Dear Corina,

Thank you for the encouraging words, we are very excited to see this project come to fruition. Rather than responding to every comment in this letter, we have included only the ones which required us to make changes or construct a response in this letter. We very much appreciate all the positive comments from yourself and the reviewers!

We would also like to note that, while you clarified that there are no space constraints at PCI RR, we felt that moving the results by measure to a supplement improved readability and placed emphasis on the overall conclusions. Therefore, we have made this change, and added references to the supplementary materials where appropriate. All changes are marked with track changes.

Lindsay J. Alley

Jordan Axt

Jessica K. Flake

1) Results: I found it extremely useful that you clarified the size of the effects in relation to what your tests were powered for (e.g., "Item 1 ("I find satisfaction in deliberating hard for long hours") was the only item above the cut-off for a medium effect, all others were small or negligible"). I noticed that some paragraphs discussed the a small effect being the cut-off, while others discussed a medium effect being the cut-off. It might be even clearer if you noted in each paragraph that the effect size cut-off related to the power/sensitivity/etc analyses you conducted at Stage 1 for each analysis, which is why it differed.

These descriptions were intended to communicate where each effect size falls relative to suggested cut-offs for the interpretation of DMACS effect sizes. We did not develop the cut-offs, and there is not one cut-off in each case, rather we are just reporting how many items were large enough to be described as small, medium, or large effects according to guidelines. To clarify this, we added the following text on p. 25 where these cutoffs are first mentioned in the results:

"Of the 10 comparisons that retained configural but rejected metric or scalar equivalence, 3 displayed DMACs effect sizes that were below the cut-off for a small effect according to suggested cut-offs for interpretation (>.20 and <.40 small, >.40 and <.70 medium, >.70 large; Nye et al., 2019)."

2) Discussion: "power in ME testing is impact by the strength of inter-item correlations" - change "impact" to "impacted"

We made this change on p. 31.

3) Discussion: "For this reason, researchers should not assume that different crowdsourced samples will be equivalent to each other, or even student samples collected in different settings". Could you please clarify what "different settings" refers to? Different countries/languages/etc.?

We changed the text as follows to clarify this:

“For this reason, researchers should not assume that different crowdsourced samples will be equivalent to each other, or even student samples collected in different settings e.g., in the lab versus online.” p. 33

4) Study design table: you could add a column to the right that shows your findings.

We have added a column to the table titled “interpretation of findings” with the following text:

RQ1: The measures we examined were non-equivalent across crowdsourced and student samples. Additionally, measures were non-equivalent across different crowdsourced samples (i.e., MTurk and Project Implicit), and some measures were equivalent across student samples collected online vs in the lab while others were not. We recommend that researchers interested in pooling or combining these samples test for measurement equivalence.

RQ2: Correcting for the non-equivalence of loadings and intercepts did not change the overall conclusions of any of the replication effects and changed the estimated effect sizes by only small amounts. While the pooling of uncorrected data from these samples is not justified, the results are robust to this practice. However, many measures displayed configural non-equivalence across samples, and data from these should not be combined, as conclusions will not be valid.

Response to Reviewer 1: Benjamin Farrar

Thank you for your kind comments, and for examining the reproducibility of our code. One error was pointed out by this reviewer:

I re-ran the code for one comparison (EMA implicit vs MTurk) as a reproducibility check, and I was able to fully reproduce the equivalence test results for this comparison, although I did note the mean age for Mturk was 34.98400 (35.0) rather than the 34.0 reported. The code is clear, and excellently commented on throughout.

We have corrected this error, thank you for pointing it out!

Response to Reviewer 2: Shinichi Nakagawa

I have reviewed stage 1 of this MS and very much enjoyed it and was looking forward to reading stage 2. I first acknowledge that I am a quantitative ecologist so I do not know the relevant field and literature. Yet, I would be able to check whether the statistical analyses conducted were sound. Also, this is my first time reviewing stage 2, but my understanding is that I check whether they followed the stage 1 plan and check for deviations. The authors conducted the

study with very minor deviations. I liked that the Discussion section had limitation and recommendation sections, which are very clearly and honestly written. Overall, I think this is a great stage 2.

Thank you, we are very glad that you are happy with the stage 2 overall.

I have one question, tho. By reading this work, I got the impression that authors are encouraging to be cautious about mixing samples. Yet, some papers in biology encourage the mixing of samples knowing non-equivalence (differences, e.g. sex and strains). I wondered what authors make of this, and there should be some related discussion. I note this mixing process is called "heterogenization", which is encouraged by an increasing number of grant agencies. There is an example paper:

Voelkl, Bernhard, et al. "Reproducibility of animal research in light of biological variation." Nature Reviews Neuroscience 21.7 (2020): 384-393.

We absolutely agree with the goal of diversification or heterogenization of samples and feel that employing large samples in psychology with greater cultural diversity is a positive move for the field. This is one of the reasons that we are such fans of projects like the Many Labs that pool diverse sources of data. In psychology, there are additional methodological hurdles to doing rigorous science, particularly when we combine psychological measures from diverse samples. In the biological research discussed in the Voelkl et al. (2020) article, the variables discussed are animal phenotypes. One example of such a phenotype is weight. If researchers were to compare animal weight gain under different conditions, and some labs had measured weight in grams and others in ounces, you would not get a coherent result if you combined these data without adjusting them to be on the same metric. When you are measuring a psychological construct, such as endorsement of moral foundations, the issue of ensuring an equivalent metric across samples becomes more complex, but nevertheless needs to be considered.

The statistical approaches that ensure different samples are on the same metric (we use multiple group confirmatory factor analysis in our paper) require first that there is evidence that the same construct is being measured in each group. Imagine being interested in an animal's size: one lab measured weight in grams and the other measured length in centimeters. There is no way you could adjust these to be on the same metric, as they are fundamentally different ways of measuring size. When there is statistical evidence that the factor structure of a measure is different across groups, this is evidence that the underlying concept is different. Perhaps moral foundations may be understood as individualizing and binding for one sample, but this is not how moral considerations are structured in a different cultural context. For many psychological constructs, these questions are open, and it isn't readily obvious if you are capturing different concepts at the outset. The hierarchical testing procedure that we employed in this study first tests whether there is evidence that the same construct is being measure across groups (i.e., both measures of weight, not weight in one group and length in the other), and then, if the same construct is being measured, proceeds

to test whether it is being measured on the same metric (i.e., grams in both, not grams in one and ounces in the other).

We absolutely encourage the combination of diverse samples, but when measures of psychological constructs are employed in research, care must be taken to ensure that these groups are comparable on the variables measured. We feel that, if it is not possible to ensure that the construct is conceptually equivalent across samples and on the same metric, it is better to analyze the samples separately. We define measurement equivalence in the introduction of our paper in slightly more formal and less metaphorical terms. I have copied the relevant passages below:

"Large replication projects such as the Many Labs present a host of measurement challenges. The international and collaborative data collection is a strength (Henrich et al., 2010), but the pooling of data from heterogeneous samples can also introduce invalidity. When samples are drawn from different populations, there is the possibility that measures exhibit non-equivalence because the items do not hold the same meaning across populations. This poses a problem for replication projects, as ME is a prerequisite for valid group comparisons and the pooling of data across samples (Davidov et al., 2014)." (p. 5)

"Also called measurement invariance, measurement equivalence is concerned with whether a particular scale is measuring the same thing in the same way across different groups. Formally, this means that, for a given level of the latent trait, the conditional distribution of the items of the measure is the same across subpopulations (Meredith & Millsap, 1992). Thus, within a latent variable modelling framework, "measuring something in the same way" means that the items of the scale are related to the latent variable in the same manner across groups. There are different levels or degrees of ME, each of which has as its focus a different aspect of the item to latent variable relationship. These hierarchical, increasingly restrictive models can be tested using multiple group CFA, allowing researchers to understand to what degree the measures function in the same way across groups. Figure 1 shows an overview of the hierarchical levels of measurement equivalence; they are described in more detail below." (p. 6)

We feel that clarifying this point and more explicitly stating our endorsement of the goal of large diverse samples in psychology is beneficial to the paper, so we have added the following to our recommendations on p. 34:

"While large-scale collaboration in psychology address limitations often found in other research, such as low power and limited generalizability, measurement differences pose a methodological challenge. **We feel that employing large samples in psychology with greater cultural diversity is a positive move for the field, and the Many Labs and other big team science projects have made important contributions to this effort. However, it is because we believe in the value of this undertaking that we want to ensure that challenges threatening the validity of conclusions from such research are adequately addressed.** For scores from measures to be validly interpretable in the context of these studies, more work is needed examining their validity in relevant groups. In advance of further large-scale replications and

other collaborations using existing measures of unknown quality, we feel that similarly large-scale collaborative construct validation research would help put any future projects on more solid methodological footing."