# Ms Title: Self-interest is overestimated: Two successful pre-registered replications and extensions of Miller and Ratner (1998)‎

**Author Names:** Cameron Brick, Adrien Fillon, Siu Kit Yeung, Meiying Wang, Hongye Lyu, Jing Yee Jane Ho, Sze Chai Wong, Gilad Feldman

# Submitted: Jan 26, 2021

**Editor First Decision—Revise & Resubmit**

Feb 17, 2021

Dear Dr. Feldman,

I have now reviewed your manuscript, “Self-interest is overestimated: Two successful pre-registered replications and extensions of Miller and Ratner (1998)‎” as well as the previous reviews and decision you submitted for streamlined review, and your response to the reviewers’ comments. I found your manuscript to be quite strong, and found only relatively straightforward issues to address. If you can address these issues, I would be happy to accept your manuscript for publication at Collabra: Psychology. I therefore encourage you to submit a revised version. In your resubmission, please include a document with a point-by-point response to the points I list here, outlining each change made in your manuscript or providing a suitable rebuttal.

One issue that is not completely resolved, to my mind, is whether “overestimation” in the kind of paradigm used here is really a sign of error/bias. It’s possible that people’s estimates of others’ behaviors are accurate, and it’s their own hypothetical self-reports that are inaccurate. You acknowledge this, but I don’t think it’s laid out as clearly as it could be in the introduction or the discussion. I think it would be helpful to readers if you could state the issue more explicitly, namely, that self-reports of hypothetical behavior could be wrong (e.g., because of impression management, self-deception, or other reasons), and therefore the discrepancy between the self-reports and the estimates of others’ behavior could reflect inaccuracy in either part of the comparison. You are choosing to describe this discrepancy as “overestimation” of others’ self-interest, and I think that’s reasonable, but I would like you to be more explicit that this interpretation rests on some controversial assumptions (i.e., taking the self-reports at face value). In addition to this ambiguity, there is another related challenge in interpreting the overestimation effects for Study 1, to my mind, which is that the two things being compared are not exactly equivalent. The self-reports are about hypothetical behavior, but from what I understand the estimates of what others would do are estimates of actual behavior, not estimates of how willing others would be to say that they would, hypothetically, donate blood (which would be a better match to what the self-reports are measuring). It seems plausible to me that if we asked participants to estimate how many other people would be willing to say yes to the hypothetical question, their estimates would be more accurate.

I would like to see more information about how the overestimation statistic (second data column of Table 5) was calculated in Study 1. You provide this info for Study 4 on page 24, but I did not see an equivalent explanation for Study 1. Moreover, I found the reporting of these results in the text and in the table (p. 17 and Table 5) quite hard to follow. In Table 5 and Table 7, please add information in the table note about what the M (second data column) is a mean of. Please also add a label to the fifth data column (between the t column and the 95% CI column) in Table 5 – I assume it is a d but the label is missing. In the text, on page 17, I could not follow what the t-tests were comparing. For every t-test you report (for both studies), please state explicitly what is being compared to what. In addition, in the last paragraph on page 17, I could not understand what the difference was between the second-to-last t-test reported (d = .85) and the last t-test reported (d = .58). The last one matches what is in Table 5, but I am not sure what the second-to-last one is testing, or how it fits with the rest of the results reported on this page and in Table 5.

Related to the point above, I am concerned about the one-sample t-tests you conducted. I am not sure what they are all comparing (for some, e.g., the last one reported on page 17, it is clearer than for others), but I think that in some cases, you are comparing participants’ estimates of what others would do to a specific fixed value based on participants’ self-reported willingness to donate blood (or, in the case of Study 4, support for a policy). If this is in fact what some of the one-sample t-tests are doing, I think this is problematic, because you are treating a value that is estimated with error (participants’ self-reported willingness to donate blood or self-reported support for a policy) as if it is a fixed value whose quantity is exactly known, without error. In other words, I don’t think these tests should be one-sample t-tests, I think you should be treating both sides of the comparison (the participants’ estimates of others’ answers, but also the participants’ self-reported positions) as estimated variables, with error. I’m not sure what the most appropriate test is to do this, but I think some kind of t-test would work.

Just in case my explanation above is unclear, let me try to illustrate the problem with doing a one-sample t-test against a specific value that is actually estimated with error. Take the case where you are comparing participants’ estimates of what heavy smokers would say about the policies to the actual answers from heavy smokers (p. 26 & table 8). Here, it is very clear that the estimate of heavy smokers’ level of support for the policies is very imprecise – this point estimate of 70.8% support (averaged across all policies) is based on only 6 respondents’ answers. Thus, the true value could be very different from this, and it would be imprudent to use the point estimate from just 6 heavy smokers as a perfectly-accurate measure of heavy smokers’ actual support for these policies. The same applies even when the uncertainty is not quite as extreme – it is not a good idea to take a statistic that is estimated with error and treat it as a precise value without error.

It is definitely possible that I am misunderstanding what you did in the one-sample t-tests, or that my concerns are misplaced for other reasons (e.g., my lack of expertise in statistics). Please feel free to push back on this if I am wrong, but I would like you to look into this issue and address it somehow, and also, regardless of what else you do, explain more clearly what values the one-sample t-tests are comparing.

You mention that the order of the studies was randomized to minimize order effects, but I could not find a statement about whether there were order effects (I also looked through the supplement and did not find this). Could you please add a short summary of whether there were order effects, and if so, what the results look like when you restrict the analyses to participants who took that study first?

It sounds like the Prolific sample is the same sample used in another published paper (Palan & Schitter, 2018). If this is the case, please make this clearer upfront (e.g., at the beginning of the method section), as this is relevant information for meta-analyses and other meta-scientific purposes.

In the Study 1 method, you refer to “blood supply in the United States” and “the American Red Cross” – was this language changed for the UK sample? Please clarify.

On page 14, I would suggest changing the heading that is currently called “Measures” to “Measures and Procedures” as this is also where you describe the order in which participants completed the measures. Also, I think your description suggests that participants completed the communalism measure first, then reported their willingness to donate blood, then estimated others’ willingness to donate blood, but this isn’t 100% clear.

You mention that one way to compare replication effects with original effects is to examine “consistency” which you define as “whether the 95% CIs cover the original effect size.” I realize you then explain why you do not report consistency. However, I think this operationalization of consistency is problematic because a replication could have a very narrow 95% CI which could lead to concluding a result is “inconsistent” even if the original effect size is very close to the replication effect size. For this reason, I think an approach like Simonsohn’s small telescopes approach is preferable. I realize that is not very relevant to your manuscript since you don’t report consistency, but the approach that you promote could influence others so I would like you to consider changing how you define consistency.

I could not understand Figure 2. The legend only includes two shades of gray but the figure seems to include three. I think this figure is quite hard to read even without this issue, so I would suggest an alternative way to visualize these results.

The description of the smoking status manipulation (manipulation 2) on page 20 was difficult to follow. I had a few questions, but rather than addressing my specific questions it might be better to try writing the description again from scratch, and checking if someone who is not familiar with your study understands it. Here were my questions: Was this manipulation only applied to people who were randomized to the self-report condition? If not, it is not clear what this manipulation looks like for participants in the estimate-others condition. In addition, I did not understand this part “if yes, participants were asked whether they were a light, medium, or heavy smoker, yes or no [could they say yes to more than one? or no to all of them? Or did they have to pick just one?], or to a five-item ordinal scale” – were there two different ways of asking about heaviness of smoking? One was a yes/no set of response options and one was an ordinal scale? This seems unlikely (that you had two different ways of operationalizing the more fine-grained measure of smoking), so I am sure I am misunderstanding something. Please rewrite this section.

Do you think there could be ceiling effects in Study 4? It seems that support for the policies is quite high. I’m not sure how this would affect the interpretation of your results, but seems worth discussing.

Smaller points: -On p.9, you write “research practices that were normal at the time but are now recognized as inflating false positives” I would suggest adding “and effect sizes” because that is especially relevant to the point you are making here. -p. 14, “More communal individuals overestimation” should be “More communal individuals overestimate”

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at [editorialoffice@collabra.org](mailto:editorialoffice@collabra.org).

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Simine Vazire Editor in Chief Collabra: Psychology

**Author Response**

Apr 2, 2021

Dear Prof. Vazire,

Thank you very much for your thoughtful and detailed review. Please see our attached revision and point-by-point response below.

There were no major changes. We replaced the confusing histograms, added text to clarify the t-tests, included order analyses (no effects), and fixed other ambiguities and issues. The only point we pushed back on is the appropriateness of the one-sample t-tests. We are happy to discuss this point further if appropriate.

The word count increased by 4%. All code and figures have been updated at OSF. We also ran the manuscript through spellcheck and statcheck again.

**Editor comments**

I have now reviewed your manuscript, "Self-interest is overestimated: Two successful pre-registered replications and extensions of Miller and Ratner (1998)‎" as well as the previous reviews and decision you submitted for streamlined review, and your response to the reviewers’ comments. I found your manuscript to be quite strong, and found only relatively straightforward issues to address. If you can address these issues, I would be happy to accept your manuscript for publication at Collabra: Psychology. I therefore encourage you to submit a revised version. In your resubmission, please include a document with a point-by-point response to the points I list here, outlining each change made in your manuscript or providing a suitable rebuttal.

1. One issue that is not completely resolved, to my mind, is whether “overestimation” in the kind of paradigm used here is really a sign of error/bias. It’s possible that people’s estimates of others’ behaviors are accurate, and it’s their own hypothetical self-reports that are inaccurate. You acknowledge this, but I don’t think it’s laid out as clearly as it could be in the introduction or the discussion. I think it would be helpful to readers if you could state the issue more explicitly, namely, that self-reports of hypothetical behavior could be wrong (e.g., because of impression management, self-deception, or other reasons), and therefore the discrepancy between the self-reports and the estimates of others’ behavior could reflect inaccuracy in either part of the comparison. You are choosing to describe this discrepancy as “overestimation” of others’ self-interest, and I think that’s reasonable, but I would like you to be more explicit that this interpretation rests on some controversial assumptions (i.e., taking the self-reports at face value).

Thanks for highlighting this assumption. We agree that exploring this more would benefit readers and the interpretation of these findings. This section was added to the introduction:

*"Importantly, self-reported willingness was used as a proxy for behavior in the original and our replication, and therefore any discrepancy between estimates and reported willingness could also be explained by differences between reported willingness and actual behavior. However, the overestimation interpretation is most plausible based on evidence from different designs (Vuolevi & Van Lange, 2009) and findings that people think of others as more selfish and less fair than themselves (Van Lange & Sedikides, 1998)."*

And this was added to the Discussion under 'Alternative Explanations':

*"Self-reported willingness to donate blood or endorse smoking policies is not equivalent to objective behaviors like blood donation or voting. The main narrative in this paper is that people over-estimate others' self-interest, but the results are also consistent with the pattern that such estimates are accurate and that self-reported willingness is inaccurate; that in actual behavior people would manifest more self-interest than they expect or are willing to report. Further studies with observed behavior would be valuable for testing this account."*

In addition to this ambiguity, there is another related challenge in interpreting the overestimation effects for Study 1, to my mind, which is that the two things being compared are not exactly equivalent. The self-reports are about hypothetical behavior, but from what I understand the estimates of what others would do are estimates of actual behavior, not estimates of how willing others would be to say that they would, hypothetically, donate blood (which would be a better match to what the self-reports are measuring). It seems plausible to me that if we asked participants to estimate how many other people would be willing to say yes to the hypothetical question, their estimates would be more accurate.

Yes, we see this difference too, yet we would like to note that we believe that it is important that replications follow assumptions and framing of the chosen target article as closely as possible, and leave addressing issues in the original design to follow-up investigations and reviewers. Therefore, to fulfill the objective of a close direct replication, we adopted this wording on purpose. However, we agree and welcome an opportunity to address possible current-day misinterpretations of Study 1 results (more than Study 4).

Therefore, we added the following to the Summary of Study 1:

*"However, the design of Study 1 has an additional confound in that participants were asked to estimate others' donation, not estimate their self-reported willingness to donate; asking the latter could have yielded more accurate estimates."*

This point was also mentioned in the transition to Study 4.

1. I would like to see more information about how the overestimation statistic (second data column of Table 5) was calculated in Study 1. You provide this info for Study 4 on page 24, but I did not see an equivalent explanation for Study 1. Moreover, I found the reporting of these results in the text and in the table (p. 17 and Table 5) quite hard to follow. In Table 5 and Table 7, please add information in the table note about what the M (second data column) is a mean of. Please also add a label to the fifth data column (between the t column and the 95% CI column) in Table 5 – I assume it is a d but the label is missing. In the text, on page 17, I could not follow what the t-tests were comparing. For every t-test you report (for both studies), please state explicitly what is being compared to what. In addition, in the last paragraph on page 17, I could not understand what the difference was between the second-to-last t-test reported (d = .85) and the last t-test reported (d = .58). The last one matches what is in Table 5, but I am not sure what the second-to-last one is testing, or how it fits with the rest of the results reported on this page and in Table 5.

Thanks for revealing these gaps. We added this text to the table captions:

Table 5

*"Overestimation M was calculated by taking the mean difference of estimates between paid and unpaid conditions and subtracting the mean difference between paid and unpaid conditions of self-reported willingness."*

Table 7

*"Overestimation was calculated by subtracting the estimate for smokers from nonsmokers within each policy, and then taking the mean of all policies within participants to yield estimated self-interest. From this value, we subtracted the actual discrepancy between smokers and nonsmokers from the self-reported condition."*

We also added information regarding the t-tests on pp. 17-18 before Table 5 and 7 and added the Cohen’s *d* label. The second-to-last t-test is a test of difference in self-interest between self and others regardless of the payment, and the last t-test is between overestimation and 0.

1. Related to the point above, I am concerned about the one-sample t-tests you conducted. I am not sure what they are all comparing (for some, e.g., the last one reported on page 17, it is clearer than for others), but I think that in some cases, you are comparing participants’ estimates of what others would do to a specific fixed value based on participants’ self-reported willingness to donate blood (or, in the case of Study 4, support for a policy). If this is in fact what some of the one-sample t-tests are doing, I think this is problematic, because you are treating a value that is estimated with error (participants’ self-reported willingness to donate blood or self-reported support for a policy) as if it is a fixed value whose quantity is exactly known, without error. In other words, I don’t think these tests should be one-sample t-tests, I think you should be treating both sides of the comparison (the participants’ estimates of others’ answers, but also the participants’ self-reported positions) as estimated variables, with error. I’m not sure what the most appropriate test is to do this, but I think some kind of t-test would work.

This is a tricky point. We talked as a team and did some reading about variance before responding. We compare to a fixed quantity for two main reasons: 1) because this is comparable to how it was done in the original paper, and 2) because in both Study 1 and Study 4, the focal willingness measures are binary—yes or no—and therefore there is not a normal distribution of willingness to test against. Rather, each participant/observation has two estimated percentages (e.g., paid "80%", unpaid "70%") and two dichotomous willingness responses (e.g., paid "yes", unpaid "yes"). The variance of a dichotomous distribution is a direct function of the mean (the proportion). E.g., even if the two blood donation willingness responses were combined across payment condition within participants, we would only have values of 0, .5, and 1, which are not appropriate for a t-test.

A separate issue is that the self-report and other-estimates in Study 4 were completed by different participants, which means that we cannot conduct a standard t-test between both distributions in Study 4 because there are no observations in which both willingness and estimates occur. One could use some sort of yoked analysis based on demographics but we suspect this would introduce more assumptions and therefore complexity than it would be worth.

Because of the above, the one-sample t-tests do seem like an appropriate compromise to test the estimates (continuous distribution) within each self-interest category (paid/unpaid; smoker/nonsmoker) against the self-reports (binary).

We believe that issues like these should be left for follow-up literature to discuss and possibly address, rather than by replicators, focusing on simply repeating what the original has done and examining reproducibility and replicability.

Just in case my explanation above is unclear, let me try to illustrate the problem with doing a one-sample t-test against a specific value that is actually estimated with error. Take the case where you are comparing participants’ estimates of what heavy smokers would say about the policies to the actual answers from heavy smokers (p. 26 & table 8). Here, it is very clear that the estimate of heavy smokers’ level of support for the policies is very imprecise – this point estimate of 70.8% support (averaged across all policies) is based on only 6 respondents’ answers. Thus, the true value could be very different from this, and it would be imprudent to use the point estimate from just 6 heavy smokers as a perfectly-accurate measure of heavy smokers’ actual support for these policies. The same applies even when the uncertainty is not quite as extreme – it is not a good idea to take a statistic that is estimated with error and treat it as a precise value without error.

Thank you for this helpful explanation. We agree that ignoring variance is not ideal and that one-sample t-tests may risk misinterpretation in small samples. We are aware of this. For example, we took pains not to analyze the smoker subgroups because of small sample sizes and used no p-values in that section.

In the revision, we modified that section to further avoid misinterpretation (new text in bold):

*"The rarity of light (n = 15), moderate (n = 18), and heavy (n = 6) smokers meant there was not enough power for inferential tests on actual vs. estimated policy support in the extension. However, visual analysis suggests that overestimation may be most pronounced when individuals consider others with stronger vested interests (here: more frequent smokers). For example, the actual policy support of moderate smokers was M = 72% and heavy smokers M = 71%, but others sharply underestimated those values (moderate smoker support estimated at M = 36%; heavy smokers M = 27%).* ***Any overestimation effect in these groups should be treated with special caution due to the small samples and the analytic choice to use one-sample t-tests against a certain value, since this does not include variance underlying that value's estimate."***

Since the other samples are much larger, for example the main findings in Study 1 and 4, we have less concern about those means mis-representing those distributions. However, we agree with the abstract point.

It is definitely possible that I am misunderstanding what you did in the one-sample t-tests, or that my concerns are misplaced for other reasons (e.g., my lack of expertise in statistics). Please feel free to push back on this if I am wrong, but I would like you to look into this issue and address it somehow, and also, regardless of what else you do, explain more clearly what values the one-sample t-tests are comparing.

We added wording in the result sections around the one-sample t-tests to help explain where the estimates came from and what each analysis is specifically testing.

We would just note that in our replications we focus on and emphasize following the target's closely, yet it is possible that there are better ways to address such issues. We could be missing something too. We will gladly further examine any further issues.

1. You mention that the order of the studies was randomized to minimize order effects, but I could not find a statement about whether there were order effects (I also looked through the supplement and did not find this). Could you please add a short summary of whether there were order effects, and if so, what the results look like when you restrict the analyses to participants who took that study first?

Here, and in our many other replications employing a similar design, we found no indication for order effects.

In the revision we added these summaries about (no) order effects:

Study 1: *"*Order effects. *Study 1 overestimation was not different based on the order of first completing Study 1 vs. Study 4, t(189) = -0.56, p = .58, 95% CI of the overestimation difference -8.89, 4.95."*

Study 4: ***"***Order effects*. Study 4 overestimation was not different based on the order of first completing Study 1 vs. Study 4, t(788) = 1.68, p = .09, 95% CI of the overestimation difference -0.40, 5.37.****"***

1. It sounds like the Prolific sample is the same sample used in another published paper (Palan & Schitter, 2018). If this is the case, please make this clearer upfront (e.g., at the beginning of the method section), as this is relevant information for meta-analyses and other meta-scientific purposes.

The samples have not been published before. This was a misunderstanding based on our earlier text. We revised this to read:

*"... for demographics of the overall Prolific population, see (Palan & Schitter, 2018). For example, the modal Prolific panel member in the larger population was employed full-time and had a bachelor's degree."*

1. In the Study 1 method, you refer to “blood supply in the United States” and “the American Red Cross” – was this language changed for the UK sample? Please clarify.

We updated the manuscript to clarify:

*"Both the US and UK samples saw the same text about the American Red Cross."*

1. On page 14, I would suggest changing the heading that is currently called “Measures” to “Measures and Procedures” as this is also where you describe the order in which participants completed the measures. Also, I think your description suggests that participants completed the communalism measure first, then reported their willingness to donate blood, then estimated others’ willingness to donate blood, but this isn’t 100% clear.

Thank you. The header was changed. It is correct that communalism was measured first (in that extension condition).

1. You mention that one way to compare replication effects with original effects is to examine “consistency” which you define as “whether the 95% CIs cover the original effect size.” I realize you then explain why you do not report consistency. However, I think this operationalization of consistency is problematic because a replication could have a very narrow 95% CI which could lead to concluding a result is “inconsistent” even if the original effect size is very close to the replication effect size. For this reason, I think an approach like Simonsohn’s small telescopes approach is preferable. I realize that is not very relevant to your manuscript since you don’t report consistency, but the approach that you promote could influence others so I would like you to consider changing how you define consistency.

Thanks for sharing this potential concern. We followed best practices in the field for evaluating and categorizing replications, and this would be an important issue to raise in a follow-up discussion with those scholars regarding what seems like the main categorization emerging as the leading standard.

This is indeed not central to the current work, and we therefore chose to revise the text to be less definitive about what Consistency means:

*"There are normally three components to the interpretation. Signal indicates a significant effect, Consistency is whether the effect size is comparable, e.g., whether 95% CIs cover the original effect size, and Direction clarifies the direction of any inconsistencies."*

1. I could not understand Figure 2. The legend only includes two shades of gray but the figure seems to include three. I think this figure is quite hard to read even without this issue, so I would suggest an alternative way to visualize these results.

Thank you for the feedback. We realized it was confusing that the unlabeled third color meant "both samples". We now present these histograms into two separate plots (MTurk, Prolific) for each of Figures 2 and 4.

1. The description of the smoking status manipulation (manipulation 2) on page 20 was difficult to follow. I had a few questions, but rather than addressing my specific questions it might be better to try writing the description again from scratch, and checking if someone who is not familiar with your study understands it. Here were my questions: Was this manipulation only applied to people who were randomized to the self-report condition? If not, it is not clear what this manipulation looks like for participants in the estimate-others condition. In addition, I did not understand this part “if yes, participants were asked whether they were a light, medium, or heavy smoker, yes or no [could they say yes to more than one? or no to all of them? Or did they have to pick just one?], or to a five-item ordinal scale” -- were there two different ways of asking about heaviness of smoking? One was a yes/no set of response options and one was an ordinal scale? This seems unlikely (that you had two different ways of operationalizing the more fine-grained measure of smoking), so I am sure I am misunderstanding something. Please rewrite this section.

We rewrote this section for clarity:

*"This extension is about a more precise measurement of smoking status. In the Prolific sample and the original study in the own-attitudes condition, participants reported their smoking status as "yes" or "no". In the MTurk sample only, participants were randomized either to that binary choice or to a five-item ordinal scale: 1 (nonsmokers: never smoked for more than 6 months), 2 (former smokers: not smoking currently, but having smoked for more than 6 months), 3 (light smokers: <10 cigarettes per day), 4 (moderate smokers: 10-20 cigarettes per day), or 5 (heavy smokers: >20 cigarettes per day). Similarly, in the Prolific sample and the original in the others-attitudes condition, participants estimated the policy support of others based on others' smoking status as "smoker" or "nonsmoker", but the MTurk sample only, participants were randomized either to that same design or to estimate others' policy preferences for each of the five categories above."*

1. Do you think there could be ceiling effects in Study 4? It seems that support for the policies is quite high. I’m not sure how this would affect the interpretation of your results, but seems worth discussing.

We are not too concerned about this, but happy to highlight it for the reader. In the Results for Study 4 we added this:

*"Note that policy endorsement was generally high, which could represent a possible ceiling effect."*

In the Discussion we added this interpretation in the section Constraints on Generality:

*"Additionally, there was a possible ceiling effect in self-reported policy endorsement in Study 4. This could have led to an artificially smaller difference between estimates and self-reported preferences due to the specific policies. That is, for a different set of policies, one might observe even more overestimation without this restriction in range."*

1. Smaller points:  
   -On p.9, you write “research practices that were normal at the time but are now recognized as inflating false positives” I would suggest adding “and effect sizes” because that is especially relevant to the point you are making here.

This now reads:

*"... are now recognized as inflating effect sizes and false positives ..."*

-p. 14, “More communal individuals overestimation” should be “More communal individuals overestimate”

Fixed: thank you.

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

We are pleased—this review cycle boosted the rigor and clarity of the manuscript. Thank you for your constructive and supportive feedback, and we look forward to hearing back from you after further consideration.

**Editor Final Decision—Accept**

Apr 10, 2021

Dear Dr. Feldman,

I have now had a chance to read over your manuscript “Self-interest is overestimated: Two successful pre-registered replications and extensions of Miller and Ratner (1998)”, along with the letter describing the changes you made. Thank you for your thorough responses and revisions based on the points I raised. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

I’ll add that I think the issue of the one-sample t-test is an interesting one, and I appreciate your careful attention to it, and also the fact that you highlighted at the beginning of your response letter that this was an issue you pushed back on - that made it very easy for me to get an overview of what changed and what didn’t, and I rarely see authors be this upfront about the overall shape of the revision. In the end, I am reassured by the fact that the peer review history will be published with this manuscript, so interested readers can see what our discussion about this issue looked like, and what factors you considered in deciding what to do in the revision. This seems like a good compromise - the complex issues we grappled with won’t be lost, and other readers can see the process and rationale.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. Our managing editor will contact you in case there are any pre-prodution file related questions. You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely, Simine Vazire Editor in Chief Collabra: Psychology