The Shape of Habits: 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 was 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, with the first behavioural repetitions leading to greater increases in automaticity than later repetitions. Furthermore, their study suggests 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. Interestingly, 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 considered 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 conducted a near-exact, multi-centre replication of the study of Lally and colleagues (2010). We adopted 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 applied 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 allowed us to determine whether the relationship between behavioural repetition and subjective automaticity is modelled best by an asymptotic curve (see H1; Table A1), and whether it indeed takes a (median) of 66 days to form a habit (see H2; Table A1).

Next to modelling habit formation, we also explored 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. In our replication study, we determined whether missing a single opportunity to perform the behaviour (i.e., lower consistency) compromised habit formation

(Armitage, 2005; see H3, Table A1), and investigated 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 H4, Table A1)(Kaushal et al., 2017; Verplanken, 2006). Furthermore, we investigated whether habit formation is affected by relevant personality factors, namely: conscientiousness (Goldberg, 1999), impulsivity (Patton et al., 1995), and personal need for structure (Thompson et al., 1989).

Methods and Analyses

Participants

The replication project was 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 planned to test 200 participants at each of the four sites (between 21-45 years). They were recruited via the universities' websites, crowdsourcing software, and social media. Each participant was paid €50 for their participation if they completed the study. If they dropped out in between the first and second meeting, or if they made less than four data entries during these first four weeks, data collection was terminated and they received 20 euros. In the case of dropout, we tested additional participants (as far as our replication budget of 800*50 euros allowed, and until the data collection deadline (23-12-2023). Each site aimed to include roughly 50% university students (and therefore a comparable sample to the original study) and 50% from a non-student population.

This study was 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 $\frac{n}{2}$ - $\frac{1.96\sqrt{n}}{2}$ th ranked value.
2. The upper 95% confidence limit is given by the 1+ $\frac{n}{2}$ + $\frac{1.96\sqrt{n}}{2}$ 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 aimed for a primary sample of 200 participants (per site). This sample size was based on exact replication of the original analysis. As our additional analysis allowed a greater number of participants to be included, this was even more powerful. Importantly, this greater sample size also allowed for more reliable results regarding the influence of individual differences.

Procedure

For the purpose of replication, we adopted the original study protocol (as illustrated in Figure 1). During this 12-week study, participants met with the experimenter three times via video conferencing. Prior to the first meeting, they provided 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.

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 had chosen a target behaviour, participants were asked how often they performed this behaviour over the past four weeks. In the present study, if they had performed it four times weekly or even more frequently, then they were asked to choose a different target behaviour. Finally, they were asked to rate anticipated effort of performing this behaviour every day and to rate their intention (see Materials for details), and they scheduled meeting 2 and 3.

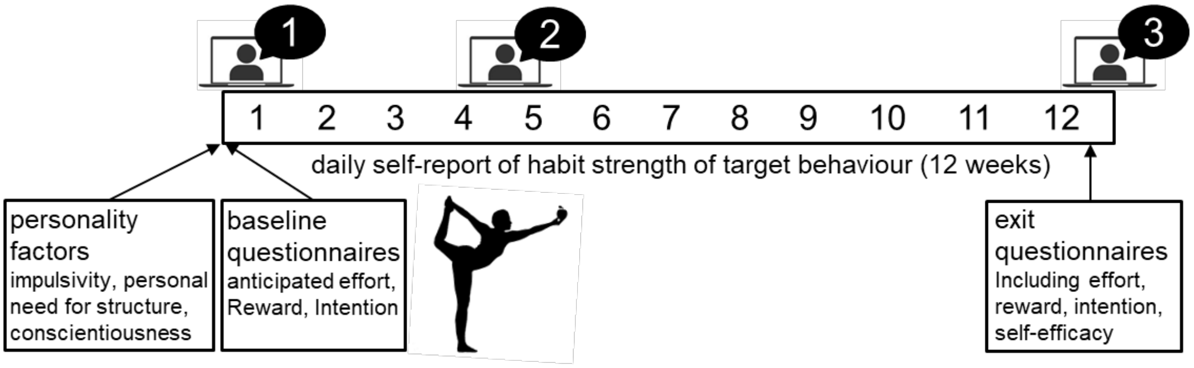
Participants were instructed to start performing the behaviour on the day after the first meeting. The first email with a link to the automaticity questionnaire was sent two days after the first meeting. It was always sent at 8.00 AM. This questionnaire asked them to rate the (current) automaticity of the behaviour using the SRHI (e.g., *'doing squats when I wake up is something I do automatically')*, and whether they performed the behaviour the previous day. If they failed to fill out the daily questionnaire, they could report whether they had performed the behaviour retrospectively for up to three previous days. Finally, if they reported that they did not perform the behaviour, they were 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 was scheduled 4-5 weeks after the first one. The main purpose of this meeting was to boost compliance. Furthermore, we asked participants whether their daily routines were disrupted by external circumstances during five consecutive days or longer (e.g., illness or holidays). If they answered in the affirmative, they were additionally 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 were emailed a link to the exit questionnaire, which they should complete before the third and final at the end of the project. After they answered the questions listed in the Materials section, we asked 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. During the final meeting, participants were thanked and reimbursed for their participation.

Figure 1

Study Procedure


1
2
3
1 2 3 4 5 6 7 8 9 10 11 12
daily self-report of habit strength of target behaviour (12 weeks)
personality factors
impulsivity, personal need for structure, conscientiousness
baseline questionnaires
anticipated effort, Reward, Intention
exit questionnaires
Including effort, reward, intention, self-efficacy

Note. At the start of the study, participants filled out questionnaires about personality factors. During the first online meeting with the experimenter, they chose an exercise / drinking / eating habit that they wanted to form during the next 12 weeks. They also reported on the anticipated effort, enjoyment (reward) and strength of their intention. Next, they were 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 had the second meeting with the experimenter and were encouraged to continue. After twelve weeks, they filled out the exit questionnaires. These questionnaires contained 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 had their third meeting with the experimenter, and were thanked and reimbursed for their participation.

Our replication was aimed to be near-exact. The main changes that we made to the original protocol are summarised in this paragraph. First of all, we conducted the three meetings online via video conferencing. Furthermore, the daily email to participants was sent automatically through the LOTUS software (www.lab.uva.nl/lotus), and the questionnaires were filled out online via Qualtrics (www.qualitrics.com). Due to technological advances since the original study, there was another difference that we could not circumvent. These days, some people receive a push message on their smartphone when receiving an email. Being reminded to complete the Self-Report Habit Index (SRHI; Verplanken & Orbell, 2003) by push-messages might affect adherence (even though these were sent early in the morning). Therefore, at the end of the study, we asked 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 were used during the screening, lab sessions, and daily habit formation were all administered using Qualtrics (www.qualtrics.com), and daily reminders were 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 assessed 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 ranged from 0-6, and, therefore, the maximal total score was 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 asked 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 added four questions to the original protocol, informed by the Theory of Planned Behaviour (Ajzen, 1991). These were 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 were: "*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 added a question about anticipated reward: "*I enjoy [target behaviour X] everyday*".

Exit questionnaire. Upon completing the 84th day of performing the selected target behaviour, participants received a final email, containing the link to a Qualtrics questionnaire. This questionnaire included the effort and intention questions that were also posed during the first meeting. Participants rated 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 posed 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) wether 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 pertained to the disruption of their daily routines (for an overview; see (https://osf.io/apwd3/).

(iii) whether they changed their original planned behaviour and/or cue during the study (and if yes, to state their new plan); (iv) and finally, we repeated the questions during meeting 2 pertaining to a disruption of their daily routines.

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 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 R^2 value was 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 R^2 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. Additionally, we will report the BIC and AIC measures of goodness-of-fit, which are nowadays considered superior to R^2 and can be of aid to the reader to assess our results.

The curve parameters will also be related to two performance variables: the number of reported repetitions (the sum of all occasions when a participant reported having performed the target behaviour) and compliance percentage (i.e., the percentage of all days for which data were reported that the participant reported having performed the behaviour).

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 analysis that circumvents unnecessary rejection of data and addresses weaknesses in the original approach in four different ways. 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 R^2 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 (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.

Comparing datasets from the four centres. In the next step, we will compare the days-distribution (histogram) of the replicated data-sets at the four sites within our consortium with each other and with the original data using Kolmogorov-Smirnov tests. If sets are found to not differ significantly, this means that the findings are highly reliable and generalizable. If sets do differ, this means that the finding of 66 days for habit formation is not generalizable. In this case, our first step will be to repeat the analyses with just the student subsamples to investigate whether this is the cause of divergent findings (as the original study tested a student sample). If we still find a significant difference, this means that future research is required to reveal the underlying cause(s), e.g.: different languages/cultures.

Behavioural complexity, consistency, and individual differences. For even more powerful analyses of potential sources in intra- en interindividual variability and to further constrain the confidence interval, we combined our data sets to investigate (as in the original study) the effects of consistency (see RQ3; Table A1) and complexity (i.e., behaviour type: drinking/eating/exercising; see RQ4; Table A1) on automaticity development. Complexity is investigated by comparing the estimated curve parameters and performance variables (i.e., number of reported repetitions and percent compliance) between relatively simple eating and drinking behaviours and more complex exercise routines using Kruskal-Wallis comparisons. The effect of consistency was investigated by comparing automaticity immediately preceding and following a single missed opportunity or 'omissions' (defined by Lally et al. as an occasion where the behaviour was reported to not have been performed but was immediately 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.

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.

Effort, intention, and reward. In addition to addressing the research questions of the original study, we also performed multiple regression analyses to determine whether anticipated effort, reward, intention, attitude, subjective norm and perceived behavioural control are correlated with the performance variables and curve parameters.

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

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

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

Hull, C. L. (1943). Principles of behavior: an introduction to behavior theory. In Principles of behavior: an introduction to behavior theory. Appleton-Century.

Hull, C. L. (1951). Essentials of behavior. In Essentials of behavior. Yale University Press.

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

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 R^2 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: Is the process of automatization of a novel routine best described by a linear or asymptotic curve?	H1: The relationship between behavioural repetition and subjective automaticity is modelled best by an asymptotic curve.	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.	A linear and asymptotic regression will be run on the individual daily Self-Report Habit Index automaticity composite scores. The R2 measure of goodness-of-fit will be used to determine whether the relationship between daily repetition and self-reported automaticity is modelled best by the linear or asymptotic curve. A Sign test will be used to determine whether this difference is significant. First, these analyses are conducted separately for each of the four datasets.	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 and RQ4).	If the asymptotic curve has a significantly higher goodness-of-fit for each of the four datasets, we consider it a reliable and generalizable finding that habit formation is best described by an asymptotic curve. A different outcome would mean that habit formation is not best described by an asymptotic curve.	The authors of the original study based their hypothesis on the theory of Hull (1943, 1951), according to which habit strength increases according to: *Habit = a (1 – 10-bN).* If we fail to replicate the original finding that habit formation is best described by an asymptotic curve, this would provide evidence against the theory of Hull (1943, 1951).
RQ2: How long does it take to form a habit?	H2: It takes a median of 66 days to form a habit.		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. We will compare the days-distribution (histogram) of the replicated datasets at the four sites within our consortium with each other and with the original data using Kolmogorov-Smirnov tests.		If sets (including the dataset of the original study) are found to not differ significantly in the individual numbers of days needed to reach a plateau of automaticity, this means that the original finding (of 66 days for habit formation) is highly reliable and generalizable. If sets do differ, this means that this is not the case. In this case, our first step will be to repeat the analyses with just the student subsamples to investigate whether this is the cause of divergent findings (as the original study tested a student sample). If we still find a significant difference, this means that future research is required to	N/A

Question	Hypothesis	Sampling plan	Analysis Plan	Rationale	Interpretation	Theory
					reveal the underlying cause(s), e.g.: different languages/cultures.	
RQ3: Do habits form faster when the behaviour is consistently performed?	H3: 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.	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 H3 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 less 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).	A significant difference in automaticity between measurements following an omission and measurements in absence of an omission (with higher automaticity in absence of an omission) indicate that performing the behaviour is important for the automatization of a routine.	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.
RQ4: Does the complexity of a novel routine negatively affect its automatization?	H4: It takes more repetitions to automatize a complex behaviour (i.e., exercise as opposed to		We will compare the estimated curve parameters and performance variables (i.e., number of reported	An a priori power analysis for an one-way ANOVA (parametric variant of Kruskall-Wallis (was conducted using G*Power	If automatization is slower and reaches a lower level for complex behaviours than for simpler ones, this suggests that complex	Complex behaviours have been proposed to automatize more slowly than simple behaviours (Verplanken, 2006; Kaushal

Question	Hypothesis	Sampling plan	Analysis Plan	Rationale	Interpretation	Theory
	simple eating or drinking behaviour).		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 is assessed via a parametric one-way ANOVA and non-parametric Kruskall-Wallis comparisons.	version 3.1.9.7 to determine the minimum number of participants required to test our H4 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. Next, these values were corrected (multiplied by 1.15) for non-parametric testing (Kruskall-Wallis): 183 (Lehman & D'Abrera, 1998). This indicated that per behaviour 61 participants would be needed. 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 0.19 (i.e., sensitivity analysis).	behaviours are harder to automatize. If this is not the case, then complexity is irrelevant for automatization.	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.