PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

What can machine learning teach us about habit formation? Evidence from exercise and hygiene

Anastasia Buyalskaya[a,1], Hung Ho[b,1], Katherine L. Milkman[c], Xiaomin Li[d], Angela L. Duckworth[c,e], and Colin Camerer[d,f,2]

Edited by Elke Weber, Princeton University, Princeton, NJ; received September 24, 2022; accepted January 27, 2023

We apply a machine learning technique to characterize habit formation in two large panel data sets with objective measures of 1) gym attendance (over 12 million observations) and 2) hospital handwashing (over 40 million observations). Our Predicting Context Sensitivity (PCS) approach identifies context variables that best predict behavior for each individual. This approach also creates a time series of overall predictability for each individual. These time series predictability values are used to trace a habit formation curve for each individual, operationalizing the time of habit formation as the asymptotic limit of when behavior becomes highly predictable. Contrary to the popular belief in a "magic number" of days to develop a habit, we find that it typically takes months to form the habit of going to the gym but weeks to develop the habit of handwashing in the hospital. Furthermore, we find that gymgoers who are more predictable are less responsive to an intervention designed to promote more gym attendance, consistent with past experiments showing that habit formation generates insensitivity to reward devaluation.

habit | machine learning | context sensitivity | predictability | nudge

Significance

Psychology and neuroscience define habit as a type of decision process in which the same behavior is executed in the same context, regardless of the outcome. We introduce a machine learning method that discovers which of many context variables are associated with behavior and infers how quickly habits form. This can include but is not limited to past choice. The method can be used to target interventions based on the specific context variables that affect behavior, that individual's predictability and how strongly their habit has been formed. This potential is illustrated by our finding that a well-powered, random-assignment intervention to increase gym attendance had a larger effect on gymgoers who were less predictably context-sensitive.

Much of human behavior is habitual. Unlike choices that are consciously deliberated, habits are the result of stimulus–response associations (1). Habits constitute a behavior in which responses are cued by context features (e.g., sensory stimuli, locations, preceding actions) that were reliably present when the habit was previously executed (2).

To date, the best evidence of automatic and context-sensitive behavior linked to habits comes from laboratory experiments on animals and humans (*SI Appendix*, section 1) and a handful of field experiments (3), But there is surprisingly little research on how human habits naturally develop outside of the laboratory over the course of weeks or months in everyday life (4). Three prior observational studies examined habit formation over substantial periods of time using daily self-reports. In a seminal study (5), 96 undergraduates ate, drank, and exercised daily in the same context for 12 wk and self-reported habit strength every day. This study, and two similar ones (6) and (7), suggest that

> "habits typically develop asymptotically and idiosyncratically, potentially differing in rate across people, cues and behaviors" (8, pg. 220).

To advance our understanding of how habits develop in natural settings, we develop a machine learning methodology that is especially well suited for analyzing panel data with repeated observations of behavior. Our Predicting Context Sensitivity (PCS) approach identifies the context variables that best predict behavior for each individual. Specifically, PCS uses a least absolute shrinkage and selection operator (LASSO) regression, a hypothesis-free form of statistical analysis which does not prespecify what variables are likely to be predictive of an individual's behavior. The LASSO technique generates a person-specific measure of overall behavioral predictability based on the variables that are predictive of that person's behavior. Predictability ranges from .5 (completely random) to 1 (completely predictable), acknowledging that habit formation is a matter of degree and is not dichotomous. The degree of habit formation can vary across people and vary within person across time. These continuous measures of predictability we generate are then used to study individual differences in predictability, and speed of habit formation, which is defined as the amount of time it takes for a given person's behavior to reach its steady state of predictability.

Because the field data we analyze do not include direct measures of automaticity of behavior, we operationalize habit as context-sensitive predictability. It is, of course, possible that other mechanisms that are not automatic account for the context-sensitive

Author contributions: A.B., H.H., and C.C. designed research; A.B., H.H., K.L.M., A.L.D., and C.C. performed research; A.B., H.H., X.L., and C.C. analyzed data; and A.B., H.H., K.L.M., A.L.D., and C.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]A.B. and H.H. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: camerer@caltech.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2216115120/-/DCSupplemental.

Published April 17, 2023.

predictability we observe (9). However, we note that our approach is consistent with a large literature operationalizing habit as we do, e.g., ref. 10.

PCS works best for large samples in which there are repeated observations of behavior for each person. For each of the 30,110 people in study 1, there are a median of 1,525 daily observations (over 4 y of gym attendance). For each of the 3,124 people in Study 2, there are a median of 3,000 observations (98 hospital shifts). In both samples, we analyze objective measures of behavior, rather than self-reports of behavior gathered after-the-fact, thereby avoiding possible errors of memory and metacognition (11).

PCS yields three important discoveries in this investigation: First, context variables are more predictive of behavior for some individuals than others. Second, contrary to common wisdom, there is no "magic number" for how long it takes to form a habit. Instead, the speed of habit formation appears to vary significantly between behavioral domain: Gym habits take months to form and handwashing habits take weeks to form. Third, consistent with prior research on nonhuman animals, more habitual gymgoers are reward-insensitive, responding less to a well-designed behavioral intervention (12).

1. Study 1: Gym Attendance

A. Data. We partnered with 24 h Fitness, a large North American gym chain, to obtain check-in and background data on a total of 60,277 regular gym users (IRB approval is described in *SI Appendix*, section 6). These users were spread across 560 gyms and consented to share their information with researchers when they signed up to be in a fitness program. The data we analyze track gym attendance for each person from the first day they attended this gym chain, which is ideal for studying the development of habits from inception. Our gym attendance dataset spans 14 y, ranging from 2006 to 2019 and it includes over 12 million data points, each corresponding to one gym check-in. Each data point has a timestamp and location of the gym visited, as well as other details about the gym (such as its amenities). We infer several other attributes not in the raw data files, such as the day of the week, and individual-level variables such as the time since gym membership creation. The total set of unique candidate context variables used to estimate predictable gym attendance includes a gym visit's month of the year, day of the week, time lag (the number of days which have elapsed since the gym goer's last visit), attendance rate in the past week, the number of consecutive days of attendance (streak), and the number of consecutive days of attendance for the same day of the week (day-of-week streak). A full list of the variables analyzed and a longer description of the data can be found in *SI Appendix*, section 2.

Our analytic sample is a subset of gym goers based on two criteria. Members without a valid gym contract (1,083) are removed. Second, we exclude participants with less than a year of data (removing 28,878 members) and too little attendance for the LASSO model to classify well due to sample imbalance (removing 206 members). This leaves N = 30,110 gym goers. We analyze the behavior of these gym goers from the first day that they joined the gym.

Table 1 provides summary statistics for the final analytic dataset. Gym members in our sample are 62% female and have a median age of 34 y. The average individual in this dataset goes to the gym every 4 to 6 d. The median number of days an individual is observed (or "has an opportunity to go to the gym") is 1,525 d, or just over 4 y.

Table 1. Summary statistics from gym attendance analytic sample

	Mean	SD	Q1	Median	Q3
Age	36.76	12.35	27.00	34.00	45.00
Female	0.62	–	–	–	–
Daily attendance	0.19	0.16	0.07	0.14	0.27
Days observed	2,020	1,453	658	1,525	3,655
Days between visits	15.77	29.74	3.69	6.89	15.22

Note: SD = standard deviation, Q1 and Q3 = first and third quartiles.

B. Analytic Approach. For each individual in our dataset, we first train a logit LASSO model to predict the likelihood of gym attendance (a binary outcome variable) day-by-day. LASSO adds a penalty to the minimized objective function (sum of squared residuals) based on coefficient absolute magnitude. This has the effect of "shrinking" many regression coefficients to zero in order to guard against overfitting when so many variables are included. To illustrate a typical difference between regular regression and LASSO, suppose a researcher believed month-of-the-year might be predictive of gym attendance but didn't have a strong hypothesis regarding which months were most predictive for which individual. A standard OLS regression would derive best-fitting nonzero coefficients for all the 12-mo variables. But it is likely that most of those coefficients are not precisely estimated, and the largest magnitudes would probably be exaggerated. Because of the LASSO penalty, many of these coefficients that are close to zero would shrink to exactly zero and the largest magnitudes would be compressed toward zero (both "zeroing out" variables and compressing them toward zero reduces the LASSO penalty). The LASSO penalty is well known in many applications to reduce false positives—which are expected when there are many coefficients—so that false positives do not degrade the fit from training to the hold-out test samples. This results in a more compact model with only the most relevant predictor variables having nonzero coefficients. For example, in our gym data analysis, about half of the month coefficients have zero LASSO coefficients (*SI Appendix*, Table S1).

It is crucial to note that LASSO shrinkage due to the magnitude penalty implies that the coefficients are not the best estimates of the true coefficient value (because they are deliberately shrunk toward zero). This is a well-known property of LASSO. It is the statistical price that is paid to guard against overfitting when exploring large variable sets (in technical terms, it reflects a "bias-variance" tradeoff). A byproduct of this property is that there is no simple conventional procedure for computing standard errors for LASSO coefficients (13) for ideas. We therefore do not say anything about the statistical significance of nonzero effects or differences in coefficients between people.

There are two major consequences of the LASSO coefficient bias toward zero. One is called "model selection consistency"—are the variables selected by LASSO those with true nonzero coefficients? The second is called "stability": When two variables are highly correlated, LASSO will often set one predictive coefficient to zero and let the other variable do the predictive work of the two variables together. Exploratory analysis seems to indicate that they are not creating a problem for our analysis, particularly since we are not trying to infer true coefficient values—we are just trying to find the most predictive coefficients. More precisely, we do not have standard errors around coefficients and therefore cannot say for how many individuals any given predictor is significant. *SI Appendix*, section 3.2 has a thorough discussion

of these issues, and the procedures we carried out to examine this problem. While we do not claim to fully resolve the problem of model selection in our setting, we hope these procedures serve as a reassurance about the directional effects of the most important predictors.

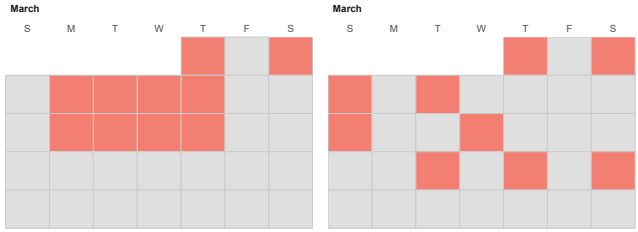
Following machine learning standards, we use approximate six-fold cross-validation, training the model on 85% (about 5/6ths) of the full-time series data for each individual (*SI Appendix*, section 3.1). We use the remaining 15% of the data (1/6th) as our "test" set. This shows how good the LASSO model is at predicting an individual's attendance on days the model did not observe. This gives us an out-of-sample test-set predictability measure, called the area under the curve (AUC), for each gym member. This serves as an objective full-sample measure of context-sensitive predictability for each gym member (i.e., how habitual an individual's behavior might be) and allows us to avoid possible errors induced by self-report measures, which rely on an individual's memories of the context cues present when they last executed a habitual behavior. This machine learning technique allows us to treat overall individual-level habit as a number from .5 to 1.

Next, we determine when, if ever, behavior becomes more habitual over time. Following (5), for each individual i, we attempt to identify $A(t)$, an exponential function of form $a - be^{-ct}$ describing daily-level habit strength as a function of time. Likewise, following (5), we define the inferred time to habit formation as the time it takes for $A(t)$ to reach 95% of its asymptote. In other words, we define "time to habit formation" as the amount of time it takes for AUC, behavioral predictability, to reach close to its steady-state value (this is seen visually as AUC increasing and eventually plateauing after a certain point in time). This is consistent with previous studies of habit formation, which find that initial behavioral repetitions cause increases in responding, with each marginal repetition delivering lower gains in until the behavior reaches its limit (14).

To infer the increase in habit strength over time, we use a different procedure than deriving AUC from full samples for each individual. In this procedure, we estimate a series of time-specific AUCs denoted $A(t)$ at period t. This procedure did not work reliably at the individual level, so we first split the gym goers into ten deciles. Deciles were based on the number of observations (sample length) for each person. The shortest-length samples are grouped into decile 1 and longest-length samples are grouped into decile 10. The purpose of this split is to ensure that individuals in a decile group are comparable in terms of their number of data points because LASSO results (as with all statistical techniques) can be sensitive to sample features such as the amount of data per individual.

For each decile of gym goers, we compute the AUC values obtained when training a LASSO model using data from periods $[t-2, t-1]$ and testing (validating) on data at period t, where each period t consists of 2 wk of data ($t = 0$ denotes the first 2 wk). The procedure starts with $t = 2$, where we train a LASSO model using everyone's first 4 wk of data (corresponding to $t = 0$ and $t = 1$), and then it computes the AUC when evaluating that model in 'test' holdout data from 5 to 6 wk. It then proceeds iteratively for increasing values of t until it reaches the end of the decile's observed time period. This procedure creates a sequence of sliding windows of equal size (4 wk of training data and two weeks of test data).

C. Results. In the LASSO training data for $N = 30,110$ gym members, the mean individual-level AUC is 0.806 (median is 0.811, interquartile range 0.750 to 0.868), where 0.5 is random and 1.0 is perfectly predictable. This indicates that the LASSO models tend to do a good job fitting the gym goers' attendance behavior. On the test datasets, these measures are slightly lower (as is expected), with a mean individual-level AUC of 0.768 (median is 0.778, and interquartile range 0.702 to 0.845).



(a) A highly predictable gymgoer. AUC = 0.964, 10 days of attendance

(b) A highly unpredictable gymgoer. AUC = 0.546, 9 days of attendance

Fig. 1. Attendance patterns during March 2018 of two individuals with comparable gym attendance rates but different AUC values. Red squares indicate attendance.

Fig. 1 *A* and *B* illustrate two gym members' monthly attendance calendars from March 2018. While the two members go to the gym about equally often in March, one is highly predictable (AUC = 0.946) and the other is not at all predictable (AUC = 0.546). This is an example that shows frequency and predictability are uncorrelated in our data (*SI Appendix*, section 3.3) Of course, including missing variables we do not have access to (e.g., someone's work schedule) could increase the predictability of the person in Fig. 1*B*, and it would also be likely to reduce the differences in predictability across people.

As shown in Table 2, the most important predictor of gym attendance across individuals is how much time has passed since the previous gym visit ("time lag"). This predictor appears almost universally important and goes in the same direction across most individuals: for 76% of gym goers in our sample, the longer it has been since they last visited the gym, the less likely they are to go on a given day. Consistent with the examples in Fig. 1, day-of-week streak is an important predictor, with 69% of the sample more likely to go to the gym on the same days of the week that they had previously attended. Among the days of the week, Monday and Tuesday are the most important, positively predicting attendance for 57% of both samples (*SI Appendix*, section 3.1), consistent with other evidence of focal "fresh start" effects (15). Since we do not have standard errors as mentioned previously, the summary above only provides suggestive evidence about the directional effects of these predictors.

Fig. 2 provides two examples of results from fitting the exponential curves to the lowest and highest decile of gym goers' AUC sequences (i.e., those with the least and most data per person). The median estimated time it takes to reach the 95% asymptote across all gym goers well fit by the exponential model is 122 to 226 d or about 4 to 7 mo (*SI Appendix*, Table S8). Model fit is not related to average frequency of gym attendance nor the age of a gym member.

D. Additional Analyses of Reward Sensitivity and Individual Differences. Insensitivity to reward change is a gold standard hallmark of strong habits in animal research. But such insensitivity has proven difficult to identify in humans, e.g., refs. 16 and 17. It has not been studied in nonexperimental large-scale field data like ours. Because we are testing for insensitivity in field data in different ways, we openly explore multiple types of reward change, and how they might affect habit measured in different ways. Specifically, we explore the effects of two methods

Table 2. Variables which LASSO identified as being most predictive across gymgoers

	Importance	Q1	Median	Q3	% zero	% positive	Predictive effect
Time lag	1.25	−1.40	−0.34	−0.02	22	2	Negative
Monday	0.36	0.00	0.11	0.50	32	57	Positive
Tuesday	0.35	0.00	0.10	0.49	33	56	Positive
Attendance last 7 d	0.34	0.09	0.29	0.47	9	82	Positive
Day-of-week streak	0.23	0.00	0.11	0.30	25	69	Positive

of reward change. The first method is to hypothesize possible changes in reward, without having a direct measure (of the kind carefully controlled in lab experiments). For gym attendance, the hypothesized reward variables are unusually good or bad weather. The second method is to use randomized interventions designed to change subjective reward to promote more gym attendance (12).

The effect of reward changes is evaluated with respect to low- and high-habit behavior across two methods—one is within-participant, and the second is between participants. The within-participant approach uses the fact that our method for estimating habit formation time divides each individual's behavior into "pre-" and "post-" habit formation periods. We test the joint hypothesis that unusually good or bad weather changes the reward value of going to the gym and has a weaker effect in the "post-habit period". This test shows no reduction in sensitivity to unusual weather—the proxy for reward change—in post-habit periods (*SI Appendix*, section 4.1).

The between-participant test uses the AUC measure of overall predictability to divide the sample into people who are low or high in context-sensitive predictability, and it tests whether responses to an experimental StepUp intervention done in partnership with 24 h Fitness are different for the two groups (*SI Appendix*, section 4.2). It is worth noting that a gymgoer who is highly predictable is not necessarily a more predictable, or habitual, gymgoer. Instead, it may be that two gym goers are equally predictable given a larger set of variables than we have. Omitting variables means that one gymgoer might be highly predictable and the other not very predictable given our feature set. This test of reward devaluation sensitivity allows us to test whether this data limitation is likely to be driving our results. If predictability does correlate positively with how habitual a gym goer is, we would expect to see more predictable gym goers responding less to reward devaluation.

Our results show that this is indeed the case. Less predictable gymgoers are significantly more likely ($P < .001$) to increase attendance in response to the StepUp intervention (*SI Appendix*, section 4.2), which means that highly predictable gymgoers respond less to the intervention. This finding provides evidence against the view that our predictability measure is largely driven by data limitations and is consistent with the experimental literature associating insensitivity to reward changes with stronger habit—measured here as predictability.

Finally, we take advantage of the size and diversity of the gymgoer sample to explore whether demographic and SES characteristics are correlated with predictability (an analysis that was pregistered on AsPredicted.org, #59014). To do this, we link our individual-level AUC predictability measures with Census data using each individual's home zip code and self-reported age and gender. We remove 2,447 people for whom we did not receive age or gender information from the gym, or whose zip code did not have data available. We then regress the AUC of the remaining sample (27,663 gym goers) on demographic characteristics. Regression results, which can be found in *SI Appendix*, section 5.4, confirm that demographic attributes are indeed predictors of AUC or "stronger habits", although most of the effects are small in magnitude. Specifically, older individuals living in more rural (low population density) areas, where a large fraction of married couples have children, have higher AUCs. Younger individuals living in more urban (high population density) areas have lower AUCs.

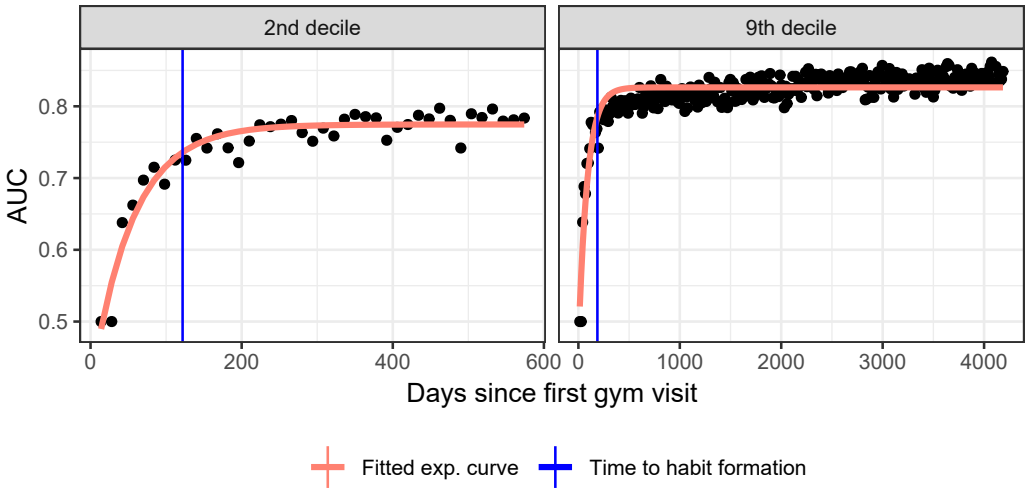


Fig. 2. Estimation of the speed of habit formation for gym attendance for the second lowest (*Left*) and second highest (*Right*) deciles by sample size. Note that the x-axes are different because, by construction, the deciles have different lengths of sampled time. The times to habit formation in these two deciles (shown by where the vertical blue line intersects the x-axis) are 121 (*Left*) and 187 (*Right*) days.

2. Study 2: Hand Washing among Hospital Workers

A. Data. We obtain hand-hygiene data from a company that uses RFID technology to monitor whether individual healthcare providers wash their hands at every opportunity to do so throughout their hospital shifts Proventix, following (18) and (19). The initial dataset tracks 5,246 healthcare workers across 30 hospitals. The dataset spans about a year, with over 40 million data points, each corresponding to whether an individual caregiver did or did not wash their hands in the face of an opportunity to do so. An opportunity is defined as a point in time when a caregiver either entered or exited a patient's room with a Proventix sanitizer present; so each room visit presents two sanitizing opportunities.

Each data point has a timestamp, as well as deidentified hospital and room locations. We further infer several other attributes about each opportunity to wash, such as the day of the week when it arose and whether the healthcare worker in question had complied with handwashing guidelines (i.e., washed their hands) in this room previously. Our unique candidate context cue variables include the time of day, time spent working, previous room and shift compliance, and indicators for whether the hospital worker is entering or exiting the room. A full list of the variables used and a longer description of the data can be found in *SI Appendix*, section 2.

Unlike the gymgoers, in this sample we do not observe hospital workers from the start of their work; we only observe their behavior after the RFID machines are installed. It is therefore likely that some degree of habit formation had already occurred, and we are not observing pure habit formation from inception. Thus, we treat the introduction of Proventix's RFID surveillance technology as a shock that may have disrupted behavior sufficiently to create somewhat new context-sensitive habits (e.g., a habit of washing when monitored). This possibility follows from the fact that hand sanitizing behavior increased by over 50% after machines were installed (19). However, we do acknowledge that handwashing may have habituated in many caregivers before the time at which we first observe their behavior in our sample.

We use two criteria to identify our final analytic sample. First, we remove any hospital workers who had fewer than 30 shifts included in our data (removing 2,115 hospital workers). Second, we remove seven workers without enough hand washing compliance data for the LASSO model to fit variability. This gives us an analytical sample of 3,124 hospital workers. Table 3 provides summary statistics about the workers in our analytic sample. The mean compliance with handwashing is 0.45 per opportunity. An average of 116 shifts are recorded per healthcare worker, and there are an average of 26 episodes (or visits to patient rooms, each with two opportunities to wash—one upon entry and one upon exit) per shift. We observe an average of 3,016 episodes per worker.

Table 3. Summary statistics from handwashing analytic sample

	Mean	SD	Q1	Median	Q3
Compliance	0.45	0.23	0.26	0.43	0.63
Total shifts	116	77	56	98	153
Rooms visited	37	33	20	29	41
Episodes (mins)	5.66	2.61	3.94	5.13	6.78
Episodes/shift	25.72	16.49	13.95	24.2	34.54
Shift length (mins)	512	214	408	581	646
Mins btwn episodes	22.42	11.5	13.95	20.12	29.00
Hrs between shifts	91.95	57.91	60.06	72.61	102.91

Note: SD = standard deviation, Q1 and Q3 = first and third quartiles.

B. Analytic Approach. We use the same machine learning approach as described in Study 1, training a LASSO model to obtain person-specific sets of coefficients and predictability measurements (AUCs) predicting when caregivers wash their hands. As in study 1, we inferred time to habit formation using the same approach of fitting an exponential curve to a sequence of AUC values for each decile of caregivers. Each period consists of two shifts rather than two weeks of gym attendance data.

C. Results. The LASSO model does a satisfactory job fitting hospital caregivers' hand-washing behavior. In the training dataset, the mean (median) individual-level AUC is 0.788 (0.783), and the interquartile range is 0.742 to 0.828. In the test dataset, these measures are only slightly lower, indicating the training overfitting is not severe the mean (median) individual-level AUC is 0.781 (0.776), and the interquartile range is 0.732 to 0.825. While our LASSO models have slightly less predictive power in this domain (compared to gym attendance), they still vastly outperform random chance at predicting hospital caregivers' hand-washing behavior.

As in study 1, the AUC measure—which can be used in any behavioral domain—is produced for each individual, and it once again serves as an objective measure of context-sensitive predictability. Furthermore, PCS again narrows down the set of context variables that are the most important predictors of hand washing at the aggregate level (Table 4). The most important and homogenous predictors of washing were a hospital worker's handwashing compliance during their last shift (a positive predictor for 100% of the hospital workers), room entry (which is a negative predictor for 77% of hospital workers, indicating most are more likely to wash their hands upon exiting, rather than entering, a room), and the room compliance of others (a positive predictor for 66% of hospital workers).

Times of day intervals were not selected by the LASSO model as predictive of most people's hand-washing behavior. However, consistent with previous research (18), the amount of time since the start of a caregiver's shift is a negative predictor of hand washing for 42% of caregivers. Again, since we do not have standard errors around these coefficient estimates, the summary above only provides suggestive evidence about the directional effects of these predictors.

As in study 1, we fit an exponential model to each decile's AUC sequences from early to later intervals. This allows us to analyze the development of predictability over time (Fig. 3). For all deciles, the median time to habit formation was on the order of a couple of weeks (habits typically formed after 9 to 10 shifts or about 220 washing opportunities). This is much faster than the habit formation measured in study 1 of gym attendance, where the median time to habit formation was 4 to 7 mo.

D. Additional Analysis of Reward Sensitivity. As in study 1, we explore whether sensitivity to reward changes is associated with handwashing habits. As in study 1, we use both a hypothesized reward change and an intervention designed to change the reward value of handwashing. The hypothesized reward change is the last opportunity a caregiver has to wash their hands during a work shift—this is a dummy variable encoding the final room visit for their shift. The hypothesis is that they are less likely to wash their

Table 4. Variables which LASSO identified as being most predictive across the handwashing sample

	Importance	Q1	Median	Q3	% zero	% positive	Predictive effect
Compliance last shift	0.77	0.66	0.70	0.92	0	100	Positive
Entry indicator	0.35	−0.33	−0.28	−0.04	18	5	Negative
Compliance within a room	0.12	0.00	0.01	0.14	33	51	Negative
Room compliance of others	0.08	0.04	0.05	0.12	32	66	Negative
Prev. room compliance	0.07	0.03	0.04	0.11	32	65	Negative

hands because it is less important to do so, for hospital hygiene, when they are leaving, or more important because they are going home. There is no statistically significant effect—this variable does not affect handwashing differently pre- and post-habit in the within-participant analysis (*SI Appendix*, section 4.1). This might be explained, in part, because we do not observe individuals from the true "start of habit formation" in our handwashing dataset or because exiting the last patient room during a shift does not truly represent a reliable reward change.

We also test the sensitivity of low- and high-predictability caregivers to various sporadic incentive interventions designed to increase handwashing. The goal of this test is to see if the stronger effect of the StepUp gym attendance intervention on low-predictability participants is replicated. However, these interventions did not have proper control groups; they were also created by the dispenser company (Proventix) rather than behavioral scientists. Unlike in the more careful StepUp design, there are no differential effects of interventions based on differences in participant predictability (*SI Appendix*, section 4.2).

3. Discussion

Despite the personal and policy relevance of understanding habits, there has been a notable absence of field studies examining the formation of individual habits using large-scale, observational data. To address this gap, we introduce a machine learning approach called PCS that identifies the context variables that best predict behavior for each individual.

In models of habit commonly used in economics and applied fields, the context predictor of greatest interest is recent past behavior (20) and (21); *SI Appendix*, section 1. (4) notes that this narrow focus on past choices is also evident "in applied social psychology, as a well as other areas such as health, social medicine, or education, and may have stalled progress in habit theory for quite some time" (p. 3).

Applying PCS to two longitudinal datasets, we identify individual differences in predictability and context cues that predict behavior for each individual. While past behavior is confirmed as a reliable predictor of future behavior, PCS also revealed other predictive variables and heterogeneity in the rate of habit formation across domains. Contrary to the idea of a "magic number" of days in which habits form, our findings suggest that developing a handwashing habit takes weeks, while developing a gym habit takes months. One possible explanation is that relative to handwashing, gym attendance is a less frequent and more complex behavior. Handwashing is more likely to involve chained sensorimotor action sequences which are more automatic (22).

Animal learning experiments have shown that reduced sensitivity to reward changes is a strong marker of habit formation (23) and (24). We tested the association between habit formation and reward sensitivity in a well-powered, random-assignment megastudy (12). We found that more habitual gym attendees are less likely to respond to interventions ($r = -.48$, $P < .001$). However, we did not find evidence of reward insensitivity in other scenarios, such as unusual weather or handwashing policies. These results suggest that reward insensitivity may be less common in human habit formation or too small to detect, for the most part, in our field data. Alternatively, our datasets may lack variables that sufficiently operationalize changes in reward. Further tests of the reward-insensitivity hypothesis are a priority for future research using field data.

In conclusion, we show that PCS is a machine learning approach well suited to the study of habit formation in natural settings. The potential of this technique for advancing personalized behavior change is hinted at by our finding that less habitual gym attendees are more sensitive to randomly assigned

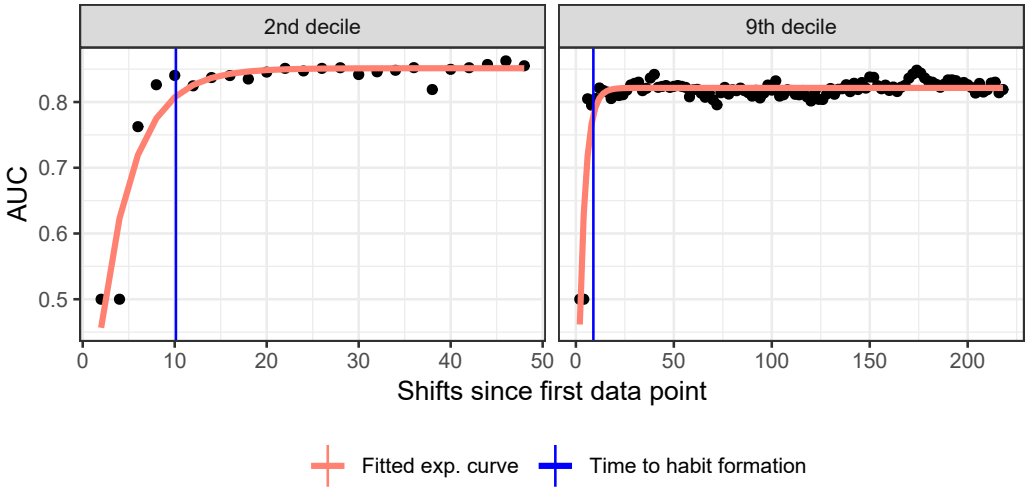


Fig. 3. Estimation of the speed of habit formation for handwashing, for the second lowest (*Left*) and second highest (*Right*) deciles of sample size. Note that the x-axes are different because, by construction, the deciles have different lengths of sampled time. The times to habit formation in these two deciles (shown by where the vertical blue line intersects the x-axis) are 10 (*Left*) and 9 shifts (*Right*; about 225 to 250 handwashing episodes).

interventions designed by behavioral scientists. Additionally, PCS may help identify the sensory, mental, and physical context variables that trigger a behavior like drug use. In conjunction with the increasing availability of large-scale, longitudinal datasets, we hope this innovative methodology inaugurates a new era in the study and personalization of behavior change interventions (25).

Data, Materials, and Software Availability. The data analyzed in this paper were provided by 24 h Fitness and Proventix. We have their legal permission to share the deidentified data. The data and code to replicate the analyses are available at https://osf.io/m8gdp/ (26).

ACKNOWLEDGMENTS. We thank Anthony Kukavica, Predrag Pandiloski, and Mira Potter-Schwartz for excellent research assistance. We thank Hengchen Dai for her help accessing and interpreting the Proventix data. We thank the Sloan Foundation G2018 11259 (CFC), the Behavior Change for Good Initiative (whose contributions were funded by the Robert Wood Johnson Foundation, the AKO Foundation, J. Alexander, M. J. Leder, W. G. Lichtenstein, the Pershing Square Fund for Research on the Foundations of Human Behavior from Harvard University and by Roybal Center grants (P30AG034546 and 5P30AG034532) from the National Institute on Aging), the Linde Institute (X.L.), and the Chen Neuroscience Institute at Caltech (A.B., CFC) for the financial support. We also thank audiences at BCFG, TA-DAH, Behavioral Ops (Shanghai), Stanford GSB, HKU, AAAI, and our internal science team meetings, for helpful ideas. We thank 24 h Fitness for partnering with the Behavior Change for Good Initiative at the University of Pennsylvania to make this research possible, as well as Proventix. Portions of this paper were developed from the thesis of A.B.

Author affiliations: [a]Department of Marketing, HEC Paris, Jouy-en-Josas 78350, France; [b]Department of Marketing, The University of Chicago Booth School of Business, Chicago, IL 60637; [c]Operations, Information and Decisions Department, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104; [d]Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125; [e]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104; and [f]Computational and Neural Systems, California Institute of Technology, Pasadena, CA 91125

1. S. Fleetwood, A definition of habit for socio-economics. *Rev. Soc. Econ.* **79**, 131–165 (2021).
2. W. Wood, D. Neal, The habitual consumer. *J. Consumer Psychol.* **19**, 579–592 (2009).
3. D. T. Neal, W. Wood, M. Wu, D. Kurlander, The pull of the past: When do habits persist despite conflict with motives? *Person. Soc. Psychol. Bull.* **37**, 1428–1437 (2011).
4. B. Verplanken, "Introduction" in *The Psychology of Habit*, B Verplanken, Ed. (Springer, 2018), pp. 1–10.
5. P. Lally, C. H. van Jaarsveld, H. W. Potts, J. Wardle, How are habits formed: Modelling habit formation in the real world. *Euro. J. Soc. Psychol.* **40**, 998–1009 (2010).
6. N. Kaushal, R. Rhodes, Exercise habit formation in new gym members: A longitudinal study. *J. Behav. Med.* **38**, 652–663 (2015).
7. M. Fournier *et al.*, Effects of circadian cortisol on the development of a health habit. *Health Psychol.* **36**, 1059–1064 (2017).
8. B. Gardner, P. Lally, "Modelling habit formation and its determinants" in *The Psychology of Habit*, B. Verplanken, Ed. (Springer, 2018), pp. 207–229.
9. K. Volpp, G. Loewenstein, What is a habit? Diverse mechanisms that can produce sustained behavior change. *Organ. Behav. Human Decision Process.* **161**, 36–38 (2020).
10. J. A. Ouellette, W. Wood, Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychol. Bull.* **124**, 54–74 (1998).
11. A. Mazar, W. Wood, Illusory feelings, elusive habits: People overlook habits in explanations of behavior. *Psychol. Sci.* **33**, 563–578 (2022).
12. K. L. Milkman *et al.*, Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).
13. J. D. Lee, D. L. Sun, Y. Sun, J. Taylor, Exact post-selection inference, with application to the lasso. *Ann. Stat.* **44**, 907–927 (2016).
14. P. Lally, B. Gardner, Promoting habit formation [Suppl 1]. *Health Psychol. Rev.* **7**, S137–S158 (2013).
15. H. Dai, K. Milkman, J. Riis, The fresh start effect: Temporal landmarks motivate aspirational behavior. *Manag. Sci.* **60**, 2563–2582 (2014).
16. S. DeWit *et al.*, Shifting the balance between goals and habits: Five failures in experimental habit induction. *J. Exp. Psychol.* **147**, 1043–1065 (2018).
17. E. R. Pool *et al.*, Determining the effects of training duration on the behavioral expression of habitual control in humans: a multi-laboratory investigation. *Learn. Mem.* **29**, 16–28 (2022).
18. H. Dai, K. L. Milkman, D. A. Hofmann, B. R. Staats, The impact of time at work and time off from work on rule compliance: The case of hand hygiene in health care. *J. Appl. Psychol.* **100**, 846–862 (2015).
19. B. R. Staats, H. Dai, D. Hofmann, K. L. Milkman, Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. *Manag. Sci.* **63**, 1563–1585 (2017).
20. G. S. Becker, K. M. Murphy, A theory of rational addiction. *J. Polit. Econ.* **96**, 675–700 (1988).
21. J. P. Dubé, G. J. Hitsch, P. E. Rossi, State dependence and alternative explanations for consumer inertia. *RAND J. Econ.* **41**, 417–445 (2010).
22. B. Balleine, A. Dezfouli, Hierarchical action control: Adaptive collaboration between actions and habits. *Front. Psychol.* **10**, 2735 (2019).
23. C. D. Adams, A. Dickinson, Instrumental responding following reinforcer devaluation. *Q. J. Exp. Psychol.* **33B**, 109–121 (1981).
24. E. Tricomi, B. W. Balleine, J. P. O'Doherty, A specific role for posterior dorsolateral striatum in human habit learning. *Euro. J. Neurosci.* **29**, 2225–2232 (2009).
25. C. J. Bryan, E. Tipton, D. S. Yeager, Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat. Hum. Behav.* **5**, 980–989 (2021).
26. H. Ho, Machine learning and habit formation. Open Science Framework. Deposited 26 February 2023.