Response to Reviewers

Dear Prof. Dienes,

Thank you for the timely review of our Stage 1 Registered Report. Below we outline our response to both your and the reviewer's helpful comments and highlight any changes in the manuscript in red font. We also upload a clean version for re-review.

We'd like to take this time to thank the reviewers for their time and all of the comments that have greatly improved this manuscript.

Yours Sincerely,

Dr Charlotte R. Pennington and co-authors

Editor's Comments

I now have two reviews which are largely positive about your submission.

1) *Sinclair-House asks you to comment on the difference between alcohol and opioids in previous studies and the bearing this may have had on different results; and Giner-Sorolla asks you to comment on the quality of the attributional manipulation if it weren't simply trying to resolve differences between previous studies.*

Response: Please see our response to Point 1 under the subsection 'Reviewer #1 Comments (Sinclair-House)', and our response to Point 1 under 'Reviewer #2 Comments (Giner-Sorolla)', which addresses both of these comments.

2) *Giner-Sorolla raises questions about Bonferroni (interestingly I made almost the exact same point about Bonferroni here https://psyarxiv.com/pxhd2 pp 19-20). On the other hand, given you give yourself some interpretational flexibility in choosing which measure is better simply by whether it yields the conclusion of a difference, I think some statistical conservatism is fine.*

Response: The mention of Bonferroni corrections were only apparent in our initial Version 1 manuscript, before we revised it in line with your helpful comments. We have now removed the two-step procedure of conducting independent samples *t*-tests followed by equivalence tests and have decided to adopt a more stringent alpha level of .01 given the number of analyses being conducted. We agree with you that some statistical conservatism is fine because we aim to use a range of different outcome measures to evaluate where any (potential) differences lie.

3) *You still have a two-step procedure which is inferentially incoherent, i.e. test whether the mean is within or outside the equivalence region only after a signfiicant result against the H0 of no effect. Why not simply test whether the x% CI is completely outside or within (or only partially within) the equivalence region? That is, drop the initial significance test again the H0 of zero effect, an H0 you have implied is of no relevance by claiming there is minimally interesting effect size.*

Response: Upon reflection, and after reading best practice guidance (Dienes, 2021; Lakens, 2017; Lakens et al., 2018), we completely agree. We have now revised our Analysis Strategy to focus on independent samples equivalence testing, as well as revising Table 1 regarding our inferences. The Analysis Strategy now states:

"Independent samples equivalence tests will be conducted on each of our RQs (see Dienes, 2021; Lakens, 2017; Lakens et al., 2018) with detailed analyses reported in supplementary materials. Allowing for direct comparisons between the current study and that of Kelly et al. (2021) and Rundle et al. (2021), these will be conducted 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. 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. Given the number of analyses, we set a conservative alpha ($p < .01$) to denote statistical significance. 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. Table 1 provides a design summary.

RQ1: Does the health condition ('drug use' vs. 'health concern') influence public stigma and discrimination?

RQ2: Does aetiological label ('chronically relapsing brain disease' vs. 'problem') influence public stigma and discrimination towards problematic substance use? Here we will focus on the 'drug use' health condition only.

RQ3: Does attributional judgement (low vs. high treatment stability) influence public stigma and discrimination towards problematic substance use? Here we will focus on the 'drug use' health condition only.

4) *For RQ3 in the design table, do not label it as exploratory (to keep things clean don't describe any exploratory analyses in the Stage 1); for the theory at stake simply state the broadest claim that is at stake given your findings (regardless of whether past research has looked at this not, which is not relevant to whether this study tests that claim). In the row for RQ3 be clear this is testing a difference of differences.*

Response: We have now revised the manuscript to remove any mention of 'exploratory' associated with RQ3 and have removed this from Table 1; for example, we now state: "Given the mixed literature regarding whether the 'brain disease' label lessens or exacerbates public stigma, and the novel inclusion of the attributional judgement factor, we do not make any directional predictions." (Page 11). We have also updated the row for RQ3 to state: "Neither Kelly et al. or Rundle et al. (2021) manipulated attributional judgement in their studies, but this is an additional factor that we recognised as a difference between the two."

5) *Finally, as we have talked about, justify minimal effects of interest by their relevance to the theory tested (in its scientific context) rather than by researcher resources, which have no inferential relation to what a theory predicts.*

Response: We have revised our power analysis, following recommendations of Lakens (2021 - "sample size justification"). It is important to note that whilst Kelly et al. (2021) had a large sample size, they did not conduct a power analysis (neither *a*-priori or sensitivity) and did not specify the magnitude of effect size which was of interest. Moreover, there is no mention of a smallest effect size of interest within this literature at current, and this is something that we'd like to include in our Stage 2 Discussion.

"Our planned sample size is informed by the effect sizes obtained from Kelly et al. (2021) and Rundle et al. (2021) . For our main effects of interest (see "Vignette development" below), Kelly et al. observed a significant effect of Cohen's $d_s \sim .15$ for perceived danger, $d_s \sim .20$ for prognostic optimism, $d_s \sim .30$ for continuing care and $d_s \sim .43$ for blame, whilst Rundle et al. observed an effect of $d_s \sim .1.03$ for Stigma Ratings. We conducted a series of sensitivity power analyses based on the two one-sided tests procedure for equivalence testing (see Dienes, 2021; Lakens, 2017). In the first, we input the smallest significant effect of $-\Delta L = -.15$ and $\Delta U = .15$ from Kelly et al., which requires 2,804 participants to achieve 90% statistical power with alpha set at .01. However, this is outside of our funding resources (see Lakens et al., 2021). For this reason, we then input the second smallest effect of $-\Delta L = -.20$ and $\Delta U = .20$, again from Kelly et al., which requires 1,578 participants ($n = 789$ per group): given that this is within our resources, this determined our planned sample size. Note that effect sizes of $d_s =/> .20$ have also been found in meta-analyses assessing the influence of the brain disease model on public stigma (Kvaale et al., 2013) meaning that the planned sample size would yield informative results with respect to the presence or absence of effect size estimates provided by this meta-analysis."

Reviewer #1 Comments (Sinclair-House)

1) *I believe this RR satisfies the relevant criteria for Stage 1 review. The scientific validity of this research question is clearly demonstrated, particularly given the seemingly conflicting findings of the studies upon which it draws. The logic and rationale of the study are coherently outlined and appear credible. The proposed hypotheses are appropriate and the research falls within established ethical norms in the field. I note the authors' response to comments at the previous review stage and the changes made in light of these. Where changes have been made, these have been beneficial and made the proposed analysis plan more rigorous. Where the decision has been made not to implement a change in line with the reviewer's comments (e.g. minimal effect of interest), this decision seems justified as a balancing of desired statistical power with the practical implications of required sample size. The proposed methodology appears to be appropriate in context and sufficiently detailed to be reproducible.*

Response: We thank the reviewer for their positive appraisal of our manuscript. We agree that the Editor's comments strengthened the methodology and analysis plan, which is an advantage of the Registered Report route. Below we respond to your additional comments.

2) *Whilst I would not necessarily expect to see it addressed in great detail, it is worth noting that one potentially important difference between the Kelly et al. and Rundle et al. approaches which may link to stigma is the use of substances with differing legal and social statuses (opioids and alcohol). Unlike the majority of addictive drugs, alcohol is easily and widely available (and widely used). Leaving aside the separate*

> *question of whether or not that should be the case, it is worth reflecting on the extent to which alcohol being the socially-acceptable face of recreational drug use impacts stigma surrounding its (mis)use. That may prove to be a relevant consideration if you find the suggestion of an effect where Rundle et al. did not.*

Response: When planning this study we had long discussions about this particular point as well as examining the evidence base. You are correct in identifying that Kelly et al. focus on opioid use whilst Rundle et al. focus on alcohol, which creates another difference between them. However, the empirical evidence suggests that both alcohol use and substance use disorders are heavily stigmatised (with A/SUD representing one of the most stigmatised clinical disorders; Kilian et al., 2021; see also Room, 2009), so we decided not to manipulate this difference and further complicate our design. Furthermore, the largest effect size found across the two studies under investigation was actually from Rundle et al. (d = 1.03) who focused on problematic alcohol use (versus a general health condition of diabetes).

We felt that this additional manipulation would unduly affect our research design: creating a fourth research question with additional analyses. On balance then, we believe that it is stronger to manipulate 'drug use concern' versus the control of health concern' rather than including an additional factor of the substance itself (alcohol vs. drug use). We have clarified our rationale now, with footnote 1 stating: "Another difference is that the two studies include different *substances* within the vignette: Kelly et al. opioid use, Rundle et al. alcohol use. Research has consistently shown that both alcohol use and substance use disorder are heavily stigmatised (Kilian et al., 2021) so we do not expect this to explain the different findings. In the current study, we therefore do not manipulate the substance itself".

Reviewer #2 Comments (Giner-Sorolla)

1) *This proposal is a fairly focussed attempt to resolve an apparent discrepancy between two studies asking whether describing drug addiction as a disease improves attitudes and reduces stigma. Differences between the studies are analysed and manipulated directly. The stigma measures are reasonable and the indirect discrimination measure by means of financial reward and punishment is an interesting touch. I think the comments of the editor have been answered thoroughly and in a well-informed way and I am convinced that we have enough statistical power to answer questions of interest.*

Response: Thank you for the positive appraisal of our manuscript: this is really pleasing to hear. Below we address your remaining concerns.

> 7) *My one sticking point, though, concerns the necessity for multiple corrections. As usually applied, Bonferroni corrections address a H0 that is not necessarily of interest, namely that all tests included are null. They are usually justified as a way to guard against one-shot "fluke" findings of significance that come about only because too many chances were taken. However, this pattern should also be evident from an inspection of the space of all findings, and from an honest summary of them. Thus, the conclusion of a study where only one out of six hypothesis tests is confirmed should not be "this one test confirmed our hypothesis, so it is true." In addition to Bonferroni corrections being mathematically unsuited for correlated effects, they also lead to absurdity -- concluding no evidence overall when each of five tests goes in the same direction and is between p = .02 and p = .05, for example. The Holm method is better*

suited for error control, for one, but I would prefer if the reader decides whether or not correction is applied, by modifying the threshold of significance rather than the p-value itself. For further reading see Mark Rubin's 2021 paper in Synthese.

Response: Thank you for this comment. In our revisions based on the Editor's initial comments, we removed any mention of Bonferroni correction because we were no longer testing interaction effects (instead deciding to make our analyses more stringent and attached individually to each research question). In our second revision, we now also drop the planned *t*-tests in favour of equivalence testing against the smallest effect that we have 90% statistical power to detect with alpha set at .01. We have decided to adjust this significance threshold/alpha from the commonly used .05 (in psychology) to .01, due to the number of research questions and dependent variables being used. We have now read the excellent paper by Rubin (2021) which states that "alpha adjustment is also inappropriate in the case of individual testing, in which each individual result must be significant in order to reject each associated individual null hypothesis". Nevertheless, we have decided to stick with our adjusted alpha based on the fact that with our proposed sample size and analytical design we have high statistical power (90%) to detect effects stringently. Lower *p*-values have been found to be more replicable and less likely to be the result of a Type 1 error (e.g., Open Science Collaboration, 2015). In agreement with the Editor, we believe some statistical conservatism is justified in this study.

8) I also should observe that the manipulations, in particular of attributional judgment, are fairly tightly focussed on resolving the conflict between the two previous studies. Looked at independently, the attributional judgment manipulation isn't that clean or obvious as a manipulation only of attributional judgment. I understand that it is derived from the wordings used in the two previous studies but I think the limitation of this approach should be acknowledged.

Response: To recap, the attributional judgement manipulation comprises all of the purple highlighted text in the vignette. That is:

"They are now in a treatment program [and] Alex is committed to doing all that they can to ensure success following treatment" [low attributional judgement]

Vs.

"They have now visited a doctor [and] The doctor tells Alex that this is potentially long-term and could get worse over time, but could also improve if they start treatment now". [high attributional judgement]

Whilst we agree that this is not the strongest manipulation compared to the other two (health concern, aetiological label), we think that it is paramount to test this research question because it represents another (subtle) difference between the manipulations used by Rundle et al. (2021) and Kelly et al. (2021). If we do not test this, support for one study over the other could be argued to be caused by *not* manipulating this factor; similarly, if we deviate from the vignette wording used between these two studies, any findings could be attributed to such wording differences. We therefore want to keep our conceptual replication as close to the original studies as possible, whilst being able to manipulate their differing factors.

In addition, our manipulation checks will allow us to examine and discuss whether participants were aware of the attributional judgement factor or not: manipulation check 3 asks "At the start of the study, you were given a description of a person named Alex. Was Alex: 'now in a treatment program' or 'visiting a doctor?'. If a large proportion of participants fail this manipulation check compared to the other factor's manipulation checks, then we can state in our Discussion that this was not a strong enough manipulation. Alternatively, if a large proportion of participants pass this manipulation check, but we do not find conclusive evidence for RQ3, we can suggest that attributional judgement does not appear to have much effect of stigmatising perceptions in our study, but future research may want to enhance this manipulation (with careful consideration of demand characteristics). We will be sure to include discussion of the limitations of our study in the discussion.

Additional comments

Please also note that, in our revision, we have removed the subsection of exploratory analyses which stated that we may explore interaction effects. The reason for this is that we have enough focused analyses as it currently stands. We would, however, be happy to include relevant exploratory analyses recommended by reviewers at Stage 2. We have also revised some wording ambiguities, with all changes denoted in red font. We have also updated our OSF Project Page with additional details of the power analyses.