Reply to decision letter reviews: #372
PCIRR Stage 2 - Newman et al. (2014) replication and extensions

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response as well as a tally of all the changes that were made in the manuscript.

Please note that the editor's and reviewers' comments are in bold with our reply underneath in normal script.

A track-changes comparison of the previous submission and the revised submission can be found on: https://draftable.com/compare/rHWGQzLlRysQ (https://osf.io/zqvc6)

A track-changes manuscript is provided with the file:
"PCIRR-S2-RNR-Newman etal 2014 replication & extension-main manuscript-track-changes.docx" (https://osf.io/6wb4z)

Response to Editor: Prof. Chris Chambers

> **The four reviewers from Stage 1 kindly returned to evaluate your completed Stage 2 manuscript, and I'm happy to report that their assessments are unanimously positive. As you will see, there are some constructive points to address concerning the reporting of results, clarification of methodological details, and potential issues for inclusion in the Discussion. Provided you are able to respond comprehensively to these points in a revision, I anticipate being able to award Stage 2 acceptance without further in-depth review.**

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit.

Response to Reviewer #1: Dr./Prof. Andrew Christy

> **I have completed my review of the Stage 2 manuscript, which is largely favorable; see the attached Word document. I would also like to thank the authors personally for undertaking this replication project; it is very useful to others, like me, who are working on these topics!**

> **Overall, I believe this study satisfies the Stage 2 criteria. I noted some minor issues in relation to criteria 2C and 2E, but these are quite minor. See below for my specific comments!**

Thank you very much for the positive and constructive feedback.

> - **Criterion 2C: Adherence to Registered Procedures**

> **.1. One set of results included on pg. 38 of the Stage 1 manuscript, namely Study 1: Neutral items and preferences (replication) does not seem to be reported in the Stage 2 manuscript (unless I have just overlooked it somehow). These results are not relevant to any of the central hypotheses, but to the extent that they were included in the Stage 1 manuscript it would seem desirable to include them in Stage 2.**

Thank you. We appreciate you going over the track changes to check that we adhered to the pre-registration Stage 1 plan.

Yes, this was not immediately obvious, but we moved that section to the supplementary materials, since we felt the manuscript was already very long and that this specific section did not add much. These can be found it in the supplementary materials Tables S7 and S8.

> - **Criterion 2E: Justification of Authors' Conclusions**

> **.2. On pg. 41, there is a small error; it is stated that neutral change was rated lower on surface-self compared to good and bad changes – I believe that should say higher, not lower (consistent with the reported means and other descriptive statements about this finding).**

Thank you very much for catching that. We are very grateful. We indeed found some oversights.

In the true-self measure, neutral change was lower than both positive and negative change, and so we changed from:

We again found support for Hypothesis 1 with a larger effect that participants were more likely to attribute true self in morally good changes (M = 68.4, SD = 20.3) than morally bad change (M = 54.9, SD = 24.8, t(801)= 12.27, p < .001) (Figure 4). When compared to neutral change (*M* = 45.7, *SD* = 22.0), true self was more likely to be revealed in morally good change, *t*(801) = 23.56, *p* <.001) while less likely in morally bad change, *t*(801)=8.73, *p*<.001.

To:

We again found support for Hypothesis 1 with a larger effect that participants were more likely to attribute true self in morally good changes (M = 68.4, SD = 20.3) than morally bad change (M = 54.9, SD = 24.8, t(801) = 12.27, p < .001) (Figure 4). When compared to neutral change (M = 45.7, SD = 22.0), true self was more likely to be revealed in morally good change (t(801) = 23.56, p < .001), and in morally bad change, (t(801) = 8.73, p < .001).

The relevant figure to see that pattern is Figure 4:
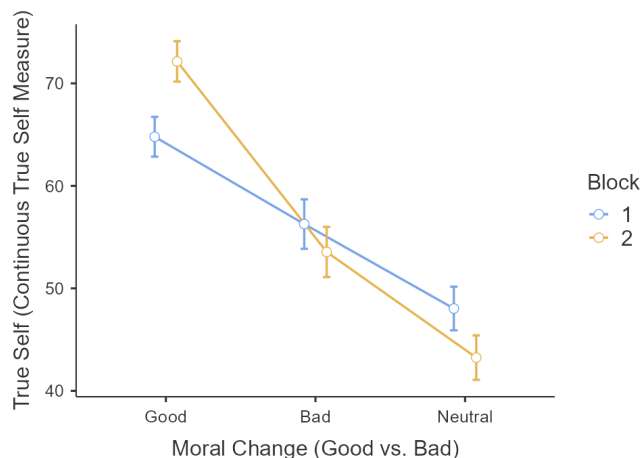

And, indeed, as you indicated, we previously had:

Yet, we found no support for differences between surface self ratings for good change (M = 47.4, SD = 23.6) and bad change (M = 47.6, SD = 24.0, t(801) = -.27, p = .962) (Figure 5). Neutral change (M=54.3, SD=21.8) was rated lower surface self compared to good (t(801) = -7.17, p < .001 and bad change (t(801) = -6.80, p < .001).

Which we revised to the following:

Yet, we found no support for differences between surface self ratings for good change (M = 47.4, SD = 23.6) and bad change (M = 47.6, SD = 24.0, t(801) = -.27, p = .962) (Figure

5). Neutral change (M=54.3, SD=21.8) was rated as <u>higher</u> surface self compared to morally good (t(801) = -7.17, p < .001) and morally bad change (t(801) = -6.80, p < .001).

The relevant figure to see that pattern is Figure 5:
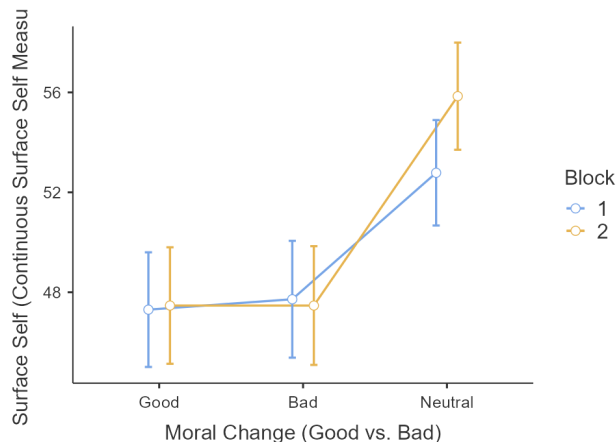
.3. Not related to the conclusions drawn from the current data – but I don't think it's appropriate to cite my 2019 paper (Christy et al., 2019) in support of the claim that essentialist cognition contributes to the belief in good true selves specifically. The studies reported in that article did not examine beliefs about the morality of true selves, but simply the belief that true selves exist; the implication is just that belief in personal true selves may be symptomatic of broader essentialist cognitive tendencies. Some of my other work has found evidence consistent with a baseline assumption of goodness about one's own (Christy et al., 2016) and others' true selves (Christy et al., 2017). Other authors (e.g., De Freitas, Cikara et al., 2017) have suggested that essentialist cognition may somehow result in this belief in good true selves, on the basis of these latter findings and others (such as those replicated here as well as Nina Strohminger's work), as well as evidence that people seem to perceive the essences of all kinds of entities, not just of persons, as being good (De Freitas, Tobia, et al., 2017). So I think that section of the Discussion could be revised to use citations more appropriately in support of the authors' claims.

Thank you. We appreciate the feedback.

We amended the citations to point to Christy et al., 2016 and Christy et al., 2017. We also went through the rest of the citations and updated references accordingly.

Response to Reviewer #2: Dr./Prof. Cillian McHugh

> **The authors conducted the study in accordance with the approved Stage 1 protocol. They provide interesting results, replicating and extending the target article. I commend the authors on this work.**

Thank you very much for the positive and constructive feedback.

> **.1. I have only 1 comment. Perhaps the authors could provide a bit more clarity on number of participants, the exclusions, and the attention checks. The authors report 803 took part and 44 were excluded. From reading the results reported, it appears that the 803 reflects the sample after the 44 were excluded? Have I understood correctly? Some clarity on this would be helpful.**
>
> **Is it possible to provide a breakdown of the number of exclusions for specific reasons? Does the "verification" refer to the attention checks or are they separate? It is not clear how participants who failed the attention checks are handled in the reporting.**
>
> **These questions are for clarity only. I have no real substantive concerns, I just think a bit more detail and clarity might be useful.**

Thank you. This needed further clarification, and so we added the following with generally more details about the participants and how they were recruited under subsection "Participants":

> We recruited a total of 803 U.S. American participants from Amazon Mechanical Turk American using CloudResearch MTurk Toolkit (Litman et al., 2017; Mage= 43.18, SD=12.76; 398 females, 393 males, 13 preferred not to disclose/other). We recruited participants with Approval Rate to be between 95% and 100% and number of tasks approved between 5,000 and 100,000. We employed the following CloudResearch MTurk Toolkit options which were considered best-practices at the time to ensure high-quality sample: Duplicate IP Block, Suspicious Geocode Block, Verify Worker Country Location, and we recruited only from CloudResearch Approved Participants. We note that 847 subjects began the survey but 44 did not proceed beyond the consent and verifications. We summarized a comparison of study characteristics between the target article and the replication in Table 4.

Also, to explain what we meant by "verification" we also added the following to the "Design and procedure" section:

Participants first read a consent form and indicated their willingness to participate, and then answered several verification questions. Participants first indicated their consent, with three questions confirming their eligibility, understanding, and agreement with study terms, which they had to answer with a "yes" and the required responses in order to proceed to the study. The three questions also served as attention checks, with a randomized display order of the options - 1) "Are you able to pay close attention to the details provided and carefully answer questions that follow?" (yes/no/not sure), 2) "Do you understand the study outline and are willing to participate in a survey with comprehension checks?" (yes/no/not sure), and 3) "Are you a native English speaker born, raised, and currently located in the US?" (yes/no). Failing any of the three questions meant that the participants did not indicate consent and therefore could not embark on the study. Upon completion of these steps, participants proceeded to begin the survey.

Therefore, to be clear, there were no exclusions. The 44 simply did not embark on the study or did not complete the study.

Response to Reviewer #3: Dr./Prof. Caleb Reynolds

This Stage 2 manuscript seems to tightly adhere to what was promised in the approved Stage 1 revised manuscript. The Introduction, including hypotheses, is nearly identical to the Stage 1 version. To my eye, the studies as reported adhere to the protocol presented in the revised Stage 1 manuscript, and I do not notice any deviations. I see no concerning patterns in the manipulation checks which would undermine interpretation, and the conclusions that the authors draw are circumspect and consistent with the evidence offered.

Consequently, I'm satisfied with the registered report in its current form, and I again commend the authors on their attention to detail. I offer a few minor optional tweaks below.

Thank you very much for the positive and constructive feedback.

.1. Abstract: The authors might specifically add "the preregistration" to the final sentence identifying the elements available on OSF

Thank you, revised to:

Pre-registration, materials, data, and analysis code were made available on
https://osf.io/9fvtq/

.2. Possible typo on p. 51: "shift between the two studies in the target *article*"?

Thank you for catching that, corrected.

.3. The OSF page might benefit from some brief explanatory text in the Wiki, as there are many files with similar names (because of the way the registration snapshots display) and I could imagine an unfamiliar reader being a bit lost

Yes, we agree, the OSF does not have the clearest interface. The OSF since changed the way they display files (now under the Files tab, not displaying pre-registration files of the pre-registration component), and they vowed to make additional changes soon, so it is difficult to anticipate what would work best. However, to atleast clarify the files that are in the project Files section we now include a README.docx file that clarifies the different files in the Data and code subdirectory.

We will also update the OSF page wiki to refer to the latest preprint.

Response to Reviewer #4: Dr./Prof. Sergio Barbosa

> **Authors designed and carried out a well-crafted replication and expansion. As is, I believe manuscript ought to be almost ready for publication. I have very few comments, none of which should be much trouble for them.**

Thank you very much for the positive and constructive feedback.

> **.1. I should have picked that up on the first round review but I just realized that stating political preferences BEFORE main data collection might bias or skew main data collection by people try to be coherent to that self-proclaimed identity or some sort of demand effect. Main data collection is quite long and effects are really solid so as to not be significantly changed by this possible bias. I don't think this is any reason to be really worried about, but one never knows with da stricter reviewer. Perhaps consider this for limitations section or come up with an possible response in case it is needed.**

Thank you for raising this. This helped us realize that our reporting may have led to a misunderstanding. To clarify, political views were collected at the end. We wrote in the "Design and procedure":

> After completing both experiments, participants rated their political views (used in the replication of Study 2) and their generalized lay-beliefs regarding true self as inherently good and inherently bad (extension).

(which can be easily verified by examining the shared Qualtrics export Word DOCX files, shared on the OSF)

We tried to find what may have led to the misunderstanding and realized that in our Tables 5 and 6 the individual differences appeared on the top, so this may have been interpreted as being presented first. To address this, we moved it to the bottom of Tables 5 and 6 and indicated clearly which sections were - "[(presented at end of Study 1)]", "[presented at the end after both studies completed)]", or "[presented after both studies completed)]".

> **.2. Authors claim that analysis wit excluded participants was not run because hypothesis were suported "Since we find support for all the hypotheses, rerunning analyses with exclusions is not needed." (p 27). I would beg to differ in that point, exclusions are there to make sure suitable data is analyzed. The reason analysis should not be run with excluded data is taht you have reason to believe that is somehow biased irrespective of subsequent results.**

It is important to put that sentence in context - we followed the pre-registered Stage 1 plan of not reporting exclusions if hypotheses were supported. "Not needed" was referring to reporting following our adherence to the pre-registration plan.

Yet, we appreciate the implicit suggestion to try and do better here, and to alleviate possible concerns about issues.

We adjusted our RMarkdown code to include a parameter that outputs the same analyses using the exclusions, and made the code with the knitted Rmarkdown outputs with the exclusions available on the OSF together with the results of the no-exclusions main analyses (filenames: RRR-Newman-etal-2014-data-analysis-with-exclusions.html/.docx). Exclusions had no major impact on the results or the conclusions reported.

We now revised to the following:

> We focused our analyses on the full sample of all participants who completed the study. We had planned to report analyses with exclusions if we failed to find support for the hypotheses (our planned exclusions were: 1) Participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale), and 2) participants who self-report not being serious about filling in the survey (self-report < 4, on a 1-5 scale). Given that we found support for all the hypotheses, we follow the pre-registered Stage 1 plan and do not report additional analyses with exclusions. As an additional exploratory analysis, with our code on the OSF, we also provided the results of our analyses with applying the exclusions, and these had no impact on our findings.

.3. Table 10 comparing results to hypothesis are not particularly straightforward to read. I take it that "signal" means that results replicated whereas "inconsistent" means results somehow differ from original results right? These terms are not easily understood and surely whether results are "signal" or not is linked to the amount of noise in observed data, not in whether they replicated prior results. I suggest changing this.

In the Table 10 note we indicated that this is based on the most commonly used criteria by LeBel et al. (2019) and using their terminology:

> The interpretation of outcomes was based on LeBel et al. (2019) on actual data. Details please see the section under "Additional tables and figures" of supplementary material.

Their article is pretty straightforward, yet as we indicated in the main manuscript - to aid readers we still included a summary of that in our supplementary materials Figure S2 to help explain it visually:

Figure S2

LeBel et al. (2019) replication evaluation criteria
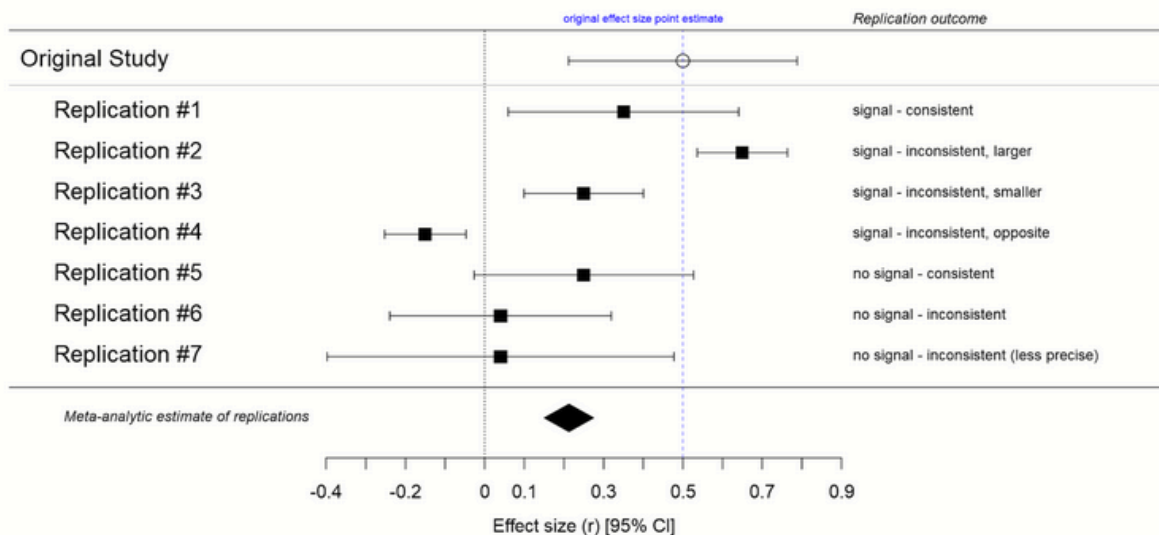
Note. Interpretation criteria for evaluation of replications outcomes by LeBel et al. (2019), if the original study detected a signal. A simplified replication taxonomy for comparing replication effects confidence intervals to target article original effect sizes.

.4. I was surprised by interaction effects of block X Moral. Maybe a bit more discussion could be offered on this. As I understood it this is not exactly expected and could be due to the choice of using blocks rather than the optimal full randomization procedure which ought to be discussed.

We understand the inclination to try and interpret this, yet we do not think that the interaction effect of the block display is of any meaningful significance and we see no need to try and over-interpret this exploratory and inconsequential finding. The pattern of findings does not vary much or change any of the conclusions. It was not expected, but it was also not unexpected, as we already raised this very methodological issue in Stage 1, and given that it had no impact on the conclusions.

We already included a very brief discussion of our view regarding this issue that this choice of blocks is indeed not ideal and should be addressed in future research:

> In our initial submission we raised concerns regarding the methodological choice in Study 1 to fix the display of items so that each block first displays four positive (/negative) changes together and then four negative (/positive) change vignettes together, followed by four neutral vignettes, which the original authors explained as contrasting certain changes against each other. There were some minor block-order effects that did not seem to impact the overall pattern of results, yet in future research it might be better to randomize the display of the vignettes within each block.

> **Other than these comments I believe this manuscript is readily suitable for publication and expect it to be accepted easily. I want to congratulate the authors on a rigourous and interesting work and look forward to seing this and their subsequent projects published.**

Thank you, that is very kind, we really appreciate the feedback.