Response to Editor

Dear Prof. Dienes,

Thank you for your time in reviewing our Stage 1 Registered Report. Your comments have been extremely invaluable in strengthening the manuscript, both with regards to the methodology and stringency of the proposed analytical tests. Below we respond to each of your comments and highlight all changes in the manuscript in red font. We also upload a clean version for peer review.

1) Your N is set by a power calculation for testing against an H0 of no effect. But you wish to interpret non-signfiicant results with equivalence tests. That means you have one system of inference for asserting there is an effect, and a different one for asserting there is not an effect of interest. This can lead to contradictions, e.g. a significant test against no effect would have led to a conclusion of no effect of interest if equivalence testing alone had been done. You may wish to use a system that is more consistent. For example, you could just use hypothesis testing against no effect with high power; or you could just use equivalence testing, generalized as an "inference over intervals" such as the rule that if a 90% CI is completely within the equivalence region there is no effect of interest; if completely outside there is an effect of interest. If using inference over intervals, the relevant way to estimate N would be the N needed for the CI to fall within the equivalence region e.g. 90% of the time when there is no effect; and outside of it 90% of the time when there is the predicted effect (see https://psyarxiv.com/yc7s5/).

Response: This is a very good point and we have consulted Daniel Lakens for some additional guidance on this, too. We have recalculated statistical power using the two one-sided t-tests (TOST) approach for equivalence testing. Note that in our analytical strategy using NHST, we have used a more stringent alpha of .01 to correct for multiple comparisons. We therefore also use this alpha within the equivalence tests, too. Equivalence will 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.

The method section now reads as follows:

"Our planned sample size is informed by the effect sizes obtained from Kelly et al. (2021) and Rundle et al. (2021) as well as time and funding constraints (see Lakens et al., 2021). For our main effects of interest (see "Vignette development" below), Kelly et al. observed a significant effect of Cohen's ds ~ .15 for perceived danger, ds ~ .20 for prognostic optimism, ds ~ .30 for continuing care and ds ~ .43 for blame, whilst Rundle et al. observed an effect of ds ~ .1.03 for Stigma Ratings. A power analysis based on the two one-sided tests procedure for equivalence testing (see Dienes, 2021; Lakens, 2017; Lakens, 2021) indicates that with 1,578 participants (n = 789 per group) we will achieve 90% statistical power using the lower and upper equivalence bounds of $-\Delta L$ = -.20 and ΔU = .20 with alpha set at .01. This is within our resources and allows us to detect and reject the second smallest effect size from Kelly et al. (2021). Note that effect sizes of ds =/> .20 have also been found in meta-analyses assessing the influence of the brain disease model on public stigma (Kvaale et al., 2013) and therefore a null result with the planned sample size would also yield informative results with respect to the presence or absence of effect size estimates provided by this meta-analysis."

2) You define d = 0.2 as your minimal effect of interest. Why wouldn't d= 0.11 be of relevance to the theory being true, given you cite such an effect as one obtained in past studies and taken seriously I presume as an effect of interest? Is there really the same minimal effect of interest for all contrasts and DVs? (THat would imply that e.g. all DVs have the same reliability.) (See paper just referenced.)

Response: Please note that in Kelly et al. (2021) the smallest effect size observed was d = .15 for perceived danger, and not .11 – we have updated this within the manuscript. However, your important point still stands:

We are bound by funding and resource constraints for this project, which has informed our power analyses. Using a lower and upper equivalence bound of $d = .20$ will provide 90% statistical power to detect the second smallest effect size observed in Kelly et al. (2021) and all other observed effects in both Kelly et al. and Rundle et al. (2021); in other words, we have statistical power to **detect four out of the five significant effect sizes** observed in these two studies. We have also changed our alpha level from .05 to .01 to correct for multiple comparisons, which requires more participants (and associated funding). With 90% power, $-\Delta L = -.20$ and $\Delta U = .20$, and alpha set at .01, we will require 1,578 participants, which is just within our funding resources. If we power based on $d = .15$, in comparison, we will require 2,804 participants, which is unfortunately not within our available resources. If an effect around $d = .15$ is indeed found, then the equivalence test is likely to suggest that our data is inconclusive. In this case, we will ensure that our discussion does not go beyond the data, suggesting that future research is (a) required to further assess how the labels and models used to describe problematic substance use influence perceived danger and (b) to evaluate what effect sizes would be deemed practically meaningful (see recent discussions by Anvari et al., 2021). In addition, a previous meta-analysis has found effect sizes of Cohen's d = ~/> .20 (Kvaale et al., 2013), which we would have sufficient power to detect. We have updated Table 1 in the manuscript as a response to this comment.

3) You say main effects will be followed by Bonferroni corrected t-tests; I am not sure what these would be given df = 1 for all main effects. Specify the family of tests and what the correction will be, if you are going this way. What I actually recommend is that you pick the one contrast that best tests each theoretical claim, and stick with that; i.e. each row i nthe design table has one test aligned to the substantial theoretical claim at stake. Other tests can be exploratory and reported as such in the Stage 2.

Response: Apologies for this error – the Bonferroni corrected t-tests were referring to the interactions. We have revised our original analytic strategy and, upon reflection, agree that too many tests were planned. We have therefore implemented your suggestion to pick one contrast for each research question that can best test the theoretical claim. This boils down to independent samples t-tests for each of our research questions, rather than the initial 2 x 2 x 2 B-S ANOVA. As you say, interaction effects and other tests can be exploratory and reported as such in Stage 2. Further, with open data, researchers would be welcome to conduct secondary analyses on this large, well-powered dataset.

The analysis section now reads as follows, and we have also updated Table 1 with regards to the theoretical models that can be tested with our research questions:

"To allow for comparisons between the current study and that of Kelly et al. (2021) and Rundle et al. (2021), we will conduct the following analyses on the five discrete subscales of

the Stigma & Attribution Assessment and the total score from the Personal & Perceived Public Stigma Measure. We will then conduct the same analyses on the reward and punishment indices of the Financial Discrimination Task.

To assess RQ1, we will conduct independent t-tests to assess whether health condition (drug use vs. health concern) influences public stigma and discrimination.

To assess RQ2, we will conduct independent t-tests to assess whether aetiological label (brain disease vs. problem) influences public stigma and discrimination. Here we will focus on the "drug use" health condition only.

To assess RQ3, we will conduct independent t-tests to assess whether attributional judgement (low vs. high stability) influences public stigma and discrimination. Here we will focus on the "drug use" health condition only.

Given the number of analyses, we will set a conservative alpha ($p < .01$) to denote statistical significance. Each test will be followed up with equivalence tests (see Dienes, 2021; Lakens, 2017) with detailed analyses reported in supplementary materials. Equivalence tests use the two one-sided tests procedure to statistically reject the presence of effects large enough to be considered worthwhile. We will use the upper and lower equivalence bounds of $-\Delta L = -.20$ and $\Delta U = .20$ based on the effect size that our design was sufficiently powered to detect. Equivalence will 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.

Potential Exploratory Analyses

It is also possible that these factors may interact with each other (e.g., a 2-way interaction between aetiological label and attributional judgement for the drug use health condition); however, given the number of planned comparisons, interaction effects aligning with the research questions will be reported only within exploratory/supplementary analyses."

4) There is some flexibility in drawing conclusions given a number of DVs are used. Is it possible at this point to relate different DVs to different theoretical questions, i.e. to make clear what conclusions you would draw given different patterns of outcome, and how that relates to the main theory?

Response: We felt that it was essential to include both the questionnaires from Rundle et al. (2021) and Kelly et al. (2021) in order to make direct comparisons but agree that without theoretical questions this can create some flexibility with regards to the conclusions. We have now amended both the Introduction and Table 1 to specifically state which theories our tests would support or contradict. For example, the mixed-blessings model (Haslam & Kvaale, 2015; Kvaale et al., 2013) suggests that the disease model of addiction can differentially influence discrete elements of stigma, such as decreased blame but increased prognostic optimism. Kelly et al.'s (2021) findings were in line with this, whereas Rundle et al. (2021) used a more general stigma questionnaire, so we will be able to pit the findings against this theory to suggest which measures are best suited to assessing whether the disease model influences public stigma towards problematic substance use (and whether or not our findings indeed support that theory).

5) When you say non-significant effects will be followed by equivalence tests, did you mean the interaction effects as well? Or, put another way, if you are going for inference by intervals, presumably you will use that same inferential method for interactions and main effects?

Response: Yes, initially this was the correct inference although our writing could have been clearer. The analytical strategy has now been updated as per above and each test will be followed up with equivalence tests.

Additional Notes:
Please note that we have also amended the order of the questionnaires and tasks – participants will now see the (randomised) questionnaires immediately after their assigned vignette, rather than the financial discrimination task, so that we can conceptually replicate Rundle et al. (2021) and Kelly et al. (2021). Our financial discrimination task will then be placed at the end of the experiment so it cannot influence responses to these questionnaires. We have also clarified some of the writing in the revised manuscript.

Yours sincerely,

Dr Charlotte R. Pennington