I appreciate the opportunity to review this report, and I’m excited to engage in the registered reports process (this is my first RR rodeo)! I thought the PCI website provided helpful questions for me to consider when reviewing this Stage 1 report, and my answers to these questions for the “Is it Worth the Hustle?” project are below in bold:

* 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?
  + **I may be a bit biased (as the authors are planning to replicate our study), but I think the design is well suited for testing the generalizability of effort moralization effects in additional societies. The hypotheses are stated clearly in the Study Design Table. The authors expect the effort moralization effects to replicate across countries, and they also expect that age will moderate effort moralization effects in each country.**
* 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?
  + **In general, I believe the protocol is sufficiently detailed, but there are a few missing pieces of information.**
  + **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.**
  + **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.**
  + **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.**
* 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?
  + **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.**
  + **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).**
* For proposals that test hypotheses, have the authors explained precisely which outcomes will confirm or disconfirm their predictions?
  + **Yes. In the Study Design Table, the authors specify what they would consider to be evidence of the effects being fully generalizable, partly generalizable, or not generalizable to the countries they sample from.**
* Is the sample size sufficient to provide informative results?
  + **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**](https://pure.tue.nl/ws/portalfiles/portal/214011492/collabra_2022_8_1_33267.pdf))**, 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.**
* 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?
  + **This is mostly addressed in my previous response. I think d = 0.4 is a justifiable estimate of the effect size, but I think powering the studies to observe smaller effects would strengthen the proposal. It would be a shame to go through all this effort (no pun intended) and be left with inconclusive findings due to insufficiently powered analyses. I understand this isn’t a trivial suggestion, but I hope the authors consider it.**
* 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](https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00781/full) or [frequentist equivalence testing](https://journals.sagepub.com/doi/full/10.1177/2515245918770963)?
  + **Yes. The analyses are structured such that null effects will not provide evidence of absence. The authors seem to recognize this.**
  + **They have also proposed some Bayesian analyses (though I must admit that Bayesian stats are not my strength, I hope another reviewer can comment on this aspect of the proposal more confidently).**
* 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, I saw no discussion of exploratory analyses.**
* Have the authors clearly distinguished work that has already been done (e.g. preliminary studies and data analyses) from work yet to be done?
  + **Yes, my understanding is that none of this work has been done yet. I was a little confused by the inclusion of some seemingly simulated data at the end of the report, but I interpreted those data as placeholders for the results that will be populated in Stage 2.**
* 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?
  + **Yes, they have described the quality control and manipulation checks they will use.**
* 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?
  + **Yes, though see my previous comments regarding reasons why I think the project may be strengthened by increasing the by-country sample size.**
* 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?
  + **The research is minimal risk and falls with established norms for the field. It was not clear whether the study has already received ethical approval from an institution.**

Here are a few additional points that I didn’t get the chance to address above:

* 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.
* 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.
* Overall, while I think there are some things that the authors could improve about this report, I think this is a valuable proposal. I wish the authors the best in any necessary revisions to their Stage 1 report, and I’m looking forward to learning more about (and hopefully reviewing) their results!
* One final point, more for the PCI team than the report authors. I do not know the authors of this project, but had I been more familiar with them or their work, allowing me to see their names could have introduced biases in my review of their manuscript. There should perhaps be more anonymity in the RR review process to minimize the chances that various biases can skew reviewers’ evaluations (at least until everything is finalized, and then everyone’s names could be revealed to one another and the public). I’m all for the accountability that open science practices can provide, but I think we may need to figure out ways to balance those considerations with measures to prevent potential biases in the review process.

Jared Celniker