Response Letter (Revision Round 1)

Is it Worth the Hustle? A Multi-Country Replication of the Effort Moralization Effect and an Extension to Generational Differences in the Appreciation of Effort

Dear Dr. Fillon (recommender), Dr. Ziano (reviewer 1), Dr. Celniker (reviewer 2), and Prof. Dr. Inzlicht (reviewer 3),

we thank you all for your effort and time contributed to our project. We have studied the reviews and recommendations carefully and will respond to all of them in the following pages. In the first section, we will address an editorial decision after initial review and revision before review circle two, points raised by multiple reviewers as well as points raised by Dr. Fillon.

Following the recommenders' suggestion, we initially focused on the review provided by Dr. Ziano, followed by Dr. Celniker and Prof. Dr. Inzlicht. We hope to have addressed all points adequately and look forward to the next review round.

While we are very appreciative of the many positive comments, for the sake of brevity, we will only address the actionable recommendations below.

Best regards

Leopold Roth (corresponding author)

Editorial decision after initial review

One-sided testing and replication criteria by LeBel et al. (2019)

After rigorously reviewing all comments, we identified a methodological issue in the implementation of the recommendation to use LeBel et al.'s (2019) replication criteria, as well as the recommendation to switch to one-way testing, which we described to the recommender of the article (Dr. Fillon) as follows before resubmitting the manuscript:

One point, which we hope to have accommodated appropriately, but which remains sub-optimal from our perspective, was the parallel inclusion of two suggestions. It was suggested multiple times to switch our analysis to one-sided testing. While we initially preferred two-sided testing, given the variance of effects in Celniker et al. (2023) study 2a-c between countries, we followed this recommendation. Yet, Dr. Ziano raised multiple times the very good recommendation to apply the replication criteria by LeBel et al. (2019) to evaluate the replication success. These criteria are very dependent on the ability to construct confidence intervals to judge the consistency and direction of the estimates (consistency: is original effect within CI of current estimate; direction: if not within CI, are boundaries of CI above, below, or in the opposite direction of original effect size). Given the way how one-sided confidence intervals are constructed (estimate [boundary, ∞]), this assessment can not be conducted by the same quality standards as with two-sided confidence intervals, from our perspective. We hope to have understood and described this challenge sufficiently to be further developed. From our perspective, this requires an editorial decision on how to proceed. We see multiple potential directions: 1. Remain with one-sided testing and the criteria by LeBel et al. (2019) and have limited interpretability, given the lack of numeric boundaries on one side of the confidence interval. 2. Remain with one-sided testing and drop criteria by LeBel et al. (2019) and judge replication success by signal and effect size as suggested in the initial submission.

3. Return to two-sided testing and apply the criteria by LeBel et al. (2019) in their full usefulness (this would be the preferable option from our perspective). Yet, we are not trained statisticians and are open to alternative procedures and always happy for suggestions. (correspondence between Dr. Fillon and the corresponding author LR)

Subsequently, Dr. Fillon made the editorial decision to integrate the third potential direction by applying the replication evaluation criteria recommended by LeBel et al. (2019), while maintaining the two-sided test in the analysis plan. We have hence adapted the manuscript and the sample size computation (final target *N* = 680).

General points (raised multiple times)

Choice of country

The question about the choice of countries for our project came up on multiple occasions in the review and we address this point in this general section for the sake of parsimony. Following the recommendations by Dr. Celniker, we reduced the number of countries for the project to Mexico and Germany. Both countries play meaningful roles in terms of economy, population, and international influence in their respective geographic region. Further, both authors speak German as their mother tongue and one of the authors (Tassilo Tissot) speaks Spanish fluently, which supports the translation and sampling process.

Improve introduction

This point was raised by most reviewers and we hope to have improved the introduction meaningfully. From our perspective, we eliminated redundancies and offered a more streamlined introduction to future readers. Additionally, we hope to have strengthened our argumentative structure which increased the comprehensiveness of the manuscript.

Comments by Dr. Fillon (recommender)

Dear Dr. Fillon,

thank you very much for your editorial efforts in progressing this registered report. We were especially happy to read your enthusiasm regarding the provided materials. We will below address your raised points chronologically.

Detailed comments:

The introduction is too long and too short: too long on unrelated topics, and too short (and with a lack of precision) on the actual topic. Both Jared Celniker and Ignazio Ziano provided examples and suggestions.

Response: Thank you very much for providing us with feedback on the introduction. From our perspective, we have streamlined the introduction in the updated manuscript and hope that the readability has increased respectively.

There is room for improvement in the method, especially for detailing the sample used, the power analysis (see Jared Celniker comments here), and the tests (as Jared pointed out, I don't understand the need for a two-sided t-test, and as Ignazio pointed out, you need to improve correction for multiple testing). Please be very clear regarding the relationship between sample size, power, and effect size of interest.

Response: Thank you very much for your summary of these points. Based on the recommendations of Dr. Ziano and Dr. Celniker, we have reduced the scope of countries, included in the project to Mexico and Germany as our resources for this project are indeed limited, as correctly addressed by Dr. Celniker. By doing so, we are able to accommodate the suggestion to decrease the effect size, used for the power analysis. We further removed the correction for multiple testing and followed the agreement, described in 'Editorial decision after initial review', keeping two-sided testing as well as integrating the replication criteria by LeBel et al. (2019). The power is now computed as follows: $d = 0.20$ at $1-\beta = .95$ and $\alpha = .05$. We hope that the updated version is sound and intelligible.

I add that, in complement to Cronbach's Alpha, I would like to see a McDonald's Omega.

Response: thank you for this suggestion. We have implemented McDonald's ω in the analysis code (see line 92 ff.)

"Jared Celniker and Michael Inzlicht asked for details regarding age. I think that when looking at the R script, things are clearer. The confusion lies in the method section, as you want a sample with a balanced distribution, while you will use age as a continuous variable. Please state that explicitly to avoid confusion and explain in the introduction why the use of age as a continuous variable can help understand better the theory tested (See Michael's review)."

Response: Thank you very much for raising our awareness on this point. We hope to have made our analysis clearer by mentioning age as a continuous predictor in the section 'Extension to age as a predictor of effort moralization'.

"In addition to the typos found by reviewers, I would like to add that on page 16 the "after" is truncated, page 17 is blank, and in table 5 footnote 2 is not displayed, at least on my version."

Response: Thank you for pointing these mistakes out. We corrected them in the updated manuscript. The '(2)' in Table 5 was not meant as a footnote but as an indication that we test a non-linear quadratic term of age as well (age and age^2). We deleted it as this was not intuitive.

In the introduction, please avoid the strong terms as "demonstrated" as researchers and findings indicate support for a theory, they don't demonstrate anything.

Response: Thank you for raising this point. We hope to have improved the introduction accordingly.

Now regarding disagreement, both Jared Celniker and Ignazio Ziano used the PCI-RR guideline for reviewing, and they mostly don’t agree with each other’s. Based on the details of their reviews, we can understand why they don't and how authors can overcome these problems. I strongly suggest the authors to firstly and extensively answer Ignazio’s review, as he provided several suggestions for improvement, before completing with answers to Jared Celniker and Michael Inzlicht.

Response: Thank you very much for providing us with editorial suggestions on navigating the submitted reviews. This was very helpful and we revised the manuscript accordingly.

Comments by Prof. Dr. Ziano (reviewer 1)

Dear Prof. Dr. Ziano,

Thank you for your detailed review of our project and for all the suggestions you made to improve our work. Please find below our responses to your comments in chronological order. We specifically valued your recommendation to integrate the replication criteria by LeBel et al. (2019), which gave the manuscript an added layer of objectivity in terms of its methodological approach.

Detailed comments:

In general, I think the writing in this paper might be more to the point. For instance, rather than starting with Aristotle, I would prefer you start with a description of what you want to do in the paper. The reader of scientific papers wants to know what you do and what you find, not read introductions. There are several sections that seem redundant and not in line with the topic, for instance, the section about faces. Your paper is not about faces, so I do not think this section should be included. Similarly, the section about phrenology, racism, and morality inferences is out of topic. Your paper is not about phrenology or racism. I would replace it with a section about the many cues that lead people to infer morality, for instance.

Response: Please refer to the comment ‘Improve introduction’ in the section on general points.

The main effect in the abstract should be explained in a clearer way. Right now, it reads ‘The effort moralization effect describes the process of deriving information through the effort invested into a given task’. I would change it to ‘The effort moralization effect is the finding that people who are exerting more effort in a job are seen as more moral’.

Response: Thank you very much for this suggestion. We adapted the abstract very closely to your recommendation.

Personally, I do not think that all the discussion of quiet quitting – a concept I find ill-defined – is warranted. If you want to test the effect of age on the EME in your paper, that is fine, justify it with scientific literature.

Response: Thank you for sharing your impression of the introduction with us. We have streamlined the introduction in this regard and hope to have raised the conciseness of our arguments. As now described in the manuscript, we will test differences in effort moralization by age and connect this analysis to the public discourse of quiet quitting, etc. which we define as reduced willingness to provide uncompensated work.

When writing, you are using too many commas. For instance, in the sentence at p.9 ‘One phenomenon, that has raised scientific psychology’s interest in recent years is the observation, that people appear to use effort, invested in given tasks, as information on the morality of agents, further summarized as effort moralization effect.’ There should be no commas.

Response: Thank you for raising our awareness on this matter. We have heightened our efforts to deliver a manuscript in a more appropriate grammatical form.

At p. 10, you write ‘building on this finding, Celniker et al. do XYZ’, but you have already described Celniker et al. in the previous paragraph.

Response: Thank you for pointing this out. We have deleted this duplicate from the manuscript.

You often combine together Bigman and Tamir 2016 and Celniker et al. 2023 as both showing an effort moralization effect, but that is not really correct. B&T look at the amplifying effect of effort. If you regard Celniker et al. 2023 as part of the findings of B&T (because it only looks at positive outcomes perhaps) then you should more clearly explain so.

Response: Thank you for sharing this impression with us. We hope to have increased the clarity of the concepts for future readers of the manuscript.

Looking at the Word documents of the survey, they still have a default placeholder in the informed consent page. See below for an example in the case of the German survey. Please double-check everything. Also note that the OSF file viewer has issue with Word (it shows the files as having many more pages than they actually have). I recommend you upload pdf files instead.

Response: Thank you for pointing this shortcoming of our materials out to us. We have adapted the materials and uploaded them as PDF files. We hope that this increased the usability in a meaningful manner.

You should not apply a within-test alpha correction for the number of tests. The tests you want to do are paired-sample t-tests (you call them dependent t-tests, which is fine) but they are all independent from each other. This analysis is not an ANOVA with multiple pairwise comparisons where the pairwise comparisons tests are dependent on the main analysis.

Response: Thank you for sharing this perspective with us. Following your suggestion, we have removed the alpha correction from our manuscript and have adapted the power computation respectively.

It is not clear to me how you translated the original text of the survey in Celniker et al 2023 into several different languages. Please document the translation in much more detail. This is a crucial passage and cannot be liquidated in a couple of sentences and a citation as you do right now.

Response: Thank you for requiring higher precision in this regard. We have extended our description of the translation process considerably and hope that this provides a more thorough impression.

You have a series of completed data in several tables (e.g., Table 6 and 7). Are they part of a pretest? Are they randomly generated data (as I suspect)? You need to specify it.

Response: Thank you for highlighting that this was not sufficiently clarified. The analysis was based on random test cases, drawn from an in-built function of Qualtrics. So far, this was mentioned below the headline of the results section. To support the readability, we changed the font color from black to red.

Why did you choose Germany, Mexico, NL, and SA? Convenience sampling or other reasoning? This is also an aspect you should describe in further detail.

Response: Please refer to the comment ‘Choice of country’ in the section on general points.

In your survey, there are questions related to participants' future of work. These questions are barely introduced in the paper. What is the point of these questions? Where do they come from? Are they validated? How will they be analyzed? Please specify.

See above for the status ladder. Please specify why you include each measure and how do you intend to analyze it.

Response: Thank you for your close study of our materials. These items about future work are self-created and won’t be part of the analysis but were included for potential future projects of master students of the authors. The same applies to the status ladder.

Is it clearly defined? ***Yes, but more detail is needed in several places.***

Response: Thank you for raising our awareness towards potential uncertainties in our design and analysis. As the point is brought up in the following comments, we speculate that this was a reference to your suggestion to use the replication criteria by LeBel et al. (2019). We are happy that you suggested these to strengthen the interpretability of the results and we hope to have implemented them correctly.

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? ***Not yet. More detail is needed, see above.***

Response: We were lacking some detail in this comment to be sure to have it answered in completeness. We speculate that this comment was in reference to your suggestion of implementing the criteria by LeBel et al. (2019) in the later comments. Hopefully, we have adapted our manuscript sufficiently from your perspective.

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? **No details are given regarding the interpretation if the findings are not conclusive or if the tests are not statistically significant, and more is needed.** *For proposals that test hypotheses, have the authors explained precisely which outcomes will confirm or disconfirm their predictions?* ***No, please see above.***

Response: Thank you for this recommendation. Given our prior implementation of Bayesian hypothesis testing in the analysis plan, we assume that this is directed towards the recommendation for the usage of the criteria by LeBel et al. (2019). We are very thankful for this suggestion and hope to have implemented in accordance with your expectations.

Is the sample size sufficient to provide informative results? ***Not so much.***

Response: While we were not entirely sure about the specific critic in this comment, we hope that our adjustments, following the recommendations of Dr. Celniker and the associated increase in sample size per country, are sufficient from your perspective as well.

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? ***Unfortunately not, and I encourage the authors to do so. In general, this being a replication, you should use the LeBel et al. (2019) guide to both a) the methodological facets of the replication and b) the quantitative comparison between the original and the replication results. At p. 20 you should also specify the coverage of both confidence and credible intervals (e.g., 95%).***

Response: We were not entirely sure about this comment as we already implemented Bayesian hypothesis testing to our analytic strategy. Yet, we are very thankful for your suggestion to utilize the criteria by LeBel et al. (2019) to further specify our planned interpretation of the results. We hope to have them included appropriately as well as the specification of using a 95% confidence interval.

Have the authors clearly distinguished work that has already been done (e.g. preliminary studies and data analyses) from work yet to be done? ***Unsure, see my comment above on Tabe 6 and 7.***

Response: We hope that our clarification about the origin of the simulated data above was sufficient.

I know that I made a lot of observations, but it seems to me that this paper needs to be refined quite a bit.

Response: Thank you again for your thorough review of our manuscript! Especially the pointer towards the replication criteria by LeBel et al. (2019) has strengthened the clarity of our proposed interpretation meaningfully.

Comments by Dr. Celniker (reviewer 2)

Dear Dr. Celniker,

thank you very much for your valuable contribution to our project. We are very excited that the original first author is involved in the review process and hope to have understood your earlier project well enough to implement this replication and extension in a meaningful manner. Below, we will address your comments in chronological order and look forward to the next round of reviews.

Detailed comments:

It wasn't clear to me how the authors planned to recruit roughly equal numbers of participants from the various age brackets they proposed. How will that be accomplished? That may be easy enough using platforms like Prolific, but I imagine it would be more difficult to control the age of respondents when recruiting on social media. A little more information on how they plan to accomplish this would be appreciated.

Response: Thank you very much for this question. We will use an in-built function of Qualtrics, which allows us to 'fill cells' by pre-defined criteria. If a cell is fully sampled (e.g., the 18 – 30 bracket makes 25% of the target sample size), participants of this age can not participate in the study anymore. To ensure this, we have a first filter question implemented right after the participation agreement. The specific age in years is assessed at the end of the survey.

I would prefer having more details about the recruitment sites the authors will be using (e.g., specific social media platforms), but if that information is detailed in Stage 2, I think it is fine.

Response: Thank you for bringing this question up. We will use two main platforms for sampling. On the one hand, we will collect data using Prolific. On the other hand, we will collect data through social media sampling, utilizing services by Meta, which offers synchronized advertisements on Facebook and Instagram in parallel.

I presume the authors are already planning on modeling age continuously, but it wasn't totally clear whether they would be modelling it continuously or as a factor variable (split by age bracket or "generation"). I would strongly suggest modelling age continuously, and a couple of words to make this more explicit would be helpful.

Response: Thank you for requiring specification at this point. We acknowledge that this was not stated clearly enough from our side. We addressed this through specification (see Extension to age as a predictor of effort moralization). The age brackets are solely implemented to achieve a broad age distribution.

I think this is the part of the proposal that could be strengthened the most. While testing whether effort moralization effects replicate in additional cultures is important, I was less clear about the motivation for choosing the specific countries that were selected. Why were Germany, Mexico, The Netherlands, and South Africa chosen? If there are theoretical reasons for selecting these countries (e.g., they represent different kinds of cultures than those in which effort moralization has been previously tested), I would like to know more about that reasoning. Alternatively, if the countries were selected out of convenience, I'd like to know that as well. To be clear, I think it's totally fine if convenience was a major factor. Convenience sampling wouldn't make me feel this study is less important, but understanding any theoretical rationale that may be operating in the background could make me feel like *this work is even more important than I do now. So being clearer about why these decisions were made seems like an opportunity the authors should capitalize on.*

Response: Please refer to the comment 'Choice of country' in the section on general points.

The empirical connection to bullshit jobs wasn't very clear to me. The materials the authors are using don't really relate to bullshit jobs. I think that concept is indirectly related to this work, but I don't see the age moderation analyses as providing evidence of increasing aversion to bullshit jobs. I think the authors could just drop the bullshit jobs reference from their title to address this. Otherwise, I'd want to know a bit more about how they see the age analyses as speaking to aversion to bullshit jobs (rather than to simply moderating the strength of effort moralization). I'd love to see work connecting effort moralization and perceptions of bullshit jobs and labor, but I don't think it's the authors' goal to do that in this proposal. For what it's worth, Study 3 from Celniker et al. (2023) would be better to make the case for age differences related to bullshit jobs, as the stimuli in that study was related to a job that could be fully automated (and would thus be closer to meeting the definition of a bullshit job).

Response: Thank you for your close readership of our arguments. These were very helpful points and we adjusted the headline of the manuscript accordingly. Our updated title reads as: 'Is it Worth the Hustle? A Multi-Country Replication of the Effort Moralization Effect and an Extension to Generational Differences in the Appreciation of Effort'. Indeed, we fully agree that the vignette of Study 6 is not a direct assessment of a potential 'bullshit job'. Rather, we refocused our attention on possible differences in the appreciation of effort with equal output across generations. We hope you find this title more appropriate for the current project.

The authors decided to use d = 0.4 as the effect size of interest in their studies. This is based on the results of our Study 6, but I think the authors may want to consider whether there may be a smaller effect size of interest (see Lakens, 2022, for more on this topic). Powering the samples to detect a d = 0.4 effect may result in some inconclusive results. For example, suppose that effects of d = 0.2 emerge across countries. From my quick and dirty power analysis, I think the proposed studies would be underpowered to detect this size of an effect. In this scenario, I think it would be reasonable to interpret the effort moralization effects as generalizing to some degree, yet the effects would likely not be statistically significant given the proposed sample size. The impact of this issue compounds when considering the age moderation analyses: if the effort moralization effects are smaller than anticipated, then the interaction analyses will be underpowered as well. Recruiting more participants requires more money, and I imagine there are resource limitations the authors have for this project. That said, if resources are indeed limited, I might suggest dropping one of the countries of interest to afford an increase in the per country sample size for the remaining samples. I think this may be a reasonable trade-off to ensure you are well powered to test the age moderation effects, especially since the rationale for studying effort moralization in the selected countries wasn't entirely clear (as I mentioned previously). Also, it was not clear to me why the authors proposed a two-sided test for the replication effects given they provided directional hypotheses. The same applies for the age moderation effects: the authors are predicting that older participants will moralize effort more than younger participants. Using one-sided tests would help reduce the number of participants needed to reach suitable levels of power for smaller effects.

Response: Thank you for sharing these points with us and suggesting a productive alternative from here. We followed your suggestions and respectively changed our research program by limiting the data collection to Mexico and Germany with increased sample sizes (*N* = 340 per country). This allows our planned comparisons to be conducted with 95% power for *d* = 0.20, using two-sided testing. The suggestion of one-sided testing was not implemented in the revision, based on its contradiction with the sufficient application of the replication criteria by LeBel et al. (2019). This decision was formed in correspondence with the recommendor Dr. Fillon and is described in the section 'Editorial decision after initial review'. Yet, by reducing the number of sampled countries and the increase of the sample size by country, we hope to have included your suggestions regarding the smallest effect size of interest suffieciently. To clarify the understanding of the regression analysis, we are not conducting a moderation, but a linear regression of age on the individual difference score of moralization. Therfore, the sample size is adequate to conduct this analysis. As acknowledged, the available resources for this project are limited and hence we are happy to have found a workable solution to accommodate your suggestions while remaining within the budget.

The literature review was a bit hard to follow. It covered a lot of interesting and related topics, but the path to the current research question was a bit confusing to me. I think the authors could narrow the scope of their introduction to hone readers in on the key findings and real-world phenomenon that motivate their study. For instance, I think that much of the "Reading virtue, measuring morals" and "Impressions of morality as a function of behavior" sections could be cut to streamline the Introduction.

Response: Please refer to the comment 'Improve introduction' in the section on general points.

The authors can take or leave this suggestion, but I thought I'd throw it out there. I recently came across some research on individual and cultural differences in difficulty mindsets (e.g., perceiving difficulty as important, Fisher & Oyserman, 2017; perceiving difficulty as improvement, Yan et al., 2023). It strikes me that some of the cultural trends that we are observing around work may reflect some generational changes in those mindsets. Perhaps *younger people are less likely to endorse certain mindsets, such as perceiving difficulty as signal of something that leads to improvement or self-growth. Differences in these mindsets, rather than differences in effort moralization, may help explain some of the trends that the authors detail in their Introduction. I'd predict that age will more strongly relate to difficulty mindsets than to effort moralization (i.e., I'm not sure the authors will find the moderation by age they hypothesize, I think people will moralize effort roughly equally across ages), but that's just a conjecture. I'd love to see whether differences in difficulty mindsets relate to effort moralization and whether age differences more strongly predict difficulty mindsets or effort moralization (there are other interesting questions in this vicinity as well). To be clear, I think this would be a completely exploratory set of analyses, but I think this study may provide a useful opportunity to integrate difficulty mindset and effort moralization research.*

Response: Thank you for offering us these interesting suggestions. For the present project, we decided not to integrate it, to keep our focal analysis parsimonious as well as due to the limited resources of the project and the associated costs of prolonged studies on Prolific. If you are interested, we would be open to discuss this as possible future project together after the completion of the current registered report to avoid conflicts of interest.

Comments by Prof. Dr. Inzlicht (reviewer 3)

Dear Prof. Dr. Inzlicht,

thank you very much for providing your expertise to our project. We will address your comments below in chronological order.

Detailed comments:

"While I see the value in this replication, the authors do not justify why they chose the countries of Germany, Mexico, Netherlands, and South Africa. I suppose they could make a WEIRD

argument, but my bet is that these data will come from wealthy students even in the developing nations, opening questions about how much of a generalization this is."

Response: Please refer to the comment 'Choice of country' in the section on general points.

"I am also unclear about the validity of the age prediction. I mean, sure, go ahead and examine this as it would be an interesting and free thing to examine. However, I do not believe it is a direct test of whether younger people are less willing to exert effort and work than older people. For that, the authors could examine actual effort willingness or even simply ask them about their values. The analysis here only gets at this indirectly, asking if young people moralize effort to the same degree as older people. Still interesting but not an examination of work ethic, per se."

Response: Thank you very much for sharing this impression with us. We hope that the updated manuscript includes a clearer description of our research question. To our understanding, we do not predict differences in work ethic by age but observe differences in effort moralization in situations of unproductive effort by age. We would not suggest that younger individuals, relative to older are less willing to exert effort in general. Rather, we predict that unproductive effort is less moralized by younger individuals. We hope to have strengthened this point in the updated manuscript sufficiently.

"I have another comment about the introduction. To put in plainly, I found it strange. There is a rich tradition in social psychology of person perception, whereby the field has tried to understand how we perceive and appraise other people. Instead of placing this study in this context, the authors talk about physiognomy, race science, and face perception, which has nearly nothing to do with these studies. My advice to the authors is to use the same framing as the Celniker paper, but to clearly frame this as a replication of that paper and then say why this replication is important, critical, etc."

Response: Please refer to the comment ‘Improve introduction’ in the section on general points.

Methods are sound. However, I have two related questions. First, the authors list 8 dependent variables, yet their Bonferroni correction indicates 5 family of tests. Can the authors explain this discrepancy. Second, and related, I imagine that there are different hypotheses for some of these dependent variables. What are they? Do the authors predict, for example, that "quality of work" will change via effort framing? I assume not, as the vignette makes clear that work quality is the same. So, some clarification needed here."

Response: Thank you for bringing up these points as they appear to have raised confusion and were not stated clearly enough on our end. In correspondence to other reviewers' comments, we have removed the alpha correction from our analysis plan. The variable about **perceived effort** is implemented, following the procedure of Celniker et al. (2023), to control for correct effort perceptions in participants and as a potential exclusion criterion (see Data Cleaning). The variables about **quality of work**, **job difficulty,** and **work value** are included, following Celniker et al. (2023, p. 73, right column). The authors assessed these measures to illustrate the strength of the effort heuristic, as all these dimensions are kept constant between scenarios in the vignette. While no difference should be observed, participants showed differences in these dimensions (*d* = 0.20 – 1.20). Last, we have five dependent measures, which are the essential targets of the study, following the procedure of Celniker et al. (2023). We have specified predictions on differences in perceived morality (**core goodness** and **value commitment**, see Celniker et al., 2023, p. 73, left column). The other variables (**competence**, **warmth**, **pay deservingness**) were shown to vary substantially between countries. We therefore have no specified hypothesis for these dimensions.

The authors include some manipulation checks, but unfortunately, they don't always do a good job of labeling them as such (see my second point above). The authors should clarify these manipulation checks and why they are important."

Response: Thank you for bringing up this point. We have included more details on the manipulation checks, by highlighting them in Table 4 as such as well as describing perceived effort explicitly as a manipulation check in the section 'Data cleaning'.

"I don't know where to say this, so I will say it here. There are spots where the authors were not sufficiently careful, leading to typos and errors. For example, on p. 7 they write "integer" but perhaps they mean integrity; p. 17 is a blank page, etc. There are few other errors like this. The authors should read their manuscript carefully to correct these errors."

Response: Thank you for bringing this to our attention. We corrected the mistakes and hope to have developed the manuscript to the highest standard, achievable for non-native English speakers.