Dear Zoltan,

Apologies for my confusion with the initial Stage 2 revisions; I now completely understand the points you are making. I have revised the manuscript accordingly and respond in detail below.

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 1:** I have now updated the analytic strategy section to be clearer in this regard (also see response 2 below). Specifically, I state that the equivalence range is based on our sample size justification and then specifically state that other researchers may specify a different smallest effect size of interest based on the effects they deem meaningful. Note that past studies which have also focused on RQ1 (labels) and RQ2 (models of addiction) have found effect sizes of $d = .15$ for self-report measures, but all of our 'equivalent' or 'inconclusive' effects were below this (RQ1 = d -.06 and .09 and RQ2 = .01, -.07 and -.11). RQ3 was our own research question so this same comment does not apply.
>
> Page 16 [tracked changes document]: "We used the upper and lower equivalence range of $-\Delta L = -.20$ and $\Delta U = .20$ based on our sample size justification and set a conservative alpha ($p < .01$) given the number of analyses. Note therefore that other researchers may specify a different smallest effect size of interest that they perceive is meaningful".

"I still see some inconsistencies in applying the logic of a CI being within or outside a suitably defined equivalence region:

> **Response 2:** The analytic strategy and the terminology we used when describing the results was confusing in this regard. In line with Lakens (2017, 2018. 2022), we have revised the strategy to state:
>
> "We refer to an effect as 'significant' if, given $\alpha = .01$, the mean difference is significantly different from zero and the 99% CI falls outside of the equivalence range; 'equivalent' if the mean difference is not significantly different from zero and the 99% CI falls within this equivalence range; and 'inconclusive' if the mean difference is not significantly different from zero but the 99% CI falls outside of the equivalence range.".
>
> We have then used these terms to clarify the Results, specifically stating in some instances whether it's the upper or lower bound of the ES which falls outside of the 99% CI and is therefore deemed 'significant'. This aligns with the automated written output of the TOSTER r-code, and we have corrected one effect size CI which was missing a minus sign. We now exemplify these using the examples you provided to document the initial inconsistencies:

1) "For the Financial Discrimination Task, participants in the drug use condition allocated greater punishment compared to those in the health concern condition, with the observed effect size (d= .14, 99% CI = .02, .26) significantly outside of the equivalence range"

But the sample mean is not significantly outside the equivalence region. If it were signfiicantly outside the equivalence region, the lower limit of the CI would be equal to or above the upper limit of the equivalence region. This problem occurs in various sentences. If by assumption values within the equivalence region are too small to be of interest - that is what an equivalence region is - then a CI that includes such values does not establish that an effect of interest was found.

> **Response 3:** Based on Lakens (2022), we deem this significant because the upper CI is outside of the equivalence region (*d* = .26). The TOSTER r-code output states: " NHST: reject null significance hypothesis that the effect is equal to zero. TOST: don't reject null equivalence hypothesis". When we compare this to an equivalent effect, the TOSTER r-code states: "TOST: reject null equivalence hypothesis".
>
> We now clarify this as follows:
>
> Page 17: "For the Financial Discrimination Task, participants in the drug use condition allocated greater punishment compared to those in the health concern condition (*d* = .14, 99% CI = .02, .26), with the upper CI significantly outside of the equivalence range."

2) "The difference for continued care (d= .01, CI = -.15, .17) was not statistically different to zero and equivalent." Also: "and for punishment (d= .007, 99% CI = -.16, .17) was equivalent" Yet, as you say elsewhere, these CIs include effect sizes previously regarded as of interest. This point needs to be explicitly made. Also you should explicitly state that your equivalence region may include values of actual interest, at a point in the discussion where you conclude equivalence."

> *Response:* It's first important to state that the previous research in this area, which this study is based upon, did not specify a smallest effect size of interest or explain what they deem is a meaningful/worthwhile effect. They use large sample sizes (e.g., *n* ~ 4000) which could find small effects (and they don't perform any power analyses). To clarify that other researchers are free to choose their SESOI, I now state:
>
> "Note therefore that other researchers may specify a different smallest effect size of interest that they perceive is meaningful".
>
> And I specify that our equivalence range is based on our a-priori sample size justification:
>
> "We used the upper and lower equivalence range of $-\Delta L$ = -.20 and ΔU = .20 based on our sample size justification and set a conservative alpha ($p < .01$) given the number of analyses."
>
> These two effect sizes and their associated CIs therefore fall within our pre-specified equivalence bounds (and the effect size estimate is negligible).
>
> I believe that with these revisions, the Results are now written consistently with the Analytic Strategy.

Yours sincerely,

Dr Charlotte R. Pennington