Since morality is perceived as a female core trait and therefore a standard for what to evaluate them on, effort might be moralized stronger for female agents, i.e. they will be judged more positively in terms of morality if they display high-effort behavior. " - I want more of an empirical backup for this claim, where does this come from? This relates to my comments in 1B, I do not totally understand what the authors take to be the theoretical claims they are testing. E.g., I am not aware that "morality is perceived as a female core trait", can the authors back this up with a citation, or is this their conjecture/theorizing? What aspect of morality are you talking about here, all moral traits or some specific moral traits?

Response: Thank you for this comment. As described in our response to the comment above, we decided to remove the theoretical models, based on your prior comments.

pg. 11: You can also make the point about what "punishment" means for the female high morality model a bit clearer. In Figure 1, it looks like low-effort females are expected to be seen as more moral than high-effort males; the "punishment", I assume, is then a greater difference between high- and low-effort moral character for females than males? It seems weird to call this "punishment", I think, I'd rather the authors be more specific about the empirical predictions rather than baking in interpretations of the predictions into the model. Including more specifics, and perhaps reconsidering some word choice, seems appropriate.

Response: Thank you very much. Please refer to our response to the comment above.

Dr. Kouassi (Reviewer 2)

Dear Dr. Kouassi,

thank you very much for your time and expertise, invested in our project. Your comments were very insightful and we hope to have addressed them according to your expectations. From our perspective, the manuscript greatly benefitted from your contribution and we are happy that you helped us to rethink some key aspects of our manuscript.

Detailed Comments Dr. Kouassi:

The paper addresses an interesting question, focusing on the "effort moralization effect". The authors announce their intention to replicate the effect in two areas: work and caregiving context. At the same time, they raise the question of whether the effort moralization effect is moderated by gender, by presenting the various predictive models available in the literature on the subject. Overall, the introduction sets out the subject well and supports its arguments correctly. The two studies presented are coherent with what the authors claim to demonstrate, and the methodology makes it possible to establish a relevant cause-and-effect relationship. The fact that the different possible predictions based on different models of moral judgment are considered is a pertinent point and enables readers to clearly visualize what can be expected from these studies. However, we have identified a number of points to bear in mind in order to optimize the registered report.

Response: Thank you very much for your positive assessment of our manuscript. We are happy that you liked our summary of potential models. Note, that based on the comments by Dr. Celniker (see above), we decided to remove the potential models from the manuscript. As Dr. Celniker argued, the empirical basis of some of the predictions was not solid enough. Hopefully, future projects in the same domain can utilize our data to develop more specified predictions.

If we expect women to be punished more than men in the moral judgment made of them for violating expectations, then shouldn't women in the low-effort condition be judged more negatively on their morality than men in the low-effort condition? According to the results of Kennedy et al. (2016), women were indeed judged more harshly than men in the event of *unethical action. Indeed, in the present studies, why would women who make no effort (and therefore violate observers' expectations) remain better judged than men (including those who make little effort) in these studies?*

Response: Thank you for pointing this out to us. This corresponds to our response above and the comments by Dr. Celniker. We therefore removed the prediction from the manuscript and refer to potential differences in a less deterministic way.

The different models presented (A, B, C and D) give readers a clear idea of the results that can be expected from the two studies. To optimize and clarify this understanding, it might be appropriate for the authors to include a paragraph that clearly indicates the situation in which we might expect to see the different models expressed. For example (this is just an example, and not necessarily what needs to be stated), mention the fact that, according to the literature, one might expect one model (e.g., model A, B or C) to express itself in the world-of-work study, while another model (e.g., model C or D) will express itself more in the case of caregiving, and/or explain which models are the most credible according to the literature in each situation. This could give greater visibility and clarity to the hypotheses.

Response: This also closely corresponds to our response above. Based on the feedback by Dr. Celniker and your comments regarding our predictions, we decided to exclude these models from the manuscript. We hope that future studies in this field can base their hypotheses on our data and make more nuanced and evidence-based predictions.

The tasks described in the care condition are all non-relational tasks (cleaning, preparing meals, etc., rather than spending time with the person, chatting, taking them for a walk, etc.). However, more relational tasks would correspond to an important aspect of caregiving, would theoretically be closely linked to the warmth dimension and could potentially be judged as a greater emotional investment (similar to studies by Johnson & Park, 2021). Indeed, including these relational tasks would better illustrate the effects insofar as caregiving tasks could be more accurately categorized as warm.

Depending on the research objectives, it might be relevant to explain why the relational aspect was not considered in the conditions of the caregiving study. And if research objectives allow, it might even be possible to include more relational tasks in the text of the caregiving study conditions.

Response: Thank you very much for raising this very relevant point. We came across this discussion during the creation of our materials as well. Our rationale was to create vignettes that are not overly stereotypical with regard to masculinity or femininity. Based on this reasoning, we changed the original vignette in the work context (Celniker et al., 2023, Study 6), which described workers in a factory setting to a more gender-neutral office setting. The same line of thinking motivated us to specifically avoid 'warm' care work, which is potentially perceived as prototypically feminine. Given that care work includes relational and non-relational tasks, we decided to focus on non-relational tasks in the vignette to elude stereotyped answering behavior based on task characteristics. To accommodate this, we added a footnote in the manuscript, which clarifies this for future readers as well.

*In line with the recommendations of various works on power analyses of interaction effects (cf. this link: https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/ and see also Sommet et al., 2023: https://doi.org/10.1177/25152459231178728 and this app www.intxpower.com), the G*Power analysis that has been carried out by the authors, although it seems appropriate for the main effects targeted, appears to be insufficient for analyzing the power of interaction effects (e.g., the interaction between gender and amount of effort produced). I therefore recommend that the authors include in the document another power analysis that takes into account the specificities of an interaction effect (using, for example, the resources proposed above or any other resource that would be relevant for a power analysis of an interaction effect). In view of the effect sizes targeted by the registered report (small), the sample size required is likely to be much larger than initially planned (initially anticipated at N= 350 per study according to the G*Power analysis currently posted). In fact, the authors will probably have to choose between these two options:*

1) Increase the sample size of the two studies to bring them into line with the recommendations on the detection of interaction effects. This would be the most desirable option, as it would maintain the ambition of the paper and thus provide solid support for the literature. However, it is also very costly, and there is a strong possibility that, due to various constraints, it will be impossible for the authors to obtain a sufficient sample.

2) Indicate in the document that it was not possible to recruit more participants, and that the sample as presented is indeed too small in relation to the recommendations made for the *detection of an interaction effect. The ambition of the registered report will therefore be scaled back, but it will still provide a contribution to the current literature.*

Response: Thank you very much for addressing this very important point, which we discussed carefully, following your recommendations. We used the intXpower tool, as suggested, which we have used in prior projects already, to double-check our sample size planning. Based on our target effect size (small: $\eta^2 = .01$, corresponding $d = 0.20$), we computed our power for a mixed design with two-sided testing and 95% power, assuming a small interaction effect. This resulted in a recommended minimum sample size of $N = 325$ (https://intxpower.com/?A=0.7&B=0.5&C=0.5&D=0.7&targetPower=95&algo=mixedTwoTailedFactorial). We included this computation in the supplemental material and our sample size computation, cited Sommet et al. (2023), and adapted our planned analysis to an interaction effect.

Yet, we are aware that different forms of interactions result in different required sample sizes. Given the exploratory nature of the specific form of the interaction term, depending on the dominance of the factors, we added a cautionary footnote to the data analysis that some forms of mean patterns could be under the threat of insufficient power, as suggested by you.

From our perspective, this refinement has helped the clarity and transparency of the manuscript, and we hope to have accommodated your comment appropriately.

There are two spaces after the sentence "The effort shown at work might be perceived as a manifestation of these stereotype dimensions."

Response: Thank you for your close readership and good eye on the formal standards. We corrected this in the manuscript.

In the 3rd line, a period is missing at the end of the sentence: "The data was collected in two separate samples at month/year [Stage 2], with participants from one study being excluded from participating in the other"

Response: Thank you very much, this was corrected as well.

The article makes an interesting contribution to the literature. In particular, it extends the "effort moralization effect" to other contexts (caregiving) and proposes a legitimate and natural link with gender differences and the associated differences in judgments. The means *employed (in terms of sample size) may perhaps curb the ambitions of these studies, but the fact remains that the document proposes interesting answers to the questions it poses, and its form allows a good understanding of the subject. On a personal note, I would recommend a resubmission and adjustments by the authors with a view to future publication.*

Response: Thank you very much for your positive assessment of our project. We hope that the changes, based on both reviews, lend further strength to the manuscript, and that we adapted everything according to your expectations.