January 28, 2018

Natasha V. Raikhel
Interim Editor-in-Chief, PNAS
500 Fifth Street NW
NAS 338
Washington, DC 20001 USA
Phone: 202-334-2679

Dear Dr. Raikhel,

We were delighted to receive the reviewers' valuable and constructive feedback and would like to thank them for their time and thoughts. Below we provide a point-by-point response to the concerns they raised as well as descriptions of how they have been incorporated in our revised manuscript.

We hope to have addressed the reviewers' concerns sufficiently and believe that our manuscript is now even stronger with the addition of these revisions.

Sincerely,

A. Zeynep Enkavi
Stanford University
450 Serra Mall, Building 420-01
Stanford, CA 94305
215-760-9290
zenkavi@stanford.edu

Point-by-point response:

Reviewer 1:

R1.1) The authors use the term "self-regulation" as an umbrella term for several related constructs that are in themselves defined broadly (e.g. impulsivity, response control, inhibition, executive functioning). It might be useful to have a more explicit justification for the scope of this construct. Many papers in these domains have argued for narrower definitions of constructs, or even task-specific mechanisms (In addition to those already cited, e.g. Cyders & Coskunipar, 2011; Karr et al., 2018; Rey-Mermet et al. 2018; Saunders et al., 2018; Sharma et al., 2014; Stahl et al., 2014).

I should say that I don't object to a broad definition; the authors note that distinctions in the literature can be inconsistent, and are not independent of the reliability issues noted here. I also acknowledge that the data driven ontology paper covers this issue in more detail. However, I think it would be useful context to discuss here how the reliable behavioral factors derived from the data relate to the constructs in the literature that would be expected (or not) to be related to (e.g.) drug use. For example, simple reaction time is one of the more reliable measures, which may improve the reliability of a factor that includes it.

However, studies on the relationship between impulsivity/inhibition and drug use would not typically include a simple RT task.

> We thank the reviewer for underlining these two issues. We agree that a paper aiming to become a go-to reference for self-regulation measure selection should at the very least provide a good list of references for the various related definitions in the literature. Therefore we included the suggested references at various points in the Introduction and Discussion sections. We explain our lack of strong commitment to any theoretical framework in the new first paragraph of the Discussion section. We also highlight the motivation for our broad selection of measures in the penultimate paragraph of section "Implications of low reliability for behavioral task DVs," which relates to the second comment on creating latent variables of reliable measures to relate to real-world behaviors of interest.

R1.2) Of the papers noted above, I think the review by Sharma and colleagues is particularly worth mentioning. They also examine the relationship between impulsive daily-life behavior and the different measures. Whereas the current manuscript suggests that self-report measures predict real world behavior and behavioral measures do not, Sharma et al. suggest that both do (but that self-report and behavioral measures do not correlate with each other).

> We thank the author for highlighting this reference as it has also been fundamental to our thinking in this paper. While the predictive validity of the measures is not a central aspect of our manuscript and is detailed in more depth in Eisenberg et al. (2018), we note some crucial methodological differences that may explain this apparent discrepancy between our results. First, it is important to note that the values in Table 7 of the Sharma et al. paper (where the prediction results are presented) are in-sample weighted correlations. For task measures these range from .00 to 0.48. This would translate to R^2's of 0 to 0.23, which is the statistic that we report. More consequentially, our analyses took several steps to isolate the true predictive ability of a measure (or latent variable), which is overestimated by the in-sample relationship. Specifically we regressed out effects of age and sex from all of our target variables, as they are often the most important predictors masking the contribution of self-regulation measures of interest, and used cross-validated out-of-sample prediction instead of in-sample correlations. Moreover our sample size was 3 to 10 times larger than the typical studies included in their meta-analysis of predictive validity.

R1.3) I think the incorporation of diffusion model parameters adds value, though there are some considerations about the suitability of the standard DDM for certain tasks. Response inhibition tasks (e.g. the Flanker component of the ANT, Simon) typically produce data patterns that are not captured well by the DDM (e.g. fast errors in only incongruent trials, negative going delta plots in the Simon task). There have been extensions to the DDM created to account for these patterns (e.g. Ulrich et al., 2015; White et al., 2011; Hübner et al., 2010).

The DDM may be sufficient as a simplification for the current purposes, as the additional parameters the extended models produce makes aggregation across tasks less straightforward. The extended models are also more time-consuming to fit. However, where the authors recommend that researchers may prefer DDM measures for their interpretability, it may be worth highlighting these extended models to readers who might otherwise apply the standard model to tasks that it is not ideally suited to.

We agree with the reviewer that this is an important point to note, and in fact we are preparing a separate paper focusing on these modeling approaches specifically. Since they are not the center of this paper, however, and given the space limitation we have included this cautionary note as a footnote in the section titled "Comparison of task measure types."

R1.4) I commend the authors for making their code available, but some additional description of the HDDM fitting would also be useful in the supplementary material:

R1.4.1- Details like the number of samples (95000?), burn-in (15000?), and if any outlier removal was used (no?)

These are now included in Supporting Information under Analyses.

R1.4.2- Is the box (/line) in the far right of Figure 6 the contrast between thresholds for the tasks where it was allowed to vary between conditions (I saw a couple from glancing at the code)? It would be useful to mention this in the caption or text as it's hard to tell the color - apologies if I just missed it.

We apologize for this confusion and have fixed the now figure 5 (previously figure 6) itself and its legend to be clearer on this.

R1.4.3- My understanding of the code is that the HDDM was fit independently to the data from each session? Alternatively, you could incorporate the effects of time in to a single hierarchical model (it isn't a diffusion model analysis, but see the approach of Rouder and Haaf, 2018). I can see an argument for doing it either way.

While the reviewer is correct and astute in suggesting fitting a single hierarchical model for data from both time points we defer these analyses to our separate paper under preparation focusing on these modeling approaches specifically.

R1.4.4- I appreciate that it's difficult to convey details about how good the model fit was for so many tasks, but it might be worth discussing how/if it was evaluated.

We agree with the reviewer that a more detailed description of the model evaluation procedures (e.g. of our results from comparing predicted versus actual response time distributions for the different procedures) would help the readers' understanding of the consequences of different procedures, however due to space limitations and because it distracts from the main message of the paper we prefer to discuss these issues in much more detail in a separate paper that is in preparation.

R1.5) If the common assumption that threshold and non-decision time do not vary between conditions in tasks with intermixed trials is true, we could interpret the contrasts between those parameters generated from the EZ diffusion model to simply be noise. If that were the case, we shouldn't expect those contrasts to be reliable.

We thank the reviewer for providing a potential explanation for the low reliability of the EZ threshold and non-decision time variables for contrasts. In our data the EZ thresholds and non-decision times for contrasts are systematically different than 0, suggesting that they are at least in part capturing a systematic difference instead of only noise. Still, these contrast measures are much less reliable than non-contrast measures. Similar to the contrasts in raw measures of

> response times and accuracies, which are less reliable than the non-contrast measures that compose them, the EZ contrast measures appear to have signal but that signal has low reliability. Because the assumption fails we present this possibility in the manuscript as a footnote on page 13. The details of our analyses on this question can however be found on: https://zenkavi.github.io/SRO_Retest_Analyses/output/reports/ResponseToReviewers.nb.html

R1.6) The authors select the ICC(3,k) form of the Intraclass correlation - equivalent to SPSS' two-way mixed (or random) ICC with the 'consistency' type and 'average measure'. I think ICC(3,k) is also equivalent to Cronbach's alpha (c.f. https://www-01.ibm.com/support/docview.wss?uid=swg21478552). I wonder if this differs from what is usually reported for retest reliability contexts? I accept the authors' comment that the choice of statistic doesn't affect the conclusions, and the data are available for other forms to be calculated, but I can comment on why we thought the single-rater form was most appropriate in our "reliability paradox" paper.

In our paper, we report ICC(2,1), or the ICC2 given by the psych package in R (in SPSS, this is two-way random for the 'absolute agreement' type and 'single measure'). As I remember, Wostmann et al. report the equivalent of the ICC3 given in R, which we also report in our supplementary material (in SPSS, two-way random for the 'consistency' type and 'single measure'). My understanding is that the average-rater form uses the overall between subject variance, whereas the single-rater form divides the between subject variance by the number of sessions. ICC(3,k) would be interpreted as the reliability of (e.g.) mean RT if you were to administer two sessions and use the average of both of them, whereas ICC(3,1) is the reliability of mean RT taken from a single session. Cronbach's alpha is typically used to look at the internal consistency of items that are in practice averaged to create a scale score, so it makes sense to use the average-rater form. In contrast, in retest reliability settings, I think we are often more interested in inferring the reliability of a measure taken in a single session; most studies will not average across multiple sessions.

The consequence is that the ICCs reported here might be higher than the form that other researchers would choose, so it says something if behavioral measures still don't reach acceptable levels. My memory of the literature is that it is often not stated specifically which ICC is used, so I am glad to see the authors do so.

> We thank the reviewer for their detailed comparison of the two types of metrics. We agree with their assessment on the suitability of ICC(2,1) in our use case and replaced the results in the main text with ICC(2,1)’s instead of ICC(3,k)’s, which would also increase the comparability of our results with their paper, a fundamental citation in our paper. The results and conclusions, as anticipated, remained the same. We also replaced the supplemental figure S3 with a more extensive set of scatter plots that now include both types of ICC as well as Pearson and Spearman correlations.

R1.7) It is stated on pg. 9 that the ICC ranges from -1 to 1. Though statistical packages report negative ICCs, it is sometimes suggested that they should be treated as zero (e.g. Bartko, 1976), as a proportion of variance cannot theoretically be negative. I am unsure what the best approach is for the current purposes, but I note it for consideration. As there don't appear to be a large number of negative ICCs, I suspect it wouldn't affect the averages much.

> We thank the reviewer for their suggestion and conducted multiple checks to address this concern. There are 21 (of 446) variables in our dataset that have negative point estimates of ICC. We ran all analyses that involved these estimates both replacing these values with 0 and removing them completely. None of our results changed with either of these cleaning procedures. Therefore we kept the results reported in the main text as is but did add a footnote on page 8 for clarification. Details of this analysis can be found here: https://zenkavi.github.io/SRO_Retest_Analyses/output/reports/ResponseToReviewers.nb.html#R1.7

R1.8) I think it is a strength of the paper overall that the authors consider the importance of trial numbers, though I'm curious why they chose the number of trials they did for their tasks (e.g. if they came from particular studies or an average)? The role of trial numbers was also something we were interested in (see our Supplementary material D: https://link.springer.com/article/10.3758/s13428-017-0935-1#SupplementaryMaterial), and is discussed by Rouder & Haaf (2018).

I think the number of trials for most tasks in the new dataset is on the higher end of what is often done in large batteries of tests, though I noted from page 55 that the Simon task consisted of 25 trials per condition, which might be on the lower end (e.g. 40 per condition in Paap and Sawi; 160 congruent and 60 incongruent in Wostmann et al.).

> As the reviewer notes number of trials is an important variable that researchers can manipulate when designing experiments. In light of the reviewer questions we included additional information that strengthens the paper as a guiding reference. First we added the number of trials used for all the papers included in our literature review. These data are openly available at https://github.com/zenkavi/SRO_Retest_Analyses/blob/master/input/lit_review_figure.csv. We highlighted these additional data points when describing the literature review in Methods.
>
> We matched the literature as best as we could while constraining the total battery length to be around 10 hours. This selection process is now described in detail in the Supporting Information under Behavioral Task Descriptions.
>
> Second we extended our analyses of the effects of number of trials on the reliability of a measurement and included a lengthier discussion of these results in the main text under 'Effects of task length on stability.'

R1.9) "For example, while high retest reliability is desirable for measures that will be used in trait-like characteristic analyses, it is neither a necessary nor a sufficient condition for the responsiveness of a measure to capture change over time"

Just an observation that I think this is a nice point that often isn't considered in reliability discussions. I've also seen it raised in psychopharmacology (Tiplady, 1992).

> We thank the reviewer for both their attention to this subtle point we tried to convey and for the citation they provide for us to highlight and contextualize this point further. We have added this citation in the last paragraph of the Discussion section and believe that it would provide an additional helpful resource for interested readers.

R1.10) In case it's of interest, we recently published a paper discussing the interpretation of individual differences in contrast measures in raw behavior (reaction time costs and error costs) in the context of choice RT models like the DDM:

Hedge, C., Powell, G., Bompas, A., Vivian-Griffiths, S., & Sumner, P. (2018). Low and variable correlation between reaction time costs and accuracy costs explained by accumulation models: Meta-analysis and simulations. Psychological bulletin, 144(11), 1200-1227.
http://dx.doi.org/10.1037/bul0000164

> We thank the reviewer for bringing this paper up as it is indeed of interest. We have read it carefully and conducted some related analyses for our paper where we examine the response time and accuracy measures from speeded information processing tasks. Due to length limitations, however, we omit these analyses from this paper and elaborate on them on the aforementioned paper under preparation.

R1.11) The authors refer to Figure 6 on page 79 - Should it be Figure S5?

> We thank the reviewer for their attention to detail and have corrected our the figure reference in the Supplementary Information.

Reviewer 2:

R2.1• The researchers were careful in their sampling procedure to try to minimize differences between participants who completed (vs. did not complete) the retest phase of the test-retest. Some simple statistical comparisons between the completers (N = 157) and non-completers/non-responders (N = 242-157 = 85) would help characterize this subset of the sample better and provide an additional data quality check.

> We thank the reviewer for this suggestion. We compared the time 1 data of participants who have been invited to complete the retest battery but have not completed to those who have. Of the 446 variables none of them were significantly different between groups after FDR correction for multiple comparisons. This information is now also added under Methods > Sample.

R2.2• It would be nice to move all or part of the example of the task length analyses (pp. 15) to the main text. It is somewhat surprising to set up that analysis (i.e., examining how reliability changes as a function of the number of trials used to estimate reliability) and then omit discussion of a result in the main text.

> We agree that elaborating further on these analyses in the main text would strengthen the paper as a guiding reference that it aims to be. To address this we have moved these findings along with additional analyses in response to Reviewer 1's questions into the main text under 'Effects of task length on stability.'

R2.3• The paragraph on pp. 17 that begins, "Although individual difference measures from tasks are not appropriate for individual difference analyses..." is a bit out of place because it veers into a different topic and summarizes an analytic procedure detailed in another paper. I was expecting a presentation and comparison of the reliabilities of raw and DDM-derived measures (paralleling the comparison of contrast and non-contrast measures in the previous paragraph). Indeed, the differential reliability of DDM parameters and raw measures is mentioned in the discussion (pp. 19), but not in the main results section.

We agree with the reviewer that this paragraph was out of place and a description of the comparison of raw and DDM parameter reliability was missing from this section. We moved the paragraph on the reliability of latent variables to the end of the Results section with a new subtitle and added a sentence on the raw vs. parameter reliability preceding the contrast vs. non-contrast comparison.

R2.4• The order of the paragraphs in the discussion is odd - the authors lead off with a discussion of what seemed to be an ancillary finding of the negative relationship between sample size and reliability estimate. It seems the headline here is the section on pp. 19 about systematic differences in the reliability of self-regulation measures.

We agree with the reviewer that the most important message of the paper is in the subsection of the Discussion titled "Systematic Differences in the reliability self-regulation measures," however we organized the Discussion section to follow the organization of the Results section. To clarify this mapping and avoiding confusions on the initial two paragraphs being mistaken for a broader summary or the article we added a new first paragraph to the Discussion section and a subtitle for the next two paragraphs.

R2.5• It felt odd that psychological theories of self-regulation were not brought to bear on the discussion of the results. Was the fact that this study was done in the domain of self-regulation as opposed to, say, working memory, have any relevance to the results? Possibly relevant here is the paper by Fujita (2011) on how self-regulation is more than just effortful inhibition of impulses. It is possible that existing task-based measures (even the broad variety sampled here) are inherently limited in the processes they capture (potentially limiting between-subjects variance), whereas survey measures get even more breadth in terms of goals, planning, motivation, beliefs, etc., that might increase between-subjects variance.

We thank the reviewer for urging us to relate our results to a broader set of findings in the psychological literature. We address this in two ways: first by providing an explanation of why we choose not to be tied to a specific theoretical framework of self-regulation in the new first paragraph of the Discussion section and the advantages of this broad approach in the second paragraph under the subheading "Implications of low reliability for behavioral task measures". We also underscore that the two questions the reviewer brings up are important questions for future study in the last sentence of the first paragraph under the subheading "Systematic differences in the reliability of self-regulation measures".

R2.6• In the first results paragraph (pp. 7), the sentence, "To our knowledge, this is the first documentation of such a bias in the literature..." is ambiguous. Are the authors referring to the small but reliable effect on sample size on reliability, or (probably) the focal difference between task and survey data? Grammatically, the "this" in that sentence refers to the former, but I believe the authors mean the latter.

We apologize for this confusion and thank the reviewer for their careful read. We had implied the former (the negative relationship between sample size and reliability in the literature) and have now clarified this to avoid further confusion.

R2.7• Point of clarification on the last sentence in the Data Quality Checks section of the results (pp. 9): the participants were recruited using mTurk, but the data were collected using Experiment Factory, correct?

> The reviewer is correct in distinguishing between the participant recruitment and data collection platforms. We have clarified this in our Methods section to avoid further confusion.