Dear Zoltan,

Thank you for initiating the timely review of our Stage 2 manuscript: "The labels and models used to describe problematic substance use impact discrete elements of stigma: A Registered Report". We are glad to hear that the reviewer found merit in our research and suggested that it be 'recommended'. Below we respond to your final comment regarding the equivalence test regions. We hope our revised manuscript, and reply below, are satisfactory.

Editor's Comment

"Just one thing on my side. As per my comments for the Stage 1, equivalence cannot be concluded unless there is justification for the equivalence limit being so small that it is only just interesting/uninteresting from the point of view of the scientific theory tested. Resource limits obviously do not provide that justification, as they are arbitrarily related to theory. Indeed several of your CIs within the "equivalence region" include effect sizes that past studies have used as support for the theory. So remove references to having found equivalence, and just e.g. leave it as a case of having estimated the possible effect size".

Response: Thank you for this helpful comment. Note that the smallest significant effect found in previous research was d = .15 and only one of our effect sizes which was found to be inconclusive was around this region (i.e., prognostic optimism, d = -.17). Previous research using traditional NHST does not justify what is a meaningful effect in this research area (although we recommend that researchers work towards). Based on this, we have revised the Results section to state that the assertions we make are based upon "our data" and "the effect size range that we could reliably detect" (given our sample size). We also highlight where an effect may indeed be classed as meaningful in previous studies. These revisions are as follows:

"Based on our data, a meaningful effect is asserted if, given α = .01 the mean difference is significantly different from zero and the 99% CI lies significantly outside of the equivalence range; equivalence is asserted if the 99% CI lies within the effect size range that we could reliably detect; and an inconclusive is asserted if the null hypothesis test and equivalence test are non-significant."

"Blame (d = -.11, CI = -.27, .07), prognostic optimism (d = -.17, 99% CI = -.35, -.004) and continued care (d = .11, CI = -.06, .28) were inconclusive, but note that this estimated effect size for prognostic optimism was similar to that found to be meaningful in previous studies."

"and this was significantly outside of our equivalence range".

"Table 1. Descriptive (M, SD) and inferential statistics for the three research questions. The first reported test result is the Welch's t-test and the second is the equivalence test based on the range of - .20 to .20. Based on this equivalence range, cells highlighted green indicate a meaningful/significant effect, yellow an equivalent effect, and red an inconclusive effect."

The other effect sizes found to be near the ES in previous research relate to our own task (the Financial Discrimination Task) and associated hypotheses. As such, they do not "include effect sizes that past studies have used as support for the theory". These were inconclusive (and the largest inconclusive effect is d = .15) and smaller than the ES found for the self-report measures.

Yours sincerely,

Dr Charlotte R. Pennington