Round 2: How long does it take to form a habit? A Multi-Centre Replication

Dear Zoltan,

Thank you and the two expert reviewers for the in-depth review and raising many insightful points. We address each of the comments in detail below, and hope that you find our revision acceptable for publication.

Kind regards,
Lukas

Dear Lukas

I now have three positive reviews back for your submission, which are as a whole enthusiastic about the planned research. In addition to other points raised by the reviewers, also address the following in your revision:

1) Wood raises the issue of whether data has already been collected. You say in your cover letter that it has not, but the verb tense used in other places implies it might be otherwise e.g. "The detailed study protocol, materials, anonymized raw data, code used in the analyses and output are permanently stored on Open Science Framework (https://osf.io/n6srx/)" (which I didn't have access to), and the frequent use of past tense for Method. To be clear, assuming data have not been collected, for Stage 1 use future tense throughout for things that have not happened, including Method (and e.g. "the data will be permanently stored.."). Future tense can then be changed to past tense in all cases when the Stage 2 is submitted. (Of course if data have been collected, than clarify this as well, and also therefore what precautions are in place for bias control.)

No data had been collected at the time of submission. However, we commenced with the data collection in early September. The decision to start data collection prior to acceptance conforms with our statement in the first submission (latest date for data collection start: August) and is based on the fact that the reviewers' feedback mainly pertained to our analysis plan. To avoid confusion, we changed past to future tense throughout the manuscript and changed the bias level of our submission.

2) In the results section indicate *exactly* what analyses you will do. The specification should be so clear that there is no analytic flexibility left. Make sure for example that anyone could fit the curves exactly the same way you will, and so they could precisely reproduce your results with your data. Be clear what the other "plausible curve" is that you will fit in addition (I know you refer to the original authors, but this information should be in your mansucript if it figures in your analytic pipeline.). be clear exactly what comparisons are done with what error control (more on this below). Another example "We will also perform multiple regression analyses to determine whether impulsivity, personal need for structure and conscientiousness were related to curve parameters and performance variables." State how many regressions you will perform. Specify the variables for each regression. If you mention analyses in the results section also put them in the Design Table and justify power for each analysis and what conclusions hang on each. Alternatively, leave out mentioning any analyses you do not want to tie down exactly (nor make sure are properly powered) and put them in a non-pre-registered section in the Stage 2. (There are other cases where the analysis needs to be tied down I have not mentioned.)

For clarification purposes, we have adjusted the manuscript as follows:

> “in addition to the exponential shape that was adopted by Lally et al., we will include another plausible (S-shaped) curve to model the relationship between repetition and subjective automaticity. This curve is based on the CDF of the Poisson distribution, $Q(\lfloor B \rfloor+1,\mu)$, and was found to be able to give a good fit to S-shaped learning curves (Murre, 2014; Murre et al., 2013). In such a curve, the initial portion is relatively flat, followed by a steeper ramp that levels off to asymptote with prolonged habit formation (see also Fournier et al., 2017); an exponential curve is a special case of this equation.”

We have decided to drop the analyses with the personality-related questionnaires (impulsivity, personal need for structure, and conscientiousness) from the present manuscript, as our central aim is to replicate the original findings. While these questionnaires were included in the original study, the results were not reported in the publication of Lally and colleagues. Therefore, we have omitted these analyses from the present manuscript (but of course we still report that these were included in the Methods section, and the results will be published in a separate future manuscript that will refer to the present publication).

3) Relatedly, in comparing linear and s-shaped functions you say you will do a sign test on R2's. But you also provide cogent reasons for why this is a flawed strategy, and hence say you will also use BIC and AIC to aid the reader. This leaves a lot of inferential wriggle room. A possible outcome is higher R2 for the s-shaped than linear function because of the difference in parameter numbers, but BIC comes down the other way. Has your manipulation check on the form of the function failed or not? Justify one measure as most suitable - and I presume it will be either BIC or AIC or related, e.g. an informed Bayes factor - and test the form of the function with that measure, stating your criteria for good enough evidence for choosing s-shape over linear functions (or one s-shape over another).

This characteristic follows from our aims to do both 1) perform the analyses as done by the original author to circumvent the critique that differences in findings may result from differences in the analysis approach, and 2) perform the improved analysis that will yield more reliable estimates. For the first aim, we will use the R-squared measure of goodness-of-fit to determine whether the relationship between repetition and automaticity is modelled successfully by an asymptotic curve, and a Sign test to determine whether there is a significant difference with a linear function. These analyses are conducted separately for each of the four datasets.

For the second aim, the improved analyses approach that falls outside the scope of replication, we will utilize the BIC exclusively, which is considered to be superior to R-squared and can be of aid to the reader to assess our results (additional goodness-of-fit measures will be reported alongside the results).

4) As per my previous point, it seems to me that providing evidence for your exponential function is more an outcome neutral test than a test of theory. You point out that the original authors were motivated by Hull's theory of habit; and thus you frame the choice of function as testing Hull's theory, by testing the predicted shape of the function defining the increase of habit strength over time. However, strength will be measured on Likert scales, therefore with fixed minimum and

maximum values. A linear function is therefore a priori ruled out if testing continues long enough. Something at least approximating an s-shape is guaranteed by the nature of the measurement. Therefore no theory can be at stake depending on the outcome of this test. Rather, obtaining something like an s-shape is a necessary precondition for your study in order to estimate when a habit has formed.

This is a valid point. Even though in the original sample most participants did not reach the extreme end point of the scale, their scores could still be skewed. We included this hypothesis and rationale in our analysis plan in the same way as the original authors of the to-be-replicated study had done, but we agree with you that an exponential shape is to be expected with a Likert scale measure with a minimum and maximum value, and therefore would not provide convincing evidence for Hull's theory. Therefore, we have decided to remove this hypothesis from our paper and study design table, and instead merely include it as an analysis step (as indeed an exponential shape is a pre-condition). We also adapted the title of our manuscript in this light, as "the shape of habits" emphasizes this aspect of our study. The new title is: "How long does it take to form a habit? A Multi-Centre Replication".

5) I take it you plan to perform all pairwise comparisons between the 5 data sets (your 4 plus original) with Kolmogorov-Smirnov tests, which is 10 tests. How will you control familywise error rate? Given that corrected alpha, determine N to control power to detect a difference you are interested in. How far apart should median number of days be before it is interesting? As Gardner points out, 66 days is just a rough figure; 59 is in the same ball park. Is there any way (other than reaching deep in one's soul) for specifying how far away from 66 would start to be interesting? (e.g. https://psyarxiv.com/yc7s5/) Once you have justified an interesting difference, use simulations to determine the power of KS to detect the effects that are just interesting. (And you should note in the paper that KS is sensitive to more than location differences, as a proviso on your analysis.)

Based on your feedback, we have re-examined and adjusted our approach in two ways:

Firstly, to determine whether we replicate the original finding of 66 days, we will fit and combine the individual curves for each of the four datasets separately to determine the corresponding 95% confidence intervals for the median of each dataset. To give you an impression of how lenient this criterion is: in the original dataset (with 39 inclusions), the 95% confidence interval was given by the values ranked 13 to 27, resulting in an interval between 48 and 81 days. In our replication attempts, we expect to include more participants (as we will sample 200 per site as opposed to the original 101), which should allow us to narrow down the confidence interval and obtain a more precise representation of the median. As we doubled the sample size, we anticipate that our confidence interval will approximately range from the 30th to 49th values, which would cover 24% of the data.

The question of replication is based on a majority rule across the four datasets, with 66 days not falling in the confidence intervals of two or more sites constituting an unsuccessful replication. If the finding does not replicate at a particular site, it means that the generalizability of the original finding is limited and that further research is needed to reveal the underlying cause(s) of the divergent findings.

For even more powerful analyses of potential sources in intra- en interindividual variability and to further constrain the confidence interval, we will combine our data sets to investigate (as in the original study) the effects of consistency (see RQ2; Table A1) and complexity (i.e., behaviour type: drinking/eating/exercising; see RQ3; Table A1) on automaticity development. The effect of omissions will be investigated by comparing automaticity immediately preceding and following a single missed opportunity or ‘omission’ (contrasting the definition of Lally et al., who defined an omission as preceded by three occasions when it had been performed) using Wilcoxon signed rank tests. Furthermore, this difference in automaticity will be compared to situations when the behaviour was performed on three consecutive days.

Similarly to Lally et al., complexity will be investigated by comparing the estimated curve parameters (i.e., a, b, and time to reach 95% of asymptote) and performance variables (i.e., number of reported repetitions and percent compliance) between eating, drinking, and exercise behaviours. To this end, we will perform five univariate ANOVA’s on each of the dependent variables. For all the ANOVA’s that turn out significant we will perform post-hoc pairwise comparisons to examine which behaviours drove the effect. We will correct for the multiple testing error rate using the Tukey-Kramer's method. Tukey-Kramer was chosen since we expect the sample sizes to be unequal between the groups (participants choose the type of behaviour themselves, therefore we have no control over the group sizes).

6) For RQ3, specify an interesting effect size in raw units: What difference in rated automaticity would be just interesting? (When several Likert ratings are combined, I find using an average rather than sum over number of ratings useful to put the final number on the same scale as the rating itself, so one has a more intuitive grasp of what one unit is.) Otherwise we just have "medium effect size" plucked out of the air; and being standardized it depends on measurement noise and reliability. But presumably what is interesting is the actual difference in automaticity.

We agree that the effect expressed as the average difference on the Likert scale makes for a more intuitive interpretation for readers. Consequently, we will report both effect size and the average difference score. Based on the abstract nature of our self-report automaticity measure and limited knowledge of its relation to real-life implications (e.g., whether the target behavior persists over a longer time interval, or the extent to which effort is reduced and mental capacity is freed up for other activities) we are reluctant to pinpoint an interesting effect size.

7) In terms of what defines an interesting effect, Takacs asks "for RQ4, would a non-significant result prove that "complexity is irrelevant for automatization"?" You can just qualify the conclusion that it applies to this particular difference in complexity. (If in addition you could 6uantify or measure the complexity difference (and I am not requiring you do) it would help place the conclusion in perspective too.) How will you control familywise error rate for number of Dvs? Determine interesting differences in raw units, then determine power for those I, taking into account the corrected alpha. Specify the IV and its levels – is it more than 2 as you are performing a KW? Will there be post hocs? What conclusions follow from different patterns? There is also inferential flexibility in specifying both ANOVA and KW. Justify one, or provide a decision procedure for choosing between them (one that does not allow wriggle room).

Our conclusions are limited to this particular difference in complexity, as such, we will qualify our conclusions to pertain to this difference specifically.

We will perform five univariate ANOVA’s on each of the dependent variables. For all the ANOVA’s that turn out significant we will perform post-hoc pairwise comparisons to examine which behaviours drove the effect. We will correct for the multiple testing error rate using the Tukey-Kramer's method. Tukey-Kramer was chosen since we expect the sample sizes to be unequal between the groups (participants choose the type of behaviour themselves, therefore we have no control over the group sizes).

Round 2: How long does it take to form a habit? A Multi-Centre Replication

1st Reviewer: Benjamin Gardner

This is a note-perfect replication of a seminal study on habit formation (Lally et al., 2010). The report meets all criteria for a Stage 1 replication study: the research questions are scientifically valid; the hypotheses are logical and plausible; the methodology is sound and feasible, and as described, permits replication. I have only two comments.

Thank you for this positive appraisal of our proposal.

8) One important methodological difference between the original study and the present study, as the authors openly acknowledge (on p6), is that, whereas the participants in the original study met with the researcher in-person in a lab, replication study participants will meet the researcher online via video conferencing. This is important because motivation is needed to initiate and maintain a habit formation attempt before habit solidifies. This difference could feasibly affect results in two ways. First, providing support, advice and/or guidance in person might be inherently more motivating than doing so online. Second, participants who are willing to travel to a lab in central London to participate (as in Lally et al's study) may be inherently more motivated than those who are only required to meet via video conferencing. Do the authors view the difference in meeting format as a problem, and if so, how might it affect their results, and to what extent might this mitigate this?

This is an important difference between ours and the original study. It is possible that this difference will influence the motivation (in ways described by the reviewer), and we will acknowledge this point in the discussion of our manuscript if compliance turns out to be low. The reason we did choose to proceed with online testing is feasibility during COVID, as well as the time constraints of the grant that funds this project. Importantly, to minimize the impact of this methodological difference with the original study, participants will still meet with an experimenter through video conferencing, who will instruct and motivate them in the same way as in the original study.

9) Hypothesis 2 focuses on testing whether habit really does peak after 66 days, as Lally et al found. This seems overly restrictive; even if Lally et al's findings are 'true', I very much doubt that a replication of this result would find habit to peak at exactly 66 days. (For example, Keller et al [2021] found a once-daily behaviour to peak in habit strength after 59 days. While not exactly 66 days, this finding intuitively appears in keeping with Lally et al's findings.) Will the authors conclude that Lally et al's findings have not been replicated if the peak habit duration is NOT 66

days? Or is there an acceptable range within which a peak other than 66 days might sit *and* Lally et al's findings be supported?

This is a valid point and we have addressed it in the reply to the editor (Q5). To determine whether the original finding of 66 days is replicated, we will assess whether 66 falls within the 95% confidence interval in a majority of the four datasets.

2nd Reviewer: Wendy Wood

This is an important research project that proposes to replicate an earlier investigation by Lally et al. (2010). The authors are correct that this earlie investigation had very few participants and consequently unstable results, despite that it has been cited over 2000 times on GoogleScholar. The opportunity to replicate this research and in addition to assess individual differences makes this a highly useful piece of research for science and for popular understanding. For these reasons, I believe it should be accepted for publication.

Thank you for this positive evaluation of our proposal.

The guidelines for evaluating registered reports we were given are:

1) The scientific validity of the research question(s): I take this as a question of how important is the research, and I answered this above.

2) The logic, rationale, and plausibility of the proposed hypotheses: This project will be informative whatever the results.

3) The soundness and feasibility of the methodology and analysis pipeline: The research has apparently already been conducted, and the data analysis is already in progress (?). For this reason, I am not going to comment in any detail on the methods and procedures except to note that it would be helpful to add a kind of intention-to-treat analysis that assesses the effects of participant attrition on the conclusions drawn.

4) Whether the clarity and degree of methodological detail is sufficient: The authors rely heavily on the original project and the protocol given, and so clarity is presumably assured.

5) The soundness and feasibility of the methodology and analysis pipeline: The research has apparently already been conducted, and the data analysis is already in progress (?). For this reason, I am not going to comment in any detail on the methods and procedures except to note that it would be helpful to add a kind of intention-to-treat analysis that assesses the effects of participant attrition on the conclusions drawn.

Apologies for the confusion. No data had been collected at the time of our first submission, and it was only in the beginning of September that we started data collection. This decision was based on our previous note in the first submission and the fact that the reviewer's feedback mainly pertained to the analyses plan. To avoid further confusion, we have changed past tense to future tense throughout the manuscript.

We agree that it would be interesting to explore reasons for dropout. Therefore, depending on the severity, we will explore this question (and the impact thereof) in additional posthoc analyses. These fall outside of the present replication attempt; hence, we do not elaborate on them here. Some insight into the effect of different degrees of participation may follow from our analysis on the effect of omissions and compliance on habit formation (RQ2).

6) Whether the clarity and degree of methodological detail is sufficient: The authors rely heavily on the original project and the protocol given, and so clarity is presumably assured.

7) Whether the authors have considered sufficient outcome-neutral conditions: In this case, the one major threat to validity is the daily report procedure. The authors do not address this or provide any insight into how they will handle it. But it leaves readers wondering whether the results would be the same if participants were not reminded each day about their behavior by the habit questionnaire.

In the original study that we aim to replicate, the experimenter offered participants to send them a daily reminder email to fill out the questionnaire, and the majority opted for this (verbal communication by Dr. Lally). However, a difference that we cannot circumvent is that nowadays many people receive a push message on their phone upon receiving an email, which renders these even more salient. To investigate whether this affected our results, we will perform exploratory analyses (see 'exit questionnaire'). The question remains whether habit formation is affected by filling out a daily habit questionnaire (in the original study as well as our replication attempt). This is a relevant issue that we will touch upon in our discussion section.

8) Although I am very favorable toward acceptance, I wonder why the authors are submitting a registered report Stage 1 for a project with already-collected data (Level 5 o 6?). How far have the authors proceeded with data collection and analysis? I think that the project is so worthwhile, it should be attractive at a number of journal outlets. So, I'm not clear why this publication format.

Apologies for the confusion, please see our reply to 1C.