Response Letter (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), Dr. Bürgler (review 3),

we thank you for your supportive, constructive and refreshing review of our first submission of our Registered Report on the German translation of the Self-Control Strategies Scale (SCSS). After carefully studying your reviews, we have updated our submission and will give a comment-by-comment answer to all raised points. These will be presented in the order in which we received the reviews.

All points raised were extremely helpful to improve the manuscript and we are very glad to be able to include your expertise prior to the conduction of the project through this format. We are looking forward to the next round of reviews and hope to have incorporated your ideas appropriately.

Best regards,

Leopold Roth (Corresponding author)

Comments by Prof. Dr. Dienes

Dear Prof. Dr. Dienes,

thank you very much for your editorial work on our Registered Report. Below, we will address the points you and the reviewers mentioned in detail in four sections. Besides, we made some general changes to the study design based on multiple comments, which we would like to outline here first.

Based on multiple comments, we reconsidered part of our data collection strategy. Our updated sampling plan can be found in Table 2 and affects parts of the Design Matrix, method and results. Two major changes have been made: (1) As suggested by Dr. Bürgler, we added a pilot study to get feedback on the understandability of the items and an abbreviated version of the introduction to the SCSS, which Dr. Bürgler suggested to dampen the focus on temptation in the scale. (2) We added two more measurement time points to Study 2 (the student sample). Data from the first time point will be used to test the proposed factor structure and allow for necessary changes before running the remaining studies, as was suggested by Dr.

Miles. The remaining three time points will be used to assess test-retest reliability over a longer period of time. In exchange, we refrain from assessing test-retest reliability in the other samples, as three measurement occasions increase the precision of the estimate considerably. Additionally, the three time points will be used to include a causal analysis by predicting changes in outcomes over time based on Dr. Bürgler's suggestions.

We are convinced that these changes as well as the other changes based on your comments led to a much improved manuscript.

Detailed Comments

"One key issue to address, emphasized by Miles, is to make sure there is an exact mapping between theories emphasized in the introduction and the inferential chain leading to conclusions, as laid out in the Design Table. In fact, all three reviewers in some way made suggestions about the framing of the study."

Response: We think this is an important and valuable point. We have restructured both the introduction and (to a lesser extent) the Design Table to achieve a better mapping between the two. Further details on the changes are outlined in the comment-by-comment response to the respective comments.

"The third reviewer wonders about some of the cut offs; for myself, I don't do this sort of work, but I wondered if you could contextualize the use of .5 and .1 for R sq for divergent and convergent validity - is there any reason why these should be taken as in general meaningful, regardless of the raw units and nature of the variables? Also could you spell out exactly how you will use AIC"

Thank you very much for pointing us towards this ambiguity. We decided to change the levels according to the commonly used interpretation standards by Cohen (1988). These interpret $R^2 < .13$ as weak, $.13 < R^2 < .26$ as moderate and $R^2 > .26$ as substantial. The manuscript was adapted to match these guidelines.

The Akaike Information Criterion (AIC, Akaike, 1974) will be used to distinguish the three different factor models by their appropriateness, given the data. The interpretation of the AIC includes several noteworthy points: a) Its absolute value is not interpretable, only the difference of values per model is meaningful; b) the lowest value indicates the preferred model; c) differences between the smallest value and the value of the given model will be interpreted, based on the recommendations by Cavanaugh and Neath (2019) (interpretation of AIC of a model based on the AIC difference to the best fitting model: < 2 substantial support, 4-7 considerable less support, > 10 essentially no support). To clarify this usage, we included a footnote in the manuscript.

Comments Dr. Miles

Dear Dr. Miles,

thank you very much for your expertise and detailed feedback on our manuscript. Below, we will respond to each point you raised in the order of the review. We believe that the manuscript has considerably improved due to your suggestions.

Detailed Comments

"a) Expand the introduction so that theory and hypotheses are presented for Aims 1 and 2. The work planned under Aims 1 and 2 extends the original SCSS paper in multiple ways, e.g. replicating the original study in a different cultural sample and extending the validation to include multiple different self-report measures related to self-control. However, the introduction focuses primarily on Aim 3 (i.e. measures of behaviour). This means that some significant contributions of the paper are not mentioned, and also means that it is difficult to evaluate some of the analyses in the Design Table because a rationale has not been given."

Response: Thank you for this very important feedback to improve the clarity of the manuscript! We have restructured the introduction so that each aim is addressed in a dedicated section ("Eight Strategies of Self-Control" for Aim 1, "Construct Validity of the SCSS" for Aim 2 and "The SCSS and Relevant Self-Control Outcomes" for Aim 3). We have also reshaped the aims slightly by incorporating the tests of the factor structure and reliability in Aim 1. Thus, Aim 2 now focuses clearly on testing the convergent and predictive validity.

"b) Expand the Design Table so that sampling plan, tests and interpretations are presented for Aim 3. In the Design Table, measures of behaviour are included within the row 'Predictive validity', with the hypothesis that "All subscales combined explain a relevant amount of variance in self-control related measures". This frames these measures as part of Aim 2, as a test of whether the SCSS is a valid measure of self-control with the ability to predict relevant behaviours.

However, the introduction and the abstract describe the research questions related to measures of behaviour differently. The introduction explains "we aim to shed light on the relevance of the different strategies of the SCSS in these domains". The aim here seems to be to explore whether different behaviours are predicted by different strategies. The introduction presents this as the primary contribution of the planned study, and separates it into a different aim (Aim 3). While I agree that this is a very valuable and interesting question, it seems to be a different question than the one answered by the planned analysis in the Design Table."

Response: Thank you for pointing out this ambiguity. We have rephrased the part on Aim 3 in the design table to match what is stated in the introduction. Now, we clearly separate between Aim 2 and Aim 3 - Aim 2 assesses convergent and discriminant validity (focusing on measures of self-control and related constructs), Aim 3 does not assess validity, but the relevance of the different strategies in different domains (focusing on the different outcomes measures). Within the design table, we have also added subheadings to highlight which aim each test relates to.

“I experienced a similar problem with the other included questionnaire measures. In the Design Table, most of these are discussed under 'Discriminant validity' (but the corresponding introduction section 'Self-Control Related Cognitions and Personality' does not discuss these measures as tests of discriminant validity; the research question is framed more in terms of investigating individual differences). The measures related to mental health are instead discussed under 'Predictive validity', but the rationale under "Self-Control Impeding Conditions" seems more about investigating which strategies are responsible for previously observed associations with self-control, although a clear research question is not stated.“

Response: Thank you for pointing out this mismatch. We have rewritten the section explaining the self-control related measures to more clearly address our aim to assess discriminant validity (in the new section “Construct Validity of the SCSS”). Besides, we have removed the part on “Self-Control Impeding Conditions” to keep the introduction more parsimonious because as you mentioned there is no clear research question connected to it. We still plan to include those measures, but in a more exploratory fashion.

“Column "Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis": I do not find statements such as "beyond all common sample size requirements for retest reliability" very helpful, as they are difficult to evaluate. I would suggest giving more detail. Can details also be provided on how the effect sizes were determined in cases where power analyses are included? The linked OSF pages show the calculations, but do not explain how the parameters were chosen.”

Response: Thank you very much for pointing us towards this lack of precision in our submission. We adjusted the design table and manuscript accordingly. For the confirmatory factor analysis (CFA), we base the argument on Kyriazos (2018) who reviewed sample size recommendations for CFA from different sources. The highest recommended sample for CFA in case of non-normal data is at $N = 1{,}000$ which we hope to oversample to $N = 1{,}800$. The recommendation is based on a set of Monte Carlo Simulations and should reach beyond classic rules of thumb. For test-retest reliability, we base our approach on Bujang and Baharum (2017). To obtain a ICC with a power of 90% and three measures per person, the highest recommended sample size was N = 83, which we hope to oversample to $N = 153$ (repeated measures).

Regarding the determination of the effect sizes, we discussed internally for a prolonged time before submission and should have described it more transparently. Following the suggestions by Lakens (2022), we defined $R^2 = .10$ as minimum effect size of interest, which we plan to detect with power > 97%. As stated in the cited review, picking the smallest effect size of interest is not a process, where all researchers necessarily arrive at the same conclusion. For us, effects below 10% explained variance are not practically meaningful enough to justify the added resources of keeping the high power constant (which still remains high when considering even smaller effects, like .07). Balancing impact and resources led us to this conclusion, which we now describe in the supplemental material to increase transparency. We have added this additional information to the supplement.

“Row "Convergent validity": The Stroop task measures something so different from the SCSS that I am not sure it is logical to use it as a test of convergent validity. As the authors point out, recent work suggests that a lack of relationship might be the most likely outcome, and I agree with the authors that this would most likely indicate "that Stroop performance is not a valid measure of self-control". Thus it doesn't seem quite right to frame this task as a test of convergent validity.”

Response: After the comments on the Stroop Task by you and Dr. Werner, we have decided to drop the Stroop task from the study. Thank you very much for making us reconsider the procedure accordingly.

“Column "Interpretation given different outcomes": various criteria here are underspecified. For example, the criteria contain the terms 'enough', 'a few', 'too much', and 'a relevant amount', when it is unclear exactly how much variance will be considered 'too much' (etc).”

Response: Thank you very much for pointing this out. We have added specific numeric thresholds for each analysis.

“d) Given that multiple samples are planned, have the authors considered testing Aims 1 and 2 first before testing Aim 3? In other words, performing factor analysis and validation of the scale in the initial samples, before going on to adapt the scale if necessary for use in predicting a broader range of behaviours in subsequent samples?”

Response: Thank you for this helpful idea! We have added another time point to Study 2 to achieve this. This time point will take place before remaining data collection. We will test the factor structure and reliability of the subscales in this sample and adapt the scale if necessary before collecting further data.

Comments Dr. Werner

Dear Dr. Werner,

thank you very much for your expertise and detailed feedback on our manuscript. Below, we will respond to each point you raised in the order of the review. We believe that the manuscript has considerably improved due to your suggestions.

Detailed Comments

“With respect to the framing of the study, I found the introduction to be quite lengthy and could've been described in a far more succinct and compelling manner. As one notable example, I'm not sure the authors need to discuss in great detail why self-control is important for each individual domain and these several pages can be combined into 1-2 paragraphs without losing any of the key points. While I appreciate that there are likely stylistic differences motivating some of my comments, I would generally encourage the authors work on better selling why this work is important and very much needed (because it is!) without getting lost in a laundry list of constructs.”

Response: Thank you very much. This was a very helpful suggestion on improving the precision of the manuscript. We have shortened the section and combined it with the section on Katzir et al.’s (2021) results concerning the relationship of the SCSS with the outcomes. We believe this paints a much clearer picture of the main message.

“As for my second point regarding divergent and convergent validity, I'm not entirely convinced that the measures used for these purposes are really all that informative. As noted below, there is quite extensive evidence now that there is a lack of convergence between self-report and behavioural measures of self-control. Thus, a lack of an association in this case wouldn't necessarily say much about the quality of the SCSS (at least I don't think). Similar with the Brief Self-Control Scale. This scale is widely understood as being a poor measure of self-control, though everyone uses is because that's what's widely available/popular. I'm not really sure what the best solution is here, as there really aren't many good measures of self-control or a comprehensive measure of strategies to compare to. Should the authors proceed with these particular measures, I would suggest toning down the interpretation of a lack of an association as evidence against the SCSS because it could just as likely (if not more likely) be problems with these comparative measures.”

Response: Thank you for these helpful thoughts regarding those measures. After the comments from both you and Dr. Miles on the Stroop Task, we have decided to drop it from the study entirely. We still see a valuable contribution in investigating the convergent validity with the BSCS, as this was also done by Katzir et al. (2021). However, we have changed the interpretation of a lack of convergence between the BSCS and SCSS to a more general indication that they measure distinct constructs rather than an indication of bad quality of the SCSS.

“In the introduction, when explaining self-control using the exercise after work example, the authors state: "In this situation, they can exert willpower to override the immediate desire in favor of their long-term goal." This is a bit picky, but I would refrain from using the term "willpower" here, since it is such a vague term that has different meanings to different people (e.g., sometimes it's referred to as a strategy, sometimes it's used as a synonym for self-control, etc.). Instead, I think the authors can make the same general point by simply saying something like "In this situation, they need to figure out a way to override the immediate desire to hang out on the couch in favour of successfully pursuing their longer-term goal to exercise more." “

Response: We agree that the term “willpower” can add confusion and make the main point unclear. Thus, we changed the sentence in line with your suggestion.

“Then, to make the introduction more succinct, I would recommend collapsing paragraphs 2 & 3. For example, in the proceeding paragraph (paragraph 2) the authors could say something like: "Historically, research has focused on a person's ability to exert willpower to successfully resolve conflicting desires [add key citations]. However, the concept of "willpower" has recently been called into question (e.g., Fujita et al., 2020; Inzlicht & Friese, 2020; Werner et al., 2022) and instead researchers have focused on more tractable strategies people can use to achieve their long-term goals. Indeed, prior research shows that people might instead set up rewards or punishments for themselves...[more generally describe the wealth of prior research examining other strategies; e.g., Mischel et al., 1989]. As a way to measure the vast range of strategies people can use to achieve their goals, Katzir et al. (2021) developed a novel scale that assesses some of the most primary strategies used in desire regulation across different domains."”

Response: Thank you for this detailed and concrete suggestion. We have changed the paragraph accordingly.

“Under the heading "The Self-Control Strategy Scale (SCSS)", I would recommend stating the scales name again in the first sentence – i.e., "The Self-Control Strategy Scale (SCSS) was developed..." I realize this may feel redundant since you have the scale name as the header of the section, but referring to the scale only as "the scale" in the written text feels a bit too informal.”

Response: We absolutely agree and changed it.

“Within the section on the "Eight Strategies of Self-Control" the authors mention that the SCSS can be grouped into three categories, including anticipatory control, down regulation of temptation, and behavioural inhibition. I'd encourage the authors to double check the original scale development paper, but I believe that was the original hypothesized factor structure but it was not supported in the data and instead the eight strategies are technically independent. While I still think these distinctions are theoretically interesting, I would recommend that the authors here make it clear where these three categories come from and whether or not they were supported in the original model from Katzier et al. (2021) (e.g., it's a bit confusing that you mention these three categories here and then jump to saying "eight factor model" in the following section).“

Response: This is a great point, which should lead to a clearer understanding of the theoretical model. It took us a while ourselves to disentangle theory and evidence in the original publication (Katzir et al., 2021). To accomodate this suggestion, we dropped the three categories from Table 1 and added a paragraph discussing the origin of the theoretical categories (based on Duckworth et al., 2016) and the evidence for empirical categories by Katzir et al. (2021). We hope this makes clear why we assume an eight-factor structure as opposed to a three-factor structure. Still, we kept the theoretical categories in the text because, as mentioned by you as well, we see it as a helpful bridge for understanding.

“In the section "Evidence for the SCSS", the authors give a nice summary of the many strategy associations across domains. Toward the middle/end of the section, they talk about how some strategies may be adaptive or maladaptive in some domains. I just wanted to point out that this is actually consistent with recent theorizing on self-control, which maps accordingly onto similar findings in the context of emotion regulation strategies (see Werner & Ford, 2023 for one such overview). I just wanted to point this out in case it is helpful for the authors' thinking, and more generally making sense as to why these patterns may have emerged.”

Response: Thank you very much for the suggestion! This is a very interesting paper that will be useful for interpreting our results as well.

“For the section "relevant self-control outcomes", the authors do a very thorough job describing in-detail the role of self-control in different key domains. That said, given that the introduction is quite lengthy already, I wonder if there is a way to make this section more succinct. For example, I don't think it's really necessary to describe how each domain is relevant for self-control, as they are all extensively studied in the literature. One recommendation would be to simply lean into the fact that self-control is highly relevant across several important life domains, including health behaviour (e.g., give key citation), performance at school and work, interpersonal relationships, and pro-environmental behaviour. In this way, you would give key citations relevant for each domain. You could then say something like "Consistent with Katzir et al. (2021), prior research provides support for

the effectiveness of certain strategies in particular domains. For example, [give a few key examples.]" Statedly differently, I would emphasize evidence for strategies across domains (especially if there are differences like in Katzir et al.) and focus much less on the general role/importance of self-control in each context. Over all, my point here is not to change the message or remove any of the base content per se, but rather help make this section more succinct (e.g., I envision all of this being maybe 1-2 paragraphs at most).”

Response: Thank you very much. This was a very helpful suggestion on improving the precision of the manuscript. We have shortened the section and combined it with the section on Katzir et al.’s (2021) results concerning the relationship of the SCSS with the outcomes. We believe this paints a much clearer picture of the main message.

“Related to the above point, for a more recent example of strategies in the context of school performance, I'd recommend checking out Duckworth, White, et al. (2016) where they empirically test the benefits of situational strategies.”

Response: Thank you very much for this useful suggestion. We have added it to the manuscript.

“For the section on metacognition about self-control, I would refrain from using the term you are defining in the definition itself (e.g., you use the term "metacognitive" twice in the definition). I would rephrase in a way that is more descriptive.”

Response: We rephrased the definition.

“Relatedly, the authors state "in order to use a broad set of self-control strategies, it is vital to know about them and how to use them." While I agree with this statement, it's not clear to me that this meta-cognition scale exactly does this. Specifically, this scale does not assess one's knowledge about specific strategies (if anything, the SCSS kind of assesses this information), but rather if people believe they generally use several strategies (i.e., it does not, however, measure whether these beliefs are actually correct). Since there is already so much ambiguity in the field, especially regarding strategies, willpower, and self-control more broadly (e.g., see Werner & Ford, 2023; Werner et al., 2022 for some overviews, among others), I would strongly urge the authors to be exceptionally careful with how they describe different constructs, both in this specific instance and in general.“

Response: Thank you for pointing us towards this inaccuracy. We have rewritten that paragraph to argue more generally that metacognitive knowledge (e.g., about one’s weaknesses) might motivate a broader range of strategies (e.g., to compensate weaknesses) and that metacognitive regulation might enhance the implementation of the strategies (e.g., planning might help to actually enact a chosen strategy). We hope this clarifies why we believe that the MCSC might be related to the SCSS.

“Case in point re: conceptual ambiguity: in the very beginning of the introduction, the authors talk about willpower as a strategy. In the section on lay theories of willpower, they further state: "Lay beliefs about willpower refer to people's beliefs about the nature of self-control (or willpower)" therefore suggesting that willpower = self-control more broadly.”

Response: Thank you for pointing out this unclarity. We absolutely agree that the use of the relevant terms should be consistent across the manuscript. Thus, we have rephrased the definition of lay beliefs about willpower to “people’s beliefs about the nature of willpower (i.e., effortful inhibition)” because willpower rather refers to effortful inhibition in the context of lay beliefs about willpower.

“The section on personality is quite sparse and it's not clear at this stage of the manuscript why this is included. Assuredly, there is far more connections between personality and self-control, simply because of its overlap with conscientiousness and grit (at minimum), yet that is not mentioned here. I'm not necessarily saying that this paragraph needs to be enhanced, but rather I would encourage the authors to consider whether it's really needed, especially given my earlier comments re: being more succinct in selling the importance of the work being done.”

Response: After further consideration, we have deleted the section about self-control and personality to keep the introduction more parsimonious. Thank you for bringing up this important improvement.

“Regarding the section on "Common measures of self-control", it is important to note that just because something is "classic" and everyone uses them doesn't inherently mean they accurately represent the construct of interest. For example, the limitations of the BSCS have been widely discussed (e.g., Wennerhold & Friese, 2023) and it's not even clear that this is a measure of self-control per se. Similar to the stroop task, which was actually not designed as a measure of inhibition despite psychologists using it as such (see Werner et al., 2022 for some discussion on this). For these two measures specifically, I'm not sure a lack of an association between the SCSS and either measure would indicate the measure is "useless" as the authors state in the table at the beginning of the manuscript. This is especially true for the Stroop task, as prior research consistently finds there is very little or no association between self-report and behavioural measures in the context of self-regulation and self-control (e.g., Dang et al., 2020; Lennerhold & Friese, 2020; Saunders et al., 2018; Eisenberg et al., 2019). Although there is less empirical critique of the BSCS, at the very least, I'm not sure it is appropriate to include the Stroop task here as a measure of convergent validity.”

Response: After the comments on the Stroop Task by you and Dr. Miles, we have decided to drop the Stroop task from the study. Thank you very much for making us reconsider the procedure accordingly. We still see a valuable contribution in investigating the convergent validity with the BSCS, as this was also done by Katzir et al. (2021). However, we have changed the interpretation of a lack of convergence to indicating more generally that they measure distinct constructs because as you mention a lack of convergence could also be due to the BSCS not measuring self-control per se.

"Similar to the personality section, the section on "Self-control impeding conditions" is quite sparse given the wealth of literature in this space. Again, I would urge the authors to make this introduction far more succinct so that the importance of this research really comes across."

Response: After further consideration, we have deleted the section about self-control impeding conditions to keep the introduction more parsimonious. We still plan to include those measures, but in a more exploratory fashion. Thank you for bringing up this important improvement.

Comments Dr. Bürgler

Dear Dr. Bürgler,

thank you very much for your expertise and detailed feedback on our manuscript. Below, we will respond to each point you raised in the order of the review. We believe that the manuscript has considerably improved due to your suggestions.

Detailed Comments

"1. While I generally like the SCSS, there is one larger issue I have with the scale, which also directly applies to this proposed study. The scale is heavily focused on dealing with temptations/desires. This is apparent in the introduction of the scale, in which self-control conflicts are described as "[...] a conflict that arises when you face a temptation/desire (e.g., our favorite dessert, going out with friends, sales in our favorite store, etc.) when in pursuit of a long-term goal (e.g., maintaining our health, being a good parent, excelling at work/school, saving money, being a faithful spouse, etc.) [...]", Katzir et al., 2021, p. 5. Furthermore, many of the individual items directly refer to temptations, for example, "I seek out situations in my life so that I will not face temptations" (Katzir et al., 2021, p. 5). Self-control, however, also includes initiating or persisting in aversive but goal-directed activities (e.g., getting off the couch to work out and not quitting after just a few minutes; see, for example, Bürgler et al., 2021; Hennecke et al., 2019; Hoyle & Davisson, 2016). The focus

on temptations of the SCSS therefore limits the applicability of the scale, as it leaves out important areas of self-control. For certain strategies it would be difficult to change this focus on temptations, as (nearly) every item used is specifically worded around dealing with temptations (e.g., for "Situation Selection/Stimulus Control"). However, one approach the authors might consider (I want to clearly state that this is merely a suggestion and not a necessity for the scale to be valuable) is to reword the temptation-focused items to shift the focus from dealing with temptations to a more general description of successful self-control (the authors could, for example, reword the item "I seek out situations in my life so that I will not face temptations" to something along the lines of "I seek out situations in my life that make successful self-control easier for me."). Such items could be added in addition to the non-reworded versions of these items, so the original wordings would not be lost. For other strategies, however, the focus on temptations is less pronounced or completely absent (e.g., "Reward"). Therefore, I would argue to at least change the introduction to the scale so that it also introduces the concept of self-control conflicts as possibly relating to initiating or persisting in aversive activities (see, for example, the introduction used in the Metacognition in Self-Control Scale; Bürgler et al., 2022). This way, at least certain strategies of the scale could be used to investigate self-control conflicts not directly related to temptations. Furthermore, I think the introduction of the SCSS is fairly long and includes non-essential information, so rewriting it could be beneficial in multiple ways."

Response: Thank you for this helpful suggestion with regard to the scale's focus on self-control conflicts in terms of resisting temptations. We agree that it would be valuable to broaden this focus as self-control conflicts can also be present in initiation of and persistence in activities, as you mention. We are hesitant to change the items themselves in order to keep them consistent with the original scale. However, we think changing the introduction as you suggested is an excellent way to allow for a broader definition of self-control conflicts at least to some degree. Thus, we will shorten the introduction and extend the focus of self-control conflicts explicitly to initiating and persisting in activities.

"2. One important variable that I felt was missing in Katzir et al. (2021) and also in the proposed study is habit strength/behavioral automaticity (as measured with the SRBAI; Gardner, 2012; Gardner et al., 2012). There seems to be a considerable association between certain measures of self-control and beneficial habits (e.g., related to nutrition, physical exercise, and studying, see, for example, Adriaanse et al., 2014; Galla & Duckworth, 2015; Gillebaart & Adriaanse; 2017, see also Gillebaart & De Ridder, 2015). Furthermore, habits play a major role in most (likely all) outcome measures of the proposed study (e.g., Gardner et al., 2011; Wood & Neal, 2016). I think it could be fruitful to assess habit-strength related to (at least some of) the outcomes, to, for example, investigate if the SCSS (or certain self-regulatory strategies included in it) are associated with stronger beneficial habits. Furthermore, it would be interesting to analyze if habit strength mediates the effects of the SCSS (or separate strategies of the SCSS) on relevant outcomes, similar to how habit strength mediated the effects of the BSCS on the outcomes in the aforementioned studies. Luckily, the SRBAI has only 4 items, so it can be included fairly easily."

Response: Thank you for this suggestion! We agree that investigating habit strength as an outcome is an interesting addition, thus we added measures about habit strength in three behaviors (healthy diet, physical activity and studying for an exam). Since mediation analyses on cross-sectional data are not very informative about an underlying (causal) mediation (Kline, 2015) and it is beyond the present project to achieve a sufficiently high sample size across several time points, we decided against running a mediation. Still, we will collect data on habit strength at two time points, thus, we will at least be able to make causal claims about the effect of the self-control strategies on changes in habit strength as a first step to the mediation.

“3. In terms of personality variables, I think the Multidimensional Self-Control Scale (MSCS; Nilsen et al., 2020) might be worth considering. The scale differentiates between six dimensions of self-control (procrastination, attentional control, impulse control, emotional control, goal orientation, and self-control strategies) as well as two higher order factors (inhibition and initiation). Having a more fine-grained assessment of individual differences in self-control could help to disentangle correlations between self-regulatory strategies (full SCSS and single strategies) and dimensions of self-control (as assessed with the MSCS). Self-control is a highly complex construct with several layers, which means that such findings could be relevant for the field.”

Response: Thank you very much for pointing us towards this interesting measure. We have discussed the inclusion of the MSCS and came to the conclusion that the current project is already assessing a broad array of interesting variables. We hope that you follow our argumentation that this is beyond the scope of our advances and we hope to include this idea in a future project.

“4. Katzir et al. (2021) reported disappointing findings on the strategy "Acceptance": "[...] we were quite surprised that acceptance was negatively associated with the BSCS, all other strategies, and most outcomes (with the exception of distance from ideal weight)", p. 12. Katzir et al. (2021) furthermore wrote: "It is possible that our acceptance items do not appropriately capture acceptance and are interpreted by participants as giving into temptation. It is also possible, however, that people do not benefit from acceptance because using it may lead to a cycle of self-control failure by normalizing such failure (De Witt Huberts et al., 2012; Prinsen et al., 2018). The nature of the relation between acceptance and self-control awaits future research". I would have liked the authors of the proposed study to address these limitations and calls for future research more directly.”

Response: We agree that it would be interesting to dig deeper into possible reasons why acceptance showed almost exclusively negative relationships to outcomes. However, we do not think this is possible within the scope of the present project which already addresses several questions. We will, however, pay special attention to the relationship of acceptance with the outcomes in our studies. As research suggests that strategies’ effectiveness might differ between domains and contexts (e.g. Werner and Ford, 2023 for an overview), we might find that acceptance is effective in some of the additional outcome domains (e.g., satisfaction). If a similar pattern of mostly negative relationships to outcomes emerges, we will discuss this further in the discussion part of the report.

“5. One relevant limitation of the SCSS is its length. Especially for fields outside the immediate area of self-control research, it may simply not be feasible to use such an extensive scale and researchers might then just default to using other self-control related scales, which would likely be the Brief Self-Control Scale (Tangney et al. 2004). Therefore, a major contribution to the field could be to create a shortened version of the scale. This shortened version might include ~2-3 items per strategy and it might also leave out certain strategies that showed less encouraging results, such as the strategy "Acceptance" (see point 4.).”

Response: Thank you very much for this point and we completely follow your argument that this would improve the field meaningfully. We did have this in mind for a follow-up project, to which we hope to invite the authors of prior studies to contribute their data as well. This would allow us to develop a short-scale which is hopefully invariant by language and country (limited to the scope of covered languages and countries). By doing so, we hope to a) benefit from the joint power of several studies and to b) avoid a method divergence, to assure that the SCSS-short would mean the same for researchers. To strengthen the point that this is an actual plan, we developed a pre-registration for secondary data analysis (https://osf.io/pfdt2). Further, we included a section in the manuscript, which will lead to the pre-registration.

“6. I wondered about the cutoffs used to assess the model fit of the scale. The authors write "Model fit will be seen as sufficient with CFI > .90, TLI > .90, RMSEA < .10, and SRMR < .10." (p. 22), without giving any reference or justification for these values. Here, more stringent cutoffs are possible, for example, CFI ≥ .95, TLI ≥ .95, RMSEA ≤ .06, and SRMR ≤ .08 (Hu & Bentler, 1999; but see also Kyndt & Onghena, 2014). I don't necessarily think that one must adhere to these stricter cutoffs, but, looking at Table 3 in Katzir et al. (2021, p. 6) it shows that they report fit indices that seem to be more in line with these more stringent cutoffs, at least for the RMSEA and SRMR. I would have appreciated some sort of justification for how the specific cutoff values were chosen.”

Response: This is a very good point and we adjusted the manuscript, based on the recommended literature (Hu & Bentler, 1999). We originally planned to be a bit more lenient with the cut-offs in the face of measurement invariance testing, but agree that the more conservative cut-offs should be preserved to assure high measurement quality.

“7. Katzir et al. (2021) note in the limitations section of their paper: "Another shortcoming is that we only used self-report measures to assess self-control related outcomes." (p. 12). In the proposed study, many of the outcome variables are still self-reports (e.g., "Healthy Diet"

"Physical Activity", and "Sleep Procrastination" pp. 18), which is perfectly fine. Importantly, objective measures such as grades are also used, which I highly appreciate. Other measures, however, are somewhere in between, such as "Steps" and "Screen Time". They could be described as objective, because they are assessed by the smartphone, but participants then report these numbers themselves, therefore, they could still be subject to, for example, false reporting and demand characteristics. This problem could be alleviated by having participants provide screenshots to verify their reports. I appreciate the fact that this will come with additional effort and might introduce some data security and privacy concerns that would need to be addressed. However, it might be worthwhile, as objective measures (in addition to subjective ones) are needed in this line of research (e.g., Smyth et al., 2022). "

Response: This is a very good and interesting idea, which has the potential to strengthen the inference greatly. To allow this approach, we have relocated the assessment of average daily steps from the Prolific sample to Study 4 (collected via social media) where participants are also asked for their screen time. To reduce drop-out rates, we would ask participants to upload their screenshots at the end of the survey. We assume that respondents who are completing the survey from their mobile phone are more likely to upload a screenshot. Thus, we deemed Study 4 which is recruited via social media to be a better fit than Study 3 which is recruited via Prolific. For both outcomes, we will do the analysis with both the data from the self-reports and from the screenshots to see whether the results differ. We hope this solution is seen as appropriate. We value the idea greatly and are looking forward to the increased insights by this assessment.

"8. Regarding the translation process: Seeing that you aim to undertake multiple studies with thousands of participants, I would suggest running a quick pilot-study (N = ~25-50) with the translated items to confirm that people understand the items and to leave them the option to give feedback on possible points of confusion. It is especially important that the somewhat lengthy introduction is understood well by the participants, even more so if you choose to reword it, as I suggested in point 1."

Response: That is an excellent suggestion that we will implement. We have added a small pilot study before the main studies.

"9. I am not quite sure about the rationale for the cutoff of 35,000 regarding the average steps per day (p. 18). While 35,000 steps is surely unreasonably high for a monthly average, I would have still liked a justification of this number."

Response: Thank you for pointing us to this arbitrary decision. The cut-off was indeed chosen by face validity and without scientific justification. Conducting literature search after your comment helped us to refine our justification here. Based on the meta-analysis by Paluch et al. (2022) on the effect of daily steps and all-cause mortality, the highest number of average steps reported was 20,784 (*SD* = 4,176), extracted from the Baltimore longitudinal study of aging. This illustrates, from our perspective, that the cut-off at 35,000 steps is justified, as it is more than 3SD above the highest reported average of 15 studies. To make this line of reasoning available to the readership, we included a respective paragraph into the manuscript.

"10. Goal importance: I appreciate that this control variable is assessed for three of the outcome variables (healthy diet, screen time reduction, and pro-environmental behavior, see Table 2). However, I did not quite understand why these outcomes were chosen and not the others as well (e.g., physical activity). As the authors themselves report, Katzir et al. (2021) wrote that "In some cases the effectiveness of a strategy also depended on goal importance, e.g., pre-commitment and behavioral inhibition were only positively related to exercise if exercise was considered an important goal." (as it is written in the report on p. 9). It seems crucial to me to assess goal importance (or to recruit people for which one can very reasonably assume that the goal in question is important, which might be the case for students and academic achievement), because if a person would rate a goal as "1 = not at all important", it begs the question if one could even call it a matter of self-control."

Response: Thank you very much for this input. We have discussed the approach of integrating goal importance about more of the outcomes in the study very closely. We decided against including goal importance for all of the measures to limit the length of the studies. We believe that goal importance will be most important concerning pro-environmental behavior (because people will likely vary most here). Besides, we have slightly changed the variables for which we assess goal importance to better represent the different domains: we will assess goal importance for two outcomes from the health domain (healthy diet, physical activity) and one from the achievement domain (studying). With these, we aim to test a possible moderating effect exemplarily. We did not include goal importance for the life / relationship satisfaction domain as we assume that these are quite universally aspired. We hope this approach appropriately incorporates your ideas into our project and we are very happy for this extension of our inference.