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

de Wit, S.1,2, Bieleke, M.3, Fletcher, P.C.4,5,6, Horstmann, A.7, Schüler, J.3, Brinkhof, L.P.1,2, Gunschera, L.J.1, Murre, J.M.J.1,2

1 Department of Psychology, University of Amsterdam, Postbus 15933, 1001 NK, Amsterdam, The Netherlands

2 Amsterdam Brain and Cognition, University of Amsterdam, Postbus 15900, 1001 NK, Amsterdam, The Netherlands

3 Department of Sport Science, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany

4 Department of Psychiatry, University of Cambridge, UK

5 Cambridgeshire and Peterborough NHS Trust, Cambridge, UK

6 Wellcome Trust MRC Institute of Metabolic Science, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK

7 Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Haartmaninkatu 3, 00290 Helsinki, Finland

**CRediT Statement**

**de Wit:** Conceptualization, Validation, Data Curation, Supervision, Project administration, Writing – Original Draft Preparation, Supervision, Funding Acquisition

**Bieleke:** Conceptualization, Supervision, Writing – Review & Editing

**Fletcher**: Conceptualization, Supervision, Writing – Review & Editing

**Horstmann:** Conceptualization, Supervision, Writing – Review & Editing

**Schüler:** Conceptualization, Supervision, Writing – Review & Editing

**Brinkhof:** Methodology, Funding Acquisition, Writing - Review & Editing

**Gunschera:** Software, Investigation, Data Curation, Project administration, Writing – Review & Editing

**Murre:** Conceptualization, Data Curation, Formal Analysis, Visualization, Writing – Original Draft Preparation, Funding Acquisition

**The Shape of Habits: A Multi-Centre Replication**

How long does it take to form a habit? This question will be addressed by an innovative study by Lally et al. (2010), in which they tracked the subjective automaticity of a novel, daily (eating or exercise-related) routine, using the Self-Report Habit Index. They showed that the gradual automatization of a novel routine is best described by an asymptotic curve, and that it takes (a median of) 66 days to reach the asymptotic ‘habit plateau’, with a range of 18 to 254 days (based on statistical extrapolation). However, these findings were based on a small sample of 39 participants, and this influential study has not been replicated yet. Therefore, the aim of the present study is to conduct a near-exact, multi-centre replication at four different locations. We aim to recruit 800 participants to increase reliability, and to allow us to investigate sources of interindividual variability in habit formation.

**Keywords:** habit formation, automaticity, behavioural complexity, behavioural consistency, individual differences

**Introduction**

Even small changes in health-promoting behaviours, e.g., diet and exercise, can have major health benefits, extend lives, and significantly reduce healthcare costs (Kelly et al., 2009; Nocon et al., 2008). However, despite good intentions, people often fail to change their behaviour accordingly. An important cause of this ‘intention-behaviour gap’ (Sheeran & Webb, 2016) may be that initial, deliberate goal pursuit requires effort and discipline. The formation of habits that can be triggered automatically by contextual cues may help to bridge the gap. In the words of William James (1890), *we must make automatic and habitual, as early as possible, as many useful actions as we can*. But how long does it take to form a new habit? And why are some people more successful than others?

These important questions were addressed by an innovative study by Lally and colleagues (Lally et al., 2010). They asked volunteers to form a new healthy (dietary or exercise) routine across 12 consecutive weeks and to report daily on the subjective automaticity of this behaviour by filling out the Self-Report Habit Index (SRHI; Verplanken & Orbell, 2003). This allowed the researchers to track and model the gradual automatization of a novel daily routine within individuals. They found that the first repetitions led to greater gain in automaticity than later repetitions, and eventually behaviour no longer became more automatic. The subjective automatization of a novel, daily routine was therefore best described by an asymptotic curve. They also found that it took participants on average (median) 66 days to reach their ‘personal habit plateau’, with a range from 18 to 254 days (based on statistical extrapolation). However, this finding was based on a small sample of only 39 participants, and this study has not been replicated yet.

We consider the study of Lally et al. (2010) an excellent candidate for replication for several reasons. First and foremost, their quantitative within-subject habit tracking and modelling procedure can be used to shed light on how habits are formed. This innovative approach has so far been adopted by only two published studies (Fournier et al., 2017; Keller et al., 2021), and it can lead to interesting theoretical insights, but importantly also inform behavioural strategies and interventions that aim to achieve lasting behaviour change, by shedding light on the number of days it takes to form a habit (Verhoeven & de Wit, 2018; Wood & Rünger, 2015). Accordingly, despite its small sample size, this study has had a major, international impact on the field of habit research and behaviour change, and continues to be widely cited in peer-reviewed scientific journals. Its influence has not been confined to psychology, but also includes, for example, public environmental occupational health, business economics and science technology. The original study has also made an impressive impact beyond the academic literature. The finding that it takes a median of 66 days is widely advertised through popular science magazines, newspapers, websites and blogs, and therefore greatly influences the public perception of how long it takes to form a habit. The finding of a (median) 66 days has informed coaching programs and mobile applications aiming to support habit formation.

 Therefore, we will conduct a near-exact, multi-centre replication of the study of Lally and colleagues (2010). We will adopt the original protocol to track and model the development of a health-related habit, but with a greater sample size and independently at four different locations (Amsterdam (The Netherlands), Cambridge (UK), Konstanz (Germany) and Helsinki (Finland)) to increase reliability and generalizability. Next to replicating the original analyses, we will apply recent advances in statistical modelling of acquisition curves to gain insight into how habits are formed and test our hypotheses (H in Table A1). This will allow us to determine whether the relationship between behavioural repetition and subjective automaticity can be modelled best by an asymptotic curve, and whether it indeed takes a (median) of 66 days to form a habit (see H1; Table A1).

Next to modelling habit formation, we will also explore potential causes for interindividual variability, including the consistency and complexity of the habit (Gardner et al., 2021). Lally and colleagues already attempted to do this but could not draw strong conclusions due to their small sample size. We will determine whether missing a single opportunity to perform the behaviour (i.e., lower consistency) compromises habit formation (Armitage, 2005; see H2, Table A1), and investigate whether it takes more repetitions to turn a complex behaviour into a habit (i.e., exercise as opposed to simple eating or drinking behaviour; see H3, Table A1) (Kaushal et al., 2017; Verplanken, 2006).

**Methods and Analyses**

**Participants**

The replication project will be conducted with a consortium consisting of: the University of Amsterdam, The Netherlands (Dr. S. de Wit and Prof. J. Murre); the University of Cambridge, UK (Prof. P.C. Fletcher); the University of Helsinki, Finland (Prof. A. Horstmann); and the University of Konstanz, Germany (Dr. M. Bieleke and Prof. J. Schüler). We plan to test 200 participants at each of the four sites (between 21-45 years). They will be recruited via the universities’ websites, crowdsourcing software, and social media. Each participant will be paid €50 for their participation if they complete the study. If they drop out in between the first and second meeting, or if they make less than four data entries during the first four weeks, data collection will be terminated and they will receive 20 euros. In the case of dropout, we will test additional participants (as far as our replication budget of 800\*50 euros allowed and before the end of the project (end of May ‘23)). Each site aims to include roughly 50% university students (and therefore a comparable sample to the original study) and 50% from a non-student population.

This study will be executed in compliance with relevant laws and institutional guidelines, aligned with the most recent Transparency and Openness Guidelines, and approved by the local ethics committees of the Universities of Amsterdam, Cambridge, Konstanz, and Helsinki. 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/>) and an overview of our study design is presented in Table A1.

*Determining the sample size.* Power analyses are commonly used to determine the sample size required for the power needed to find, e.g., the smallest still relevant effect. However, this traditional approach is not appropriate for this replication study since the main analysis does not directly compare groups. Instead, it takes all individual curves together to determine the median number of repetitions to reach the plateau of automaticity. The precision of the median of Lally et al. can be determined by taking the 95% confidence interval for the median, using the following equations (Conover, 1980; Hazra, 2017):

1. The lower 95% confidence limit is given by the $\frac{n}{2} $- $\frac{1.96√n}{2} $th ranked value.
2. The upper 95% confidence limit is given by the 1+ $\frac{n}{2} $+ $\frac{1.96√n}{2} $th ranked value.

For the number of participants for whom the nonlinear model was a good fit in the original analysis (N = 39) and who could thus be used to determine the number of repetitions needed to reach a plateau of automaticity, the 95% confidence interval for the median is given by the values ranked 13 to 27 (which covers approximately 36% of the data). Narrowing down this rather large interval will allow us to obtain a more precise representation of the median. To this end, we propose to increase the sample size by factor 2 (N = 78), resulting in an interval ranging from the 30th to 49th value that will cover only 24% of the data. We estimate that this can be achieved by including 61% more participants, based on the fact that in the original study only 39% of the initially included participants (39 out of 101) could eventually be used in the main analyses, either due to voluntary withdrawal or analysis-based exclusion. Therefore, we aim for a primary sample of 200 participants (per site). This sample size is based on exact replication of the original analysis. As our additional analysis will allow a greater number of participants to be included, this will increase the precision of the median estimate . Importantly, this greater sample size also allows for more reliable results regarding the influence of individual differences.

**Procedure**

For the purpose of replication, we will adopt the original study protocol (as illustrated in Figure 1). During this 12-week study, participants meet with the experimenter three times via video conferencing. Prior to the first meeting, they provide informed consent and completed the Barratt Impulsiveness Scale (Patton et al., 1995), the Personal Need for Structure scale (PNS; Thompson et al., 1989), and conscientiousness items from the International Personality Item Pool (Goldberg, 1999). Although these three questionnaires were not included in the original publication, they were in fact part of the original study protocol. The results of these questionnaires fall outside the scope of the present replication attempt, and will be analysed for a separate publication, which will refer to the present manuscript.

Subsequently, during the first meeting, participants selected a healthy eating/ drinking/ exercise behaviour that they were motivated to perform daily and turn into a habit. This target behaviour should meet four criteria. It should be *something... (1) they don’t already do; (2) they can do in response to an event in their day (a situational cue) (3) that only occurs once a day but every day (4) they believe they can realistically achieve every day*. For example: *‘doing squats when waking up’*. When they have chosen a target behaviour, participants are asked how often they performed this behaviour over the past four weeks. In the present study, if they performed it four times weekly or even more frequently, they will be asked to choose a different target behaviour. Finally, they will be asked to rate anticipated effort of performing this behaviour every day and to rate their intention (see Materials for details), and they schedule meeting 2 and 3.

Participants will be instructed to start performing the behaviour on the day after the first meeting. The first email with a link to the automaticity questionnaire will be sent two days after the first meeting. It will always be sent at 8.00 AM. This questionnaire asks them to rate the (current) automaticity of the behaviour using the Self-Report Habit Index (SRHI; Verplanken & Orbell, 2003; e.g., *‘doing squats when I wake up is something I do automatically’)*, and whether they performed the behaviour the previous day. If they fail to fill out the daily questionnaire, they can report whether they have performed the behaviour retrospectively for up to three previous days. Finally, if they report that they did not perform the behaviour, they will be asked to indicate the reason from a *list (I was not in the situation/the cue did not occur; I forgot; I wasn’t prepared; I chose not to do it; Other).* Again, the latter question was not included in Lally et al. (2010), but it was in fact part of their study protocol.

The second meeting will be scheduled 4-5 weeks after the first one. The main purpose of this meeting is to boost compliance. Furthermore, we will ask participants whether their daily routines were disrupted by external circumstances during five consecutive days or longer (e.g., illness or holidays). If they answer in the affirmative, they will additionally be asked to indicate the start and end date, the circumstances (serious illness / holidays / other), and the extent to which this interfered with performing the target behaviour during that period.

Following the complete 12 weeks of the study, participants will receive an email with a link to the exit questionnaire, which they should complete before the third and final at the end of the project. After they have answered the questions listed in the Materials section, we will ask them to fill out several additional questionnaires, to investigate relevant individual differences. Those data are not part of the replication and will therefore be analysed for a separate publication, which will refer to the present manuscript (and will also include the impulsiveness, PNS and conscientiousness questionnaires). During the final meeting, participants will be thanked and reimbursed for their participation.

**Figure 1**

*Study Procedure*

*Note.* At the start of the study, participants will fill out questionnaires about personality factors. During the first online meeting with the experimenter, they will choose an exercise / drinking / eating habit that they want to form during the next 12 weeks. They will also report on the anticipated effort, enjoyment (reward) and strength of their intention. Next, they will be asked to perform this behaviour daily and to report every morning on habit strength (using items of the Self-Report Habit Index). After the first four weeks, they will have the second meeting with the experimenter and will be encouraged to continue. After twelve weeks, they will fill out the exit questionnaires. These questionnaires contain questions regarding effort, reward, intention, and self-efficacy (as well as several additional questionnaires at the end that are not part of the replication attempt). Finally, they will have their third meeting with the experimenter, and will bethanked and reimbursed for their participation.

Our replication is aimed to be near-exact. The main changes to the original protocol are summarised in this paragraph. First of all, we will conduct the three meetings online via video conferencing. Furthermore, the daily email to participants will be sent automatically through the LOTUS software ([www.lab.uva.nl/lotus](http://www.lab.uva.nl/lotus)), and the questionnaires are filled out online via Qualtrics ([www.qualtrics.com](http://www.qualtrics.com)). Due to technological advances since the original study, there is another difference that we cannot circumvent. These days, some people receive a push message on their smartphone when receiving an email. Being reminded to complete the SRHI by push-messages might affect adherence (even though these were sent early in the morning). Therefore, at the end of the study, we will ask participants whether they receive push messages or not, to allow us to conduct a control analysis to determine whether this affected the results. We also added several questions regarding their behavioural intentions to the first meeting, and regarding the target routine to the third meeting. We have uploaded the detailed study protocol on OSF (<https://osf.io/n6srx/>), with all minor changes to the original protocol highlighted, and signed approval by Dr Lally, the author of the original publication.

**Materials**

The questionnaires that are used during the screening, lab sessions, and daily habit formation are all administered using Qualtrics (www.qualtrics.com), and daily reminders are sent via the in-house Lotus software (<https://www.lab.uva.nl/lotus/help/>).

*Impulsivity.* The Barratt Impulsiveness Scale (BIS-11; Patton et al., 1995) is a 30-item self-report instrument designed to assess the personality and behavioural construct of impulsiveness. Participants rate the items (e.g., “*I am self-controlled*”) on a four-point Likert scale (ranging from rarely/never to almost always/always). The BIS-11 has demonstrated acceptable internal consistency across a range of cultures (Cronbach’s alpha range .71-.83).

*Personal Need for Structure.* The Personal Need for Structure questionnaire (PNS; Thompson, Naccarato, & Parker, 1989) encompasses 12 items and assesses the degree to which people are motivated to structure their environment in simple and unambiguous ways. Participants rate items (e.g., “*I enjoy being spontaneous*”) on a six-point Likert scale (ranging from strongly disagree to strongly agree). The scale is thought to capture two factors, factor one concerns the wish for structure whereas factor two concerns the reason for a lack of structure. Overall, the PNS has demonstrated acceptable internal consistency (Neuberg & Newsom, 1993).

*Conscientiousness.* Self-reported conscientiousness was measured with ten corresponding items of the International Personality Item Pool (IPIP; Goldberg, 1999), a measure commonly used to assess the big-five personality factors. Participants rate the ten items (e.g., “*I like order*”) on a five-point Likert scale (ranging from very inaccurate to very accurate). The IPIP, as well as the conscientiousness subscale specifically (Cronbach’s alpha = .86), have demonstrated good internal consistency (Cronbach’s alpha = .79 - .87) and test-retest reliability (Goldberg, 1999; Ypofanti et al., 2015).

*Self-reported habit strength.* We will assess the subjective experience of habitual behaviour with the Self-Report Habit Index (SRHI; Verplanken & Orbell, 2003). While the SRHI entails twelve items in total, we make use only of the seven that were included in the original analyses (Lally et al., 2010). These items include: “*I do automatically; I do without having to consciously remember; I do without thinking; That would require effort to not to do; I start doing it before I realize I’m doing it; I would find hard not to do; and I have no need to think about doing*”. Participants indicate their agreement with the statements, with respect to the selected target behaviour, on a seven-point Likert scale (ranging from strongly disagree to strongly agree). Scores range from 0-6, and, therefore, the maximal total score is 42. The SRHI has exhibited good psychometric properties across a range of contexts (Cronbach’s alpha = .81 - .95; Morean et al., 2018; Verplanken & Orbell, 2003).

*Effort, intention, and reward.* During the first meeting, we will ask participants to rate (from 1 [easy] to 5 [difficult]) “*how easy/difficult would it be for you to do [target behaviour X] everyday*”. Furthermore, we add four questions to the original protocol, informed by the Theory of Planned Behaviour (Ajzen, 1991). These have been incorporated to measure intention, attitude, subjective norm, and perceived behavioural control, and participants indicated their agreement with the respective statements on a seven-point Likert scale (ranging from strongly disagree to strongly agree). The corresponding questions are: “*I intend to (target behaviour X) every day”* (intention); *“target behaviour X) every day is good for me”* (attitude)*; “the people in my life whose opinion I value would approve of me (target behaviour X) every day”* (subjective norm)*; “I believe that I have control over whether or not I (perform target behaviour X) every day*” (perceived behavioural control). Furthermore, we have added a question about anticipated reward: “*I enjoy [target behaviour X] everyday*”. These variables will be investigated in an exploratory fashion.

*Exit questionnaire.* Upon completing the 84th day of performing the selected target behaviour, participants will receive a final email, containing the link to a Qualtrics questionnaire. This questionnaire includes the effort and intention questions that are also posed during the first meeting. Participants rate the following items on a five-point Likert scale (ranging from strongly disagree to strongly agree): *Doing this everyday got easier over time; During the study my enjoyment of the behaviour increased; During the study my desire to do the behaviour increased; During the study my belief in my ability to do the behaviour increased; This is now a habit.*Subsequently, we will pose several additional open questions (that were not part of the original study) to determine: (i) whether they received push messages on their phone upon receiving an email; (ii) whether they changed the planned target behaviour and/or cue throughout the study (and if yes, to state the new plan/cue); (iii) what the main obstacle was for forming the target habit (open question); (iv) what advise they would offer for other people; (v) how consistently they performed the target behaviour on each day; (vi) whether they interpreted the habit questionnaire to pertain to behaviour instigation or execution; (vii) whether preparatory behaviours posed an obstacle; (viii) whether they noticed effects of the new routine and whether those were motivating or not; (xi) and finally, several questions repeated from meeting 2 that pertain to the disruption of their daily routines (for an overview; see (https://osf.io/apwd3/).

**Planned analyses**

*Replication of original habit modelling.* Firstly, we will duplicate the data processing and fitting process in the original paper exactly, using SPSS (for detailed description and explanations, see Lally et al., 2010). In agreement with the original paper, participants are excluded from analyses if: (i) they fail to provide data beyond day 60; (ii) SPSS is unable to find an optimal solution after 100 iterations attempting to fit the curve equation to the data; (iii) the model generates a zero value for the *b* parameter of the fitted equation *y = a – be-cx*, which implies no learning (a flat curve); (iv) the modelled asymptote score is below 21 (indicating a lack of habit) or higher than 49 (which is an unrealistic value); (v) if the R2 value is below 0.7.

In accordance with the original analyses, both a linear and nonlinear, asymptotic regression (i.e., *y = a – be-cx*) will be run for the remaining participants. In this equation, *y* stands for automaticity, *x* for day of the study, *a* for the asymptote (or ‘automaticity plateau score’), *b* for the difference between the asymptote and the modelled initial value of *y* (when x = 0), and *c* is the rate constant that represents the rate at which the maximum is reached. These parameters are restricted to positive values and the corresponding starting values are: a = 27, b = 23, and c = 0.6. Subsequently, the number of repetitions needed to reach a plateau of automaticity (95% of asymptote) will be calculated using the following (inverse) equation: -[ln(a/20/b)]/*c* (see H2; Table A1). Following the original study, the R2 measure of goodness-of-fit will be used to determine whether the relationship between repetition and automaticity is modelled most successfully by an asymptotic curve, and a Sign test will be used to determine whether this is a significant difference (see H1, Table A1). These analyses are conducted separately for each of the four datasets. To investigate whether Lally’s et al. (2010) estimate of the time it takes to reach a ’habit plateau’ is generalizable across the research sites we will check whether 95% confidence intervals contain the original median value (66). We will use the majority rule – if at least three out of four yield confidence intervals containing the original value of 66 we will conclude that we have replicated the original finding. Conversely, if the original value of 66 falls outside the confidence intervals of more than one research site, this means that the finding is not replicated.

*Additional habit modelling.* The equation above that was used by Lally and colleagues could fit only 48% of the participants (*N* = 39), as a result of which there was a large confidence interval around the finding of 66 days to habitual behaviour. Therefore, we will perform an additional ‘improved analysis’ alongside the replication analysis.The novel approach circumvents unnecessary rejection of data and addresses weaknesses in the original approach in four different ways. This ‘improved analysis’ falls outside the scope of our replication attempt and takes the following form. First, 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 (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. Second, Lally et al. used a Sign test to show that a more complex (asymptotic) model led to an increased R2 compared to a simple, linear regression model. This, however, is unsurprising as more complex models usually fit the data better and a Sign test does not penalise the goodness-of-fit for adding a parameter to the model. This will be remedied by using the Bayesian Information Criterion (BIC), which is also more appropriate for nonlinear models for other reasons (see Spiess & Neumeyer, 2010). Third, Lally and colleagues used SPSS to fit the exponential model simultaneously with a freely varying initial value and asymptotic final value. For our additional analyses, we will instead make use of Mathematica (Version 12), which includes powerful optimizers to prevent unwarranted rejection of data. Fourth, even though the maximum possible score of their measure of automaticity was 42, they allowed the asymptote to take on values as high as 49 (an impossible value). In our additional analysis, we will use 42 as the maximum score.

*Behavioural complexity, consistency.* For even more powerful analyses of potential sources in intra- and 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. For all the ANOVA’s that turn out significant we will perform post-hoc pairwise comparisons to examine which behaviours drive 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).

**References**

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Armitage, C. J. (2005). Can the theory of planned behavior predict the maintenance of physical activity? *Health Psychology*. https://doi.org/10.1037/0278-6133.24.3.235

Calin-Jageman, R. J. (2022). esci: Effect Sizes and Confidence Intervals. R package version 0.1.1. https://thenewstatistics.com/itns/esci/jesci/

Conover, W. J. (1980). *Practical Nonparametric Statistics*. John Wiley and Sons.

Dienes, Z. (2021). Obtaining Evidence for No Effect. https://doi.org/10.31234/osf.io/yc7s5

Fournier, M., d’Arripe-Longueville, F., Rovere, C., Easthope, C. S., Schwabe, L., el Methni, J., & Radel, R. (2017). Effects of circadian cortisol on the development of a health habit. *Health Psychology*. https://doi.org/10.1037/hea0000510

Gardner, B., Arden, M. A., Brown, D., Eves, F. F., Green, J., Hamilton, K., Hankonen, N., Inauen, J., Keller, J., Kwasnicka, D., Labudek, S., Marien, H., Masaryk, R., McCleary, N., Mullan, B. A., Neter, E., Orbell, S., Potthoff, S., & Lally, P. (2021). Developing habit-based health behaviour change interventions: twenty-one questions to guide future research. *Psychology and Health*. https://doi.org/10.1080/08870446.2021.2003362

Hazra, A. (2017). Using the confidence interval confidently. *Journal of Thoracic Disease*. https://doi.org/10.21037/jtd.2017.09.14

James, W. (1890). *The principles of Psychology*. Holt.

Kassambara, A. (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. https://CRAN.R-project.org/package=rstatix

Kaushal, N., Rhodes, R. E., Meldrum, J. T., & Spence, J. C. (2017). The role of habit in different phases of exercise. *British Journal of Health Psychology*. https://doi.org/10.1111/bjhp.12237

Keller, J., Kwasnicka, D., Klaiber, P., Sichert, L., Lally, P., & Fleig, L. (2021). Habit formation following routine-based versus time-based cue planning: A randomized controlled trial. *British Journal of Health Psychology*. https://doi.org/10.1111/bjhp.12504

Kelly, M. T., Rennie, K. L., Wallace, J. M. W., Robson, P. J., Welch, R. W., Hannon-Fletcher, M. P., & Livingstone, M. B. E. (2009). Associations between the portion sizes of food groups consumed and measures of adiposity in the British national diet and nutrition survey. *British Journal of Nutrition*. https://doi.org/10.1017/S0007114508060777

Lally, P., van Jaarsveld, C. H. M., Potts, H. W. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology*, *40*(6), 998–1009. https://doi.org/10.1002/ejsp.674

Goldberg, L. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower level facets of several Five-Factor models. In *Personality psychology in Europe (Vol. 7)*.

Morean, M. E., DeMartini, K. S., Foster, D., Patock-Peckham, J., Garrison, K. A., Corlett, P. R., Krystal, J. H., Krishan-Sarin, S., & O’Malley, S. S. (2018). The Self-Report Habit Index: Assessing habitual marijuana, alcohol, e-cigarette, and cigarette use. *Drug and Alcohol Dependence*. https://doi.org/10.1016/j.drugalcdep.2018.01.014

Murre, J. M. J. (2014). S-shaped learning curves. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-013-0522-0

Murre, J. M. J., Chessa, A. G., & Meeter, M. (2013). A Mathematical Model of Forgetting and Amnesia. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00076

Neuberg, S. L., & Newsom, J. T. (1993). Personal Need for Structure: Individual Differences in the Desire for Simple Structure. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.65.1.113

Nocon, M., Hiemann, T., Müller-Riemenschneider, F., Thalau, F., Roll, S., & Willich, S. N. (2008). Association of physical activity with all-cause and cardiovascular mortality: A systematic review and meta-analysis. In *European Journal of Preventive Cardiology*. https://doi.org/10.1097/HJR.0b013e3282f55e09

Patton, J., Standord, M., & Barratt, E. (1995). Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, *51*(6), 768–774. https://doi.org/10.1037/t05661-000

Sheeran, P., & Webb, T. L. (2016). The Intention–Behavior Gap. *Social and Personality Psychology Compass*. https://doi.org/10.1111/spc3.12265

Spiess, A. N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach. *BMC Pharmacology*. https://doi.org/10.1186/1471-2210-10-6

Thompson, M., M., Naccarato, M.E., Parker, K. E. (1989). Personal Need for Structure Scale (PNS) [Database record]. *APA Psych Tests*. https://doi.org/https://doi.org/10.1037/t00912-000

Verhoeven, A., & de Wit, S. (2018). The Role of Habits in Maladaptive Behaviour and Therapeutic Interventions. In *The Psychology of Habit*. https://doi.org/10.1007/978-3-319-97529-0\_16

Verplanken, B. (2006). Beyond frequency: Habit as mental construct. *British Journal of Social Psychology*. https://doi.org/10.1348/014466605X49122

Verplanken, B., & Orbell, S. (2003). Reflections on past behavior: A self-report index of habit strength. *Journal of Applied Social Psychology*, *33*(6), 1313–1330. https://doi.org/10.1111/j.1559-1816.2003.tb01951.x

Wood, W., & Rünger, D. (2015). Psychology of Habit. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-122414-033417

Ypofanti, M., Zisi, V., Zourbanos, N., Mouchtouri, B., Tzanne, P., Theodorakis, Y., & Lyrakos, G. (2015). Psychometric properties of the International Personality Item Pool Big-Five personality questionnaire for the Greek population. *Health Psychology Research*. https://doi.org/10.4081/hpr.2015.2206

**Appendix A**

**Table A1**

*Study Design Table*

| Question | Hypothesis | Sampling plan | Analysis Plan | Rationale | Interpretation  | Theory  |
| --- | --- | --- | --- | --- | --- | --- |
| RQ1: How long does it take to form a habit? | H1: It takes a median of 66 days to form a habit. | We aim to test 200 participants at each of the four sites, until the data collection deadline (end of Dec ‘22). In agreement with the original paper, participants were excluded from analyses if: (i) they failed to provide data beyond day 60; (ii) SPSS was unable to find an optimal solution after 100 iterations attempting to fit the curve equation to the data; (iii) the model generated a zero value for the b parameter of the fitted equation y = a – be-cx, which implies no learning (a flat curve); (iv) the modelled asymptote score was below 21 (indicating a lack of habit) or higher than 49 (which is an unrealistic value); (v) if the R2 value was below 0.7. | The number of days needed to reach a plateau of automaticity (95% of asymptote) will be calculated for each participant, and the median value will be determined for each of the four datasets. Additionally, we will compare the 95% confidence intervals around the median estimates of the replicated datasets at the four sites within our consortium.  | Our (replication) analysis takes all individual curves together to determine the median number of repetitions to reach the plateau of automaticity (RQ1). To narrow down the 95% confidence interval for the median, and thereby obtain a more precise representation, we aim to increase the original sample size by factor 2. We estimate that this can be achieved by including 61% more participants, based on the fact that in the original study only 39% of the initially included participants (39 out of 101) could eventually be used in the main analyses. Therefore, we aim for a primary sample of 200 participants (per site). This sample size was based on exact replication of the original analysis.Importantly, this greater sample size also allows for more reliable results regarding the influence of individual differences (RQ 2/3). | We will use the majority rule to determine if the original finding of 66 days for habit formation is reliable and generalizable. If at least three out of four research sites yield 95% confidence intervals containing values between 56-76 we will conclude that the finding is replicated.Conversely, if more than one research site’s 95% confidence interval does not fall within 56-76 , this means that the finding is not replicated. |  N/A |
| RQ2: Do habits form faster when the behaviour is consistently performed? | H2: Missing a single opportunity to perform the behaviour (i.e., lower consistency) compromises habit formation. | Automaticity measurements preceding (X1) and following (X2) a single missed opportunity (i.e., an occasion where the behaviour was reported to not have been performed but preceded by three consecutive days of performing the behaviour) are compared using a Wilcoxon signed rank test. Additionally, we will also compare the automaticity scores preceding a single missed opportunity (X1) and the second day following this missed (X3) opportunity (only when automaticity scores were also available for X2). Finally, this difference in automaticity was compared to situations when the behaviour was performed on three consecutive days (i.e., without a miss in between), again with a Wilcoxon signed rank test. The error rate of these three tests will be corrected using the Bonferroni method. We will calculate the effect size for Wilcoxon signed rank test (r) using the formula z / sqrt (n). For hypothesis testing, we will construct Bonferroni-corrected 95% confidence intervals around the estimated effect sizes for each comparison using bootstrapping. This will be done using the rstatix package in R (​​Kassambara, 2021).  | An a priori power analysis for a Wilcoxon signed rank test was conducted using G\*Power version 3.1.9.7 to determine the minimum number of participants required to test our H2 hypothesis, Results indicated the required sample size to achieve 80% power for detecting a medium (0.50) effect, at a significance criterion of α = .05, was 28. Based on Lally’s observation that on average, for each participant 2.5 missed opportunities were found, this suggests 11 (medium effect) participants would be needed for the first analyses. However, based on Lally’s experience, we can expect to find fewer occasions where both X2 and X3 automaticity data will be available (i..e., 1.2 per participant on average). Hence, this suggests that for the second and third analysis, a minimum of 23 participants is required. Given that data of all sites will eventually be combined (N = 800), our sample size will suffice and even allow to detect an effect size as small as 0.09 (i.e., sensitivity analysis). To test the hypotheses, we will follow the approach specified by Dienes (2021) - by specifying a minimal interesting effect of interest for the theory. Therefore, not obtaining a significant effect will not be treated as evidence against the theory. We will use the smallest effect we are powered to detect (i.e., 0.09) as our minimal interesting effect size. The effect size will be converted to raw units once the data is collected and this will be included in the Stage 2 of our Registered Report. | If any of the lower bounds of the three confidence intervals around the effect sizes are above the minimal interesting effect (with higher automaticity in absence of an omission), this would indicate that performing the behaviour consistently is important for the automatization of a routine. Conversely, if the lower bounds of confidence intervals are below the minimal interesting effect, then this indicates that performing the behaviour consistently is irrelevant for automatization.  | James (1890) suggested that consistent performance is vital for habit formation see also, (Armitage, 2005). If omissions fail to affect automatization in our study, this will provide evidence against this theory. |
| RQ3: Does the complexity of a novel routine negatively affect its automatization? | H3: It takes more repetitions to automatize a complex behaviour (i.e., exercise as opposed to simple eating or drinking behaviour). | We will compare the estimated curve parameters (a, b, and time to reach 95% of asymptote) and performance variables (i.e., number of reported repetitions and percent compliance) between participants who chose eating, drinking, and exercise target behaviours. The former two are thought to be relatively simple, whereas the latter is considered a more complex behaviour. The difference between the behaviours will be compared by performing 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 looking at 5 dependent variables with the Bonferroni method. For the pairwise comparisons in each of the ANOVAs we will use the Tukey-Kramer correction. We will use Cohen’s D as a measure of effect size for the pairwise comparisons. The 95% confidence intervals around these effect size estimates can be calculated using a bootstrapping procedure in the ESCI library in R (Calin-Jageman, 2022).  | An a priori power analysis for a one-way ANOVA was conducted using G\*Power version 3.1.9.7 to determine the minimum number of participants required to test our H3 hypothesis. Results indicated the total required sample size to achieve 80% power for detecting a medium (0.25) effect, at a significance criterion of α = .05, was 159. Given that data of all sites will be combined (N = 800), our sample size will suffice and even allow to detect an effect size as small as 0.12 (for each pairwise comparison). Similarly to H2, we will use the smallest effect we are powered to detect as the minimal interesting effect size (i.e., 0.12). The effect size will be converted to raw units once the data is collected and this will be included in the Stage 2 of our Registered Report. | If automatization is slower and reaches a lower level for complex behaviours than for simpler ones, this suggests that complex behaviours are harder to automatize. If this is not the case, then complexity is irrelevant for automatization. Specifically, if any of the lower bounds of the five confidence intervals around the estimated effect sizes for comparisons between simple and complex behaviours are above the minimal interesting effect, this would indicate that complexity of the routing is important for its automatization. Conversely, if the lower bounds of confidence intervals are below the minimal interesting effect, then this would indicate that complexity of the routing is irrelevant for its automatization. | Complex behaviours have been proposed to automatize more slowly than simple behaviours (Verplanken, 2006; Kaushal et al., 2017). If automatization is equally fast for relatively complex exercise behaviours and simpler eating and drinking behaviours, this would provide evidence against this theory. |

*Note.* This table summarises the main research questions (RQ) and hypotheses (H) of our replication, as well as our sampling plan, analysis plan, rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis, interpretation of different outcomes, and theory that could be proven wrong by the outcomes.