Response Letter (Stage 2, Round 1)

Title: Registered Report: Self-Control Beyond Inhibition. German Translation and Quality Assessment of the Self-Control Strategy Scale (SCSS).

Dear Prof. Dr. Dienes (editorial handling), Dr. Miles (review 1), Dr. Werner (review 2) and Dr. Bürgler (review 3),

thank you very much for providing your time and expertise to this project. With great interest, we have studied your feedback and comments on the current state of the Stage 2 manuscript. We are eager to further strengthen the readability and conclusiveness of the draft and are confident that the implementation of your recommendations improved the manuscript considerably.

As in prior rounds, we address each comment in chronological order below (apart from the comments on data exclusion by Dr. Werner, which we pooled). We hope to have accommodated all concerns appropriately and are looking forward to further feedback.

Best regards

Leopold Roth (corresponding author)

Comments by Prof. Dr. Dienes

Dear Prof. Dr. Dienes,

thank you very much for your continued commitment to leading our project through this process. We are very happy to receive this amount of positive feedback on our project and hope to further strengthen our manuscript by including all addressed points by the reviewers.

Detailed comments:

The reviewers are in general happy with the Stage 2, it is satisfying to see such a well thought out piece of work come to fruition, but the reviewers do have important comments to address. As an editor, I would like to especially highlight a point made by reviewer 1: The results are complex, and determining whether you had strictly followed the Design Table is not easy. Could you lay out your results according to each row of the Design Table, in the same order, and in that way make clear you have addressed each question in the precise way you said you would, reporting no more outcome measures (in the main results section) as you listed, and drew your conclusions according to the precise method you had set out (e.g. spell out the AIC comparison of models). That way anyone can scan the Design Table and read its corresponding implementation in the Results section in a convenient way.

Response: Thank you for advising us on the presentation of the results. We think that the manuscript benefitted from this input. We have restructured the results and discussion section accordingly and addressed the individual aims and respective hypotheses more clearly. We repeated the precise criteria wherever justified to lend further transparency to each test. We hope that these actions increase the comprehensiveness of our manuscript and will make it easier to read and communicate the findings. We are happy to integrate any further feedback on this matter to further strengthen the clarity and conciseness.

Comments by Dr. Miles

Dear Dr. Miles,

thank you very much for your very helpful comments and observations, regarding our manuscript. We hope to sufficiently address your comments below as well as in the updated draft.

Detailed comments:

To assess this criterion, I would find it helpful for the authors to provide additional information to explain why so much data was excluded. If I am reading Table 3 correctly, a very large proportion of responses were excluded in Studies 4 and 5 (58% and 53%). I visited the link on p14 to find the number of exclusions by reason and sample (https://osf.io/aup93/), but could not find this information. Could the authors point me to where I can find it on OSF, or add it if it has not yet been uploaded?

The pre-registered exclusion criteria were clear and were quite stringent regarding responding incorrectly to attention checks, so it does not necessarily seem surprising to me that a lot of data was excluded. However that does seem like a very large amount of unused data, and I am curious about what happened here and what the authors' interpretations were. For example, I would be interested to hear whether the authors feel confident that all of the excluded data is poor quality, and whether they think there is any chance that the exclusions systematically relate to questionnaire scores (e.g. are people with lower inhibition more likely to miss an attention check?)

Personally I think it would be useful to provide more detail on this in the manuscript as well as in the response letter.

Response: Thank you very much for pointing us toward this source of confusion. As correctly noted by you, we forgot to create the pre-registered table to document the exclusions by reason.

We now added this table to the supplemental material and added an in-text note that the majority of exclusions were due to incomplete participation. The remaining exclusion criteria only accounted for relatively small fractions of the exclusions (14 % across all studies). This trend is specifically evident in the social media samples, which, from our perspective, isn't surprising, as the initial sample size also included everyone who e.g. accidentally clicked the link to the survey. The majority of the incomplete cases (57% in study 3 and 62% in study 4) completed at most the first 7% of the study meaning they only got to the introduction of the SCSS.

To further clarify the procedure, we added a line in Tables 2 and 3, documenting the number of complete cases before the application of the remaining exclusion criteria, which is then documented in the line below (effective N).

We hope these measures clarify the exclusion process and draw a more conclusive picture of the data acquisition.

I did not note any problems with this criterion. However, one thing I'd note is that the hypotheses in the introduction were very clearly stated, but the results section does not refer specifically to the hypotheses, and the subheadings also don't use wording that is consistent with the introduction. A clearer correspondence between these two sections would make it easier for readers to follow along as the aims are tested and to understand whether all the hypotheses were supported by the data.

My suggestion would be to use the three aims stated in the Design Table as the subheadings (i.e. internal structure and reliability, convergent and discriminant validity, relationship to

self-control outcomes) and underneath these headings to refer specifically to the numbered hypotheses.

Response: Thank you for pointing this out. This comment is very consistent with the editorial recommendation by Prof. Dr. Dienes. We addressed these points in our reply above. We hope to have increased the clarity and comprehensiveness of the manuscript by restructuring our results and discussing the defined aims and hypotheses. We are happy for further feedback to make the findings as clear as possible.

Related to this, I was not sure it worked well to start the results section with a section titled 'Study 1 (Pilot)'. Study 1 is really quite different from the other studies (i.e. methodological development rather than hypothesis testing), and this section doesn't test any hypotheses (or really present any results), it is more like preliminary work that allows for the other hypotheses to be tested. Perhaps a different/more descriptive subheading would work better to make this clear, or even to move this description of scale development to the method (where the SCSS measure is described).

Response: Thank you for sharing this impression with us! We changed the title of the section to be more descriptive of its purpose, by renaming it ("Pilot: User Experience and Design Adaptation of Questionnaire") and referring to Study 1 as Pilot now throughout the manuscript. Consequently, we renamed Studies 2-5 to Studies 1-4 to highlight the different purposes of the pilot study and the main studies.

I generally found these informative and clearly explained. However, it seems to be assumed that the reader will already know what a lasso-regularized network model is and how to interpret it. I did not know these things and would have appreciated more information!

Response: Thank you very much for pointing out this lack of clarity! This was very valuable feedback to us. We added some notes to it, which hopefully explain what this method does and why its application is useful. Further, we added a reference to this section.

I thought the conclusions were appropriate. The descriptions of findings seemed accurate and the authors' interpretations seemed reasonable, and the authors were also careful to make relevant caveats clear, e.g. that data could often be interpreted in multiple ways and that interventions are not necessarily justified based on this data.

In the discussion section, again I thought using wording consistent with the earlier parts of the paper would be helpful, i.e. starting by using the three key aims as subheadings (internal structure and reliability, convergent and discriminant validity, relationship to self-control outcomes) before moving on to other topics which are more exploratory or speculative.

Response: Please refer to our response on the connection of the design table, hypotheses, results, and discussion above. We hope to have addressed all critical points sufficiently and are looking forward to receiving further feedback on this matter.

Comments by Dr. Werner

Dear Dr. Werner,

thank you very much for providing us with your feedback and expertise on the current manuscript. All of these points were very helpful and we addressed them, hopefully sufficiently, below. We pooled your comments on data quality to avoid repetition. We also answered a similar comment by Dr. Miles above.

Detailed comments:

Comments on data quality:

Comment 1: *In table 3, it looks like there was a lot of data lost to quality checks. In some studies, the retention rate is as low as 42-53%. This amount of data loss makes me very hesitant about the quality of the data overall (e.g., was there something wrong with the sampling methods? Maybe the study was too long, considering the number of variables assessed?). At the very least, this needs to be explicitly addressed in the discussion and any conclusions should be tempered as a result. I might also suggest doing some investigation as to why this issue with data quality exists (e.g., there appears to be a lot of attention checks, are most people failing the same checks within a single scale? Are people failing one check or many?).*

Comment 2: *With the above point in mind, I went back and saw that the samples with the lowest data quality seem to be those collected through social media. While I think that makes a little bit more intuitive sense, I think finding some reference regarding social media data quality is imperative here (e.g., is this amount of data loss similar to what other studies experience?).*

Comment 3: *Finally, considering the issues with data quality, I would also recommend (a) re-running any pooled sample analyses without those samples, and (b) re-run any quality assessments (e.g., such as those listed in Table 6) for individual samples. This would serve as a bit of a sanity check if the findings remain the same without these seemingly problematic samples.*

Response: Thank you for pointing our attention to this issue. Dr. Miles addressed this as well above and we created the forgotten table which documents the exclusions by sample and reason in the supplemental material. As evident from there, the majority of exclusions are not based on the attention checks but on incomplete participation. From our perspective, this appears intuitive in social media samples, as the initial sample size is inflated by participants who accidentally/out of curiosity clicked on the link during social media usage. The majority

of the incomplete cases (57% in study 3 and 62% in study 4) completed at most the first 7% of the study meaning they only got to the introduction of the SCSS. After excluding incomplete participations, as pre-registered, the remaining exclusions, based on the other criteria do not appear overly large (about 11 % across all studies). To further clarify this to readers, we added a line in Table 2 and Table 3, documenting the number of complete cases, before applying the remaining exclusion criteria. We hope this clarification is a sufficient argument, as to why we are currently not concerned that the data quality is below the standards in the psychological literature, but rather substantially strong, given the rigor of attention assessment.

Yet, we followed the recommendation to test whether one sample was specifically influential on the model fit and computed the hypothesized 8-factor model with the pooled data recursively while excluding each of the larger samples once from the analysis:

Response Table 1

Model fit without samples 3,4 and 5

excluded sample	CFI	TLI	RMSEA	SRMR
none	.927	.920	.044	.059
2	.926	.918	.044	.062
3	.925	.917	.044	.058
4	.929	.921	.043	.059

To our interpretation, the added analysis doesn't indicate that a certain sample is prone to bias the findings meaningfully. We hope the clarification of the transition of sample sizes from collected through completed to effective is sufficiently documented and that the additional factor models lend further trust to our data quality.

Comments on remainder of manuscript:

For table 3, it would also be helpful to have a row indicating the type of sample (e.g., prolific, student, etc.) used in each study. This way this information is easily accessible when comparing metrics across samples.

Response: Thank you for your feedback on the readability of our manuscript. This is very valuable and we equipped the table (now Table 2) with more extensive information

Regarding the lack of an association between strategies and habits. I'm not as deeply entrenched in the habits literature, but this really interesting paper on how strategies and habits can be synergistic may be worth considering for the discussion (see citation below). Generally speaking, the question of whether strategies predict habits is interesting, though one that is not super obvious to me (and certainly not for construct validation) and certainly one that does not have strong empirical support, to my knowledge. However, this recent paper talks about how habits and strategies are often treated as different processes, but that strategies can be used to support more complex habits. Perhaps this newer framing regarding the interplay between strategies and habits might be useful for the authors' discussion.

- *Saunders, B., & More, K. R. (2024). Some habits are more work than others: Deliberate self‐regulation strategy use increases with behavioral complexity, even for established habits. Journal of Personality.*

Response: Thank you for providing us with this very interesting source. We integrated it into our discussion section.

Given that the field is shifting its emphasis from what strategies are "good" vs. "bad", I am rather cautious about some of the conclusions using this language in the discussion. The SCSS is designed to capture general/habitual strategy use, but it is most likely the case that some strategies are good in certain situations and may backfire in others - while we can start

to capture this at the habitual level when assessing moderators (which is currently beyond the scope of the current study), I suspect these distinctions are more likely to emerge when capturing actual strategy use (e.g., when actively managing a self-control episode in-the-moment). Perhaps some discussion around this would help give greater context to the current findings (see one of the original conceptual pieces by Bonanno & Burton (2013), some discussion of this issue focusing specifically on reappraisal by Ford & Troy (2019), or a recently adapted version by Werner & Ford (2023) that bridges some of those concepts to self-control).

Response: Thank you very much for lending us your expertise on this aspect. We changed the discussion of the strategies, specifically punishment to a certain degree. Given the clinical evidence that punishment cognitions and behaviors can be risk factors for self-harm behavior, we refrained from stating that this strategy has the potential to be helpful as the potential risks outweigh the benefits from our perspective.

We strengthened the notion that no strategy is likely to be beneficial in all circumstances and referenced the suggested literature (Bonano & Burton, 2013; Werner & Ford, 2023).

For the descriptive tables (e.g., Tables 3-4), including percentages rather than raw number distributions would be a little bit easier to read.

Response: Thank you for your feedback on the presentation of the descriptive results. We adjusted the information accordingly, which is now presented in percentages and hence hopefully easier to compare.

I think some of the tables are numbered incorrectly as there are two table 4s.

Response: Thank you very much for your close readership of the manuscript! We corrected this mistake and made the respective changes throughout the manuscript.

For the limitations and future directions section, I would recommend including citations where appropriate. While I agree with most of what was said, the recommendations are fairly consistent with what other more established work has stated as well.

Response: Thank you very much. We hope the current state of the discussion appropriately accommodates this comment. Yet, we are looking forward to further feedback in case of potential gaps from our side.

Comments by Dr. Bürgler

Dear Dr. Bürgler,

thank you very much for contributing your time and expertise to the second stage of this registered report. We are very thankful for your comments and hope to have addressed each of them sufficiently.

Detailed comments:

I do have some concerns with Behavioral Inhibition (BI) and that the discussion now portrays this as a highly effective strategy (e.g., on p. 36 "behavioral inhibition still related to the highest number of outcomes […] This is good news because previous research put a strong emphasis on behavioral inhibition. Our results show that this focus is not unwarranted."). While this is backed by the results, I do interpret them somewhat differently. I do have ever increasing doubts that one can actually call BI a strategy (also considering the items used in the SCSS) and think it should better be understood as an outcome (see also Werner, Inzlicht, et al., 2022). For example, a person might respond to the item "I find it easy to keep myself from acting on unwanted desires" with high levels of agreement (indicating high levels of BI) not only because they are better at resisting desires through sheer willpower, but because they might have successfully used any of the strategies before this point that might have downregulated the level of desire they feel in the moment. Therefore, to

*some extent, this "strategy" might show to be highly effective because successfully using any of the preceding strategies is assessed through this "strategy" as well. BI would then not only be conflated with the use of previous strategies but with the at least partially successful use of them (perhaps because they were implemented context-sensitively). Additionally, some people might in general perceive certain desires to be less tempting, and such individual differences might play into this as well. The issue that BI might measure an outcome and not a process is furthermore supported by the finding that BI was related to the BSCS strongly (with .74*** and well above any other strategy, with the second highest only showing a correlation of .35*** with the BSCS). This is noteworthy because for the BSCS, similar concerns have been raised, i.e., that it might measure an outcome (i.e., successful self-control) rather than the underlying process of how this outcome was achieved (see Bürgler et a al., 2022). I furthermore found it fascinating that BI was significantly positively associated with habit strength across all behaviors and timepoints. Quite often it was the only, or one of few, strategies that showed significant associations with habit strength and in all cases it showed the strongest. Habitual behavior is enacted automatically, with minimal cognitive effort, awareness, and control (Gardner & Rebar, 2019). Therefore, theory would strongly suggest that BI, if it actually measures the degree to which the "strategy" effortful behavioral inhibition is used, should be negatively associated with habit strength. If this is not the case, it might be a sign that the subscale BI measures something it is not supposed to measure (e.g., an outcome rather than a process). Subsequently, it might be appropriate to temper some of the wording when discussing the efficacy of this strategy and to discuss some of these potential issues and open questions.*

Response: Thank you very much for this very interesting point. After longer engagement with measures of self-control, we are currently at a similar stage of the discussion and wonder, whether the measures of behavioral inhibition couldn't be conflated by assessing self-control

success. We address your comment in two ways. First, we modified the discussion to a) include your and our thoughts on behavioral inhibition as an outcome and b) recommend keeping this in mind when assessing behavioral inhibition measures with the given or similar items. Second, we computed all regression models from Table 11 without behavioral inhibition and presented these in the supplemental material. Considering these secondary models, we see stronger associations of other strategies and outcomes but many associations also remain non-significant and small.

Regarding strategy repertoire, the authors wrote "This indicates that it might be useful to have a broad repertoire of strategies" (p. 37). Would it not be possible to provide these analyses with the current data, similar to Werner, Wu et al. (2022) as an additional exploratory analysis (perhaps reported in the SOM)? It appears to me that calculating a repertoire size (e.g., using a sum index, see Werner, Wu et al., 2022) might be a good way to utilize data gathered from the SCSS in general. This might be especially relevant when considering the following point.

Response: Thank you for raising this interesting point. We discussed the computation of repertoire scores and decided that the already very long manuscript might not benefit in readability when including this further topic as an exploratory post-hoc analysis. Especially as this touches a field, reaching beyond liberal definitions of scale validation and in the light of the editor's comments that the result and discussion section should rather increase their clarity with regards to the pre-registered analysis. Yet, we can imagine well to do this analysis as secondary data analysis in a follow-up project but also offer our documented open data and code to anyone, who is interested in doing this. We hope these arguments are sufficient to justify, why we refrained from including this analysis. Rather, we modified the referenced section and more directly pointed toward the respective literature on repertoires.

Some of the strategies are described as "maladaptive" (e.g., on p. 37). Here, I find it important to discuss that describing some strategies as maladaptive is somewhat dangerous (see the "fallacy of uniform efficacy" in emotion regulation research; Bonanno & Burton, 2013). It is likely that strategy efficacy depends on many factors, for example, the context and the person using the strategy (see Hennecke & Bürgler, 2020). Therefore, strategies should be used flexibly and context-sensitively, and strategies that might be effective in some situations might not work in others (e.g., Wenzel et al., 2023). The SCSS assesses strategy use very broadly across situations (this issue was briefly discussed in the limitations on p. 41), which is why such nuances are lost. I think it is valuable to provide insight into what strategies are generally effective, but it might be important to clearly describe it as such a "general efficacy". Subsequently, it might be important to discuss the possibility that this might not necessarily mean that they are maladaptive in every situation and for every person.
Response: Thank you for raising this interesting point. We extended the paragraph by mentioning that it can not be guaranteed that a certain strategy can not be productive for any individual or situation.

Given the clinical evidence of self-punishment behavior and cognitions as a risk factor for self-harm and non-suicidal self-harm behaviors, we refrain from stating that it can be helpful but hope that the extension of our prior statements is sufficiently clear on the absence of absolute generalizability. Following your and Dr. Werners' recommendation, we included the respective work by Bonanno & Burton (2013).

Regarding the additions made from the previous version, there was one small addition for which I would have liked a short explanation (perhaps in a footnote), which was the addition regarding the PHQ-9 "We only assessed eight of the items, not including the measure for suicidal thoughts and tendencies." (p. 19).

Response: Thank you for this suggestion. We added a footnote which hopefully delivers further insight into our decision to exclude item 9 from the assessment.

When introducing the different strategies on p. 9, it might be confusing to some readers to see "changing environments" as part of situation selection strategies and not situation modification, as they are commonly descried as two separate groups of strategies that can be subsumed under "situational strategies" (e.g., Duckworth et al., 2016, p. 40). Similarly, on p.12 and p.13, two strategies are described as "situation selection", which I would clearly describe as situation modification, i.e., "getting rid of one`s dryer" and "turning off the wifi automatically to go to bed earlier". Therefore, a small addition in the introduction of the strategies might be needed to explain that you consider both situation modification and situation selection strategies under the term "situation selection", similar to what Katzir et al. (2021) did, by treating both situation modification (or "stimulus control") and situation selection strategies as one type of strategy, because they "load onto the same factor and were therefore combined into one final subscale (situation selection)" (p. 5).

Response: Thank you for this very helpful comment which we hope will serve even more clarity to the conceptual framework. As suggested, we added a notion to this in the introduction and referenced Katzir et al. (2021) accordingly.

Regarding the thresholds used for the R2 and adj. R2 (e.g., "R2 < .26" on p. 23). While the same thresholds were defined in Stage 1, I have now realized that there appears to be no reference to justify those exact thresholds.

Response: Thank you for highlighting this point, lending further clarity to our criteria. We referenced now the respective place in the manuscript to Cohen (1988), whose interpretation guidelines we also applied at other points in the manuscript (see note below Design Matrix). We hope this added information sharpens the interpretability of the draft.