Dear Zoltan,

Thank you for your patience and guidance here; we have learnt a lot throughout this process, and it has been an enjoyable one. We have now revised the manuscript to remove any inconsistencies by using dichotomous claims. Below I outline my response to your comments, which have significantly improved the manuscript.

**Editor's Comment**

"We are getting there but a few points. In a few places you say things like:

"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"

The upper limit of the CI is not "significantly outside" the equivalence range, it is just "outside" of it. That phrasing makes it sound like .26 is significantly different from .2 - but that has not been tested. What matters is what one can say about the population mean; and the sample mean is not significantly outside the equivalence region, so the set of possible population means, i.e. what is inside the CI and hence cannot be rejected, includes values deemed too small to be interesting i.e. practically equivalent to zero. You are allowed to say the drug use condition allocated greater punishment compared to those in the health concern condition; but the following clause should be deleted. (Also consider other similar sentences.) But given your claims about what effects are meaningful, and your Stage 1 decision procedure, even this phrasing is not the one you should use (see below).

You say "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" - but this is not the rule that has been applied, at least on a straightforward interpretation of what "falling outside" means, and given the logic of wanting to exclude all meaningless values in order to conclude there was a meaningful one; and also given how such reasoning has proceeded since Greenwald (1975) onwards; and given the need for consistency in a decision procedure for when a difference is regarded as good enough to be meaningful. If this rule were applied then you would declare a meaningful difference if the bottom limit of the CI were above the upper limit of the equivalence region (and vice versa, for a CI below the equivalence region). That rule, as I have just stated it, makes perfect sense. In the stage 1 you phrased it thus: "Equivalence will therefore be asserted if, given $\alpha = .01$, the 99% confidence interval of the mean difference lies within this equivalence region, and rejected if the 99% CI lies outside of this region" and that states what I have just stated - equivalence is only rejected if the CI lies outside the equivalence region. Now I see you haven't distinguished "completely" from "partially" - but "partially" leads to self contradiction in what erffects are regarded as meaningful. Thus, the striaghtforward interptation of what you have stated that preserves self consistency is that outside means completely outside - and I presume you wish to avoid self contradiction. Conversely, equivalence only is declared when the CI lies within the equivalence region. THis means you suspend judgment in all other cases - namely if the CI overlaps both the equivalence region and the region of values deemed meaningful. Notice this condition for suspending judgment is different from the one you have just stated; for example, if a sample mean is significantly different fropm zero, yet the CI spans equivalent and meaningful values, you suspend judgment. This rule needs to be consistently applied.

Concerning whether an effect size of .15 is meaningful for some tests but not others, in the absence of a principled argument for why, this seems arbitrary. So bear this in mind in how you interpet results.

**Authors Response**

I now understand that our inferences were inconsistent because we were not applying dichotomous claims. For example, I cannot claim a meaningful effect if (a) the effect size is below the one I specified for the equivalence range (e.g., d = .18 when the range was d = -.20 to .20) or (b) the 99% CIs include values that are both within and outside of the equivalence range. In such cases, judgement is reserved,

and these differences are inconclusive. The manuscript has been revised in the following ways to remove these inconsistencies and ensure correct inferences are made:

First, we have reverted the Analysis Strategy back to the accepted Stage 1 manuscript wording whilst further clarifying what is meant by an 'inconclusive effect':

"We interpret an effect as *meaningful* 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 99% CI overlaps both the equivalence range and the range of values deemed meaningful.

We then apply this logic to the results, the inferences for RQ1 remain the same, except for that of 'punishment', which has changed from meaningful to inconclusive:

"For the Financial Discrimination Task, the difference in reward (*d* = -.03, CI = -.14, .09) was not significantly different and equivalent, and for punishment (*d* = .14, CI = .02, .26) was inconclusive."

The inferences of RQ2 remain the same, except that for social distance, which has now been included in the sentence with other inconclusive results:

"The difference for social distance (*d* = -.20, CI = -.36, - .03), prognostic optimism (*d* = -.07, CI = -.24, .09), danger (*d* = -.11, CI = -.28, .05), and public stigma (*d* = -.18, CI = -.34, -.02) were inconclusive."

For RQ3, however, many of the results now change from meaningful to inconclusive; the 99 CI lies both within and outside of the equivalence range:

"The differences for social distance (*d* = -.23, CI = -.40, -.06), danger (*d* = .19, CI = .03, .37), blame, (*d* = -.11, CI = -.27, .07), prognostic optimism (*d* = -.17, CI = -.35, -.004) and continued care (*d* = .11, CI = -.06, .28) were all inconclusive. Similarly, the difference for public stigma (*d* = .31, CI = .14, .48) was inconclusive, as although the effect size estimate was outside of the equivalence range, the CIs included values that were within it."

Table 1 has been updated accordingly to these changes. Please note, that exploratory findings using different SESOIs (supplementary material) are now even more consistent with these confirmatory results. All code, data and supplementary materials have therefore been updated and uploaded to the OSF.

The first two paragraphs of the Discussion have not changed. The third paragraph has been revised to reflect the changes in findings, acknowledging that Kelly et al. were able to detect smaller effect sizes (due to greater statistical power), but that the direction of the results are inconsistent with the current study:

"Whilst Kelly et al. were able to detect smaller effect sizes than the current study (e.g., perceived danger, *d* = .15), and some of our confidence intervals include effect sizes around this region that others may deem meaningful, the direction of these findings for all but one of the subscales (prognostic optimism) are contrary".

In paragraph 4, we then revise the discussion of findings for RQ3, focusing on the inconclusive results and recommendations for future research:

“Although not considered explicitly within either, a key difference between the two previous studies relates to the scope for attributional judgments afforded by the vignettes (see Davies, 1997; Kingree et al., 1999): in Kelly et al. the individual with problematic substance use is described as receiving treatment with a high likelihood of success (high stability condition), whereas in Rundle et al. they are described as seeking treatment with a variable outcome (low stability condition). When manipulating these factors in the current study, we found that the differences for social distance ($d$ = -.23), danger ($d$ = .19), blame ($d$ = -.11), prognostic optimism ($d$ = -.17), continued care ($d$ = .11), and public stigma ($d$ = .31) were inconclusive. As such, whilst the effect size estimates for some of these effects were outside of our equivalence range, and align with that of previous research (Kelly et al., 2021; Rundle et al., 2021; see also Kvaale et al., 2013), their confidence intervals overlapped both the equivalence range and values deemed meaningful. Future work in this area should therefore explicitly define their smallest effect size of interest, justify which effects are practically meaningful (see Anvari et al., 2022), and ensure that they have sufficient statistical power to reliably detect these effects. Furthermore, researchers should scrutinise whether the vignettes they use inadvertently manipulate other potentially confounding factors that may impact results”.

We also update Paragraph 5 with regards to the inconclusive effect for punishment on the Financial Discrimination Task:

“The influence of health condition was inconclusive for punishment and equivalent for reward indices on this task, and the influence of both aetiological label and attributional judgment was equivalent for punishment and inconclusive for reward.”

Finally, we have updated the Abstract with regards to the inconclusive results whilst ensuring that this meets the word count guidelines for the journal we wish to submit to if this manuscript is Recommended:

“Findings for attributional judgement were either inconclusive or statistically equivalent.”

We again thank you for your time and diligence which has made our manuscript considerably stronger. All co-authors have approved the final version of this manuscript for re-submission.

Yours sincerely,

Dr Charlotte R. Pennington and co-authors.