*Supplementary Materials for:*

# Context-sensitivity of behavior in field data: Using machine learning to study habit formation in natural settings

Anastasia Buyalskaya Hung Ho Xiaomin Li Katherine Milkman
Angela Duckworth Colin Camerer

March 21, 2023

## Table of Contents

**1 Literature Review** **3**

1.1 Overview . . . 4
1.2 Psychology . . . 6
1.3 Computational Neuroscience . . . 10
1.4 Economics . . . 13
1.5 Political Science . . . 16

**2 Dataset Descriptions** **17**

2.1 Hand Washing Data . . . 17
2.2 Gym Attendance Data . . . 18
2.3 Description of Context Variables . . . 18
2.3.1 Hand washing data . . . 18
2.3.2 Gym attendance data . . . 19

**3 Analysis Details** **19**

3.1 Individual LASSO Regressions . . . 20
3.2 Model Selection Challenges in LASSO . . . 23
3.3 AUC vs Frequency . . . 29
3.4 Speed of Habit Formation . . . 30

**4 Field Tests of Insensitivity to Reward Devaluation** **32**

4.1 Within-subject Field Tests of Insensitivity to Reward Devaluation Pre- and Post-habit . . . . . 33

4.2 Between-person Predictability Reactions Toward Incentivised Intervention . . . . . 38

4.3 Sensitivity Analyses . . . . . 44

**5 Additional Analyses: Demographic Predictors of AUC** **45**

5.1 Motivation . . . . . 45

5.2 Variable List . . . . . 46

5.3 Correlation Matrix . . . . . 46

5.4 Regression Results . . . . . 47

**6 Human Subjects Protections** **48**

6.1 Gym Attendance . . . . . 48

6.2 Hand Washing . . . . . 49

**7 Review of Habit Formation Studies** **49**

7.1 Summary of Previous Habit Formation Studies . . . . . 49

# 1 Literature Review

Since habit naturally crosses disciplinary boundaries, the most promising understanding of it is likely to come from integrating evidence and methods across disciplines (1) (pg. 42). That is our approach. The purpose of the following section is to highlight key papers from the major disciplines we take evidence and methods from. Specifically, this section summarizes how habit is studied in psychology, computational neuroscience, economics, and political science.

We first present a summary table comparing how these literatures have addressed the different hallmarks of habitual behavior in Table S1. We mark in bold what we view as being the "best practices" of measurement. Some of the attributes - such as time to habit formation - don't have an ideal best practice yet. We follow the table with more detailed reviews of each field's major contributions.

## 1.1 Overview

Table S1: General practices and measures of habit features across different social and natural sciences. Bold typeface indicates our subjective valuation of **best practices**

| | **Psychology** | **Computational Neuroscience** | **Economics, Political Science** | **This Paper** |
|---|---|---|---|---|
| **Methods** | Lab & field experiments, field data | Lab experiments | Econometrics, field experiments | Machine learning, field data |
| **Data types** | Self-report surveys (SRHI), behavior | Behavior, **neural activity** (small samples), **lesions** | Behavior, often **large samples** | Behavior, **large samples and time span** (no attrition) |
| **Length of habit formation** | Thought to be around 2 months[1], **estimated from increase in self-report** | Simple motor habits, trained from <1 hour to several hours | Field experiments often assume 1 month[2] | **Estimated from increased predictability** |
| **automaticity** | measured by **self-report[3]** and newer implicit measures | **Response times, changes in brain activity[4]** | Typically not inferrable from choices alone | No data |
| **Reward devaluation insensitivity (RDI)** | Measured in **"extinction" test[5]** after devaluation by feeding to satiety etc. | Optogenetics modulated habit.[6] Only weak evidence for short-run (1 hr) human tasks[7]. Evidence of rat-human homology[8] | Not tested | No evidence of RDI tested with hypothesized ecologically-relevant reward changes. **StepUp gym attendance after reward increase is negatively correlated with predictability** |
| **Context sensitivity** | "habit discontinuity hypothesis (HDH)"[9], **habits more easily changed when context changes** | Not studied to our knowledge | Behavior reduced by context change (HDH) | Large set of context variables permits estimation of **individual specific context-sensitivity** |
| **Individual differences** | Most studies aggregate across individuals[10] | Learning parameters can be different | Most studies aggregate across individuals | **AUC measures differences in predictability** across people. Can compare context variables across people |

Footnotes for Table S1:

[1] This estimate comes from (2), who compiled panel observational data using self-report questionnaires and fit individual-level habituation models. They estimated that it took anywhere from a few weeks to over half a year to reach the 95% asymptote of behavior.

[2] One of the first papers in economics to empirically test habit formation used one month (3). While they do not explicitly justify the choice of one month, they do state "Our results indicate that it may be possible to encourage the formation of good habits by offering monetary compensation for a sufficient number of occurrences, as doing so appears to move some people past the "threshold" needed to engage in an activity." Some later studies continued to use one month (4). Acknowledging that this might be too short of an intervention interval for habits to form. (5) wrote "An alternative explanation is that some subjects would have experienced an increase in postintervention attendance if the intervention period had been longer".

[3] BFCS ("Behavior Frequency x Context Stability") measure (6) is a self-report index covarying past behavior frequency with measured contextual variables. It is designed to test behaviors which are repeated frequently in familiar contexts are more likely to become habitual. The most commonly used self-report measure is the SRHI ("Self-Report Habit Index"). A subscale has been designed specifically to capture automaticity and is called SRBAI ("Self-Report Behavioral Automaticity Index"), including questions like "I do this behavior without thinking."

[4] Classic work by (7), recording from rodent brains during maze-running, found concomitant changes in response times, increases in performance, and reductions in neural signals of imminent reward. We think of these as hallmarks of automaticity but there may be other markers in humans (e.g. degraded memory for past activity, or misattributions of cause from habit to internal states see (8)).

[5] The classic extinction test paradigm was developed by (9). In an important field experiment with humans akin to the extinction test, (10), found that individuals continue to eat stale popcorn beyond satiation if they are in a specific context which cues the habitual behavior (specifically, watching a film or eating with their dominant hand).

[6] See (11).

[7] See (12).

[8] The only paper to find evidence of reward devaluation insensitivity with humans in fMRI (13) has not been directly replicated. The same feeding-to-satiety paradigm and analogous versions with token-money reward devaluation have not reliably shown devaluation after short-run training (14). Establishing robust devaluation insensitivity for humans is an active area of research (15).

[9] The HDH is the hypothesis that behavioral interventions aimed at changing habits often work best after a lifestyle change (16). It seems to be an open question whether this is due to changes in prices and opportunities, or to subtler forms of context-sensitivity in which missing context cues do not trigger associations (or cravings, for drugs of abuse).

[10] As noted in (17), "Theory has also been inadequately tested at the individual level. Most (80 studies; 98% of all the study sample) studies have exclusively modelled between-person variation in habit, based on aggregates of

individuals' habit scores. Yet, habitual action is inherently idiosyncratic, based on personally acquired behavioural responses to personally meaningful cues. Within-person effects cannot be reliably interpreted from aggregations of processes that differ between people."

---

## 1.2 Psychology

Psychologists define habit as a behavior which is prompted automatically by contextual cues as a result of learned context-action associations (18). This definition combines two key attributes of habitual behavior which guide a lot of the psychology research: automaticity, and predictable context-sensitivity.

Some habit researchers make further distinctions about what should be considered a habit. For example, (17) argues that the initiation and performance of a behavior are distinct. He classifies behavior into one of three types: habitually initiated but consciously performed (his example: riding a bike to work every morning), consciously initiated but habitually performed (his example: exercising at the gym), or habitually initiated and habitually performed (eating a snack in the afternoon). This is a sensible distinction but without measures of automaticity we cannot apply it to our data. It is simply a reminder that repeated behaviors that we call habits need not be unconscious or automatic.

*Context-Sensitivity*

The focus on context-sensitivity came from evidence that habits arise when context-stable behavioral repetition creates a "transfer" from (internal) goals to (external) associations with environmental cues (19). In the language of animal learning and instrumental conditioning, an S-R-O relation in which an association is developed between a stimulus (S), the response (R) it elicits, and a reward outcome (O), becomes habitual as an S-R relation.

Habits are not "innate" to the behavioral repertoire in the way reflexes are (e.g. one is not born with, but must develop, the habit of tooth-brushing, - unlike the reflex of being startled by something unexpected, which is present in newborns at birth). Instead, most habits begin as goal-directed behaviors. Eating solid food using a fork, for example, begins as a very deliberate goal-directed behavior in small children (one which requires a lot of motor and cognitive control in the beginning). It may take months or even years for eating to become an automatic motor sequences with little need for cognitive control, such that a habit can form. In adults, who have cheaper cognitive control and can eat "mindlessly," the behavior of eating is ripe for developing associations with context and reward independent of nutritional goals. Specifically, the "trigger" to eat is often transferred to context elements of the environment which reliably co-occur with the behavior. For example, people who snack frequently in a stable context are no longer driven by an internal motivation to eat, but rather by an environmental cue (20).

The range of possible context cues is usually idiosyncratic, because they are likely to vary by the type of behavior and by individual. The context cues most often studied in psychology tend to be physical time, space or social cues which are easily measurable (such as the location in which a behavior occurs, the day of the week, the time of day, whether other people were present when the behavior was executed, etc.) (21). However, one can imagine less easily

measurable contextual cues, such as a specific mood, sensory input or a memory, as triggering a habit. These may be harder to measure objectively, for example relying on individuals' recollection of a memory or ability to verbally describe a feeling, but are still important. In clinical studies for example, stress and visual cues that induce craving states are often measured given their importance for behavior (22–24). Psychology and applied psychology (e.g., health behavior research) are disciplines that are the most focused on, and seek to measure, context-sensitivity.

*Automaticity*

The other attribute that psychologists seek to measure to determine whether a behavior is truly a habit is automaticity (25, 26). A behavior is considered automatic is if it "brought to mind by cognitive processes largely outside of conscious awareness" (21) (pg. 14).

An early start on this definition came from (27), who presented four criteria of automatic behavior. The first is awareness of the cognitive process which gives rise to the behavior. The second is intentionality – or control – over the initiation of the cognitive process. The third is efficiency – automatic processing requires fewer mental resources. And the fourth is control – the ability to stop or alter the cognitive process after it has begun. Even if there was an easy way to measure all four of these, Bargh noted that not all of the criteria need to be met in order for a behavior to be considered automatic. In fact, a behavior which only meets two or three may still be automatic, confusing the definition even further still. More recent theoretical models of automaticity have maintained the view that it is a multidimensional construct, continuing to emphasize the unintentionality, uncontrollability, and unconscious execution of behavior (28).

Animal learning studies also illustrate how simple theories of automaticity and habit are often hard to evaluate conclusively. (29) trained rats on a two-lever-press paradigm for 20 days or 60 days, then tested for automaticity and sensitivity to reward devaluation. The more extensively trained rats performed the rewarding lever presses more often and more quickly (by these measures, their behavior became more automatic). But both groups exhibited similar insensitivity to reward devaluation and a difference in apparent goal-directed control of the two different levers.

*Measurement*

Next, we'll examine the most common measures used in psychology to assess context-sensitivity and automaticity of behavior. Classic research on habits (e.g. in animal learning studies) inferred habit from observed behaviour in response to cues. However, psychological research with humans has typically used self-reported answers to questions about a participant's own behavior. In a meta-analysis looking at 136 empirical studies which applied ideas from the habit literature to health behaviors over the years 1998-2013, (17) found that self-report scales are still the main methods used to measure habitual behavior. Two scales dominate the literature.

The first scale – relied on by 88% of the studies in Gardner's meta-analysis – is the SRHI, or "Self-Report Habit Index" (30). Its popularity stems from the fact that the questionnaire is short (a 12-item scale), direct and has become the standard in psychology. One of the questions asks the subject to rate their agreement with the following statement on a Likert scale: "I do this behavior without thinking." A subscale of SRHI was designed specifically to

capture automaticity and is called SRBAI ("Self-Report Behavioral Automaticity Index").

Accurately self-reporting habits and automaticity relies on good memory meta-cognition (our thinking about our thinking). This raises a fundamental question about limits to accuracy of self-reports about habit and their automaticity. This question has been debated extensively in the applied psychology literature (31–34).

However, as (35) argue, the SRHI infers habit from reflections on *symptoms of* habitual responding– such as proceeding without effort, conscious awareness, conscious intention, etc.– rather than assuming people have deeper insight into habitual regulation. There could be other limitations of self-report, such as misconstrual of items. It could also be that asking people to reflect on the characteristics of behavioral performance may lead them to report behavioral frequency rather than habit symptoms. However, (36) found from think-aloud protocols that only 10% of responses indicated problems.

Another popular measure – used by 12% of the studies in Gardner's review – was Ouellette and Wood's (1998) BFCS ("Behavior Frequency x Context Stability") measure. This is a self-report index co-varying past behavior frequency with context stability. This measure is based on the assumptions outlined earlier that behaviors which are repeated frequently in familiar contexts are more likely to become habitual (18). The questions aim to assess both directly, phrased as "how often do you do this behavior?" and "when you do this behavior, how often is this cue present?"

The BFCS self-report clearly depends on accurate memory recall and metacognition. It is conceivable that more habitual activities are less well remembered. Consider the example of checking a mobile phone. Using a smartphone app which calculated true frequency of phone use, (37) were able to track the phone behavior of 27 participants over the course of 14 days. They found that there was no correlation between true phone-checking behavior and a self-report measure called the Mobile Phone Problem Use Scale ("MPPUS"). The MPPUS is a 27-item questionnaire that includes items such as "I can never spend enough time on my mobile phone". This anecdote points to another fault with self-report measures: they are inherently retrospective, relying heavily on hindsight. But memory degrades quickly – with the details of a morning becoming foggy as one enters their afternoon – meaning the timescale at which these questionnaires are administered is crucial.[1]

A more systematic review comes from (38), who ran a meta-analysis of 47 studies to measure the link between logged and self-reported digital media use. To evaluate the association between self-reported and logged media use, 66 effect sizes from 44 studies were considered (n=52,007) and correlations were calculated with robust variance estimation (RVE). Their analysis concluded that self-reported media use has a positive but medium-magnitude relationship with logged (objective) measurements (r=0.38, 95% CI = 0.33 to 0.42, $p$ <0.001). Furthermore, problematic media use showed a slightly smaller association with usage logs (r=0.25, 95% CI = 0.20 to 0.29, $p$ <0.001). These studies, along with another relevant critique from (39), point out the challenges of self-reporting some aspects of habits from memory.

Besides these two most common scales, two other measures in Gardner's meta-analysis were used in just one study

[1] A modern technique which the smartphone makes available is real-time experience sampling where people are prompted to discuss situational cues and whether they are executing a habit.

each (1% of the sample each; proportions add to more than 100% because of rounding). The EHS ("Exercise Habit Survey"), used in one study, is similar to BFCS. The other measure was an association test, designed to measure cue-behavior associations underpinning habitual behaviors (an implicit association test).

So while psychologists have identified two important elements of habitual behavior - context-sensitivity and automaticity - there have been some concerns about how good their current measurement tools are as proxies for true habitual behavior (1).

*What behaviors can become habitual?*

Psychologists study habits across a range of behavioral domains. Popular domains of study include activities which are done frequently: eating, exercising, and hygiene behavior. However there is some debate around how complex a behavior can be before it can no longer be considered a candidate for becoming habitual. This is in part due to research which has demonstrated that simpler actions like drinking water tend to become habitual more quickly than complex actions like exercise routines (2). The idea is also evident in animal learning, in which chained motor sequences are slower to form habits (7).

Focusing on the two behaviors covered in this paper, hand-washing seems to be ripe for becoming habitual because it involves a short motor sequence. (40) (pg. 248) suggest that hand-washing habits "minimize[e] cognitive resources required for a given behavior to ensure that it can be performed with a maximum of patients and/or for when such resources are especially needed".

Whether exercise can become habitual is more debatable (41). Physical activity, particularly travelling to a gym for exercise, is different from other familiar habitual behaviors. Two differences worth noting are that it is a multi-step behavior, not a simple motor action, and that it takes a long time to perform. However, the type of exercise which is done inside a gym is often a relatively straightforward motor action as well. Running on a treadmill, rowing, lifting weights – while requiring "control" and "awareness" and hence not meeting the definition of automaticity – are simple enough that many gym goers are able to multi-task while doing them – as is obvious by watching gym-goers listening on their headphones, holding a conversation, reading or watching TV while they exercise. Secondly, the other attribute of habitual behavior, context-sensitivity, is likely present for gym goers. Location, other people, time of day, or biological states (for very regular exercisers) are likely candidates for cuing the decision to attend the gym.

*Speed of Habit Formation*

Behavior goes from being goal-directed to being habitual through frequent repetition in a context-stable state. Many researchers have been interested in this habit formation process, and in particular, how long it takes for a habit to form. However, answering this question using traditional tools from psychology is difficult because it requires a significant amount of data collection (obtaining regular SRHI responses over many days, as an example). This requires researcher time and persistent longitudinal engagement by subjects. Hence, only a handful of studies have been done to answer this question (2, 42, 43).

A seminal study is (2). The researchers collected SRHI measures for 82 subjects daily over the course of 12 weeks

for an eating, drinking or physical activity behavior chosen by the subject. (2) then fit a curve to each individual's self-report scores through time in order to measure the time it took them to reach 95% of the asymptote (their definition of when something became a habit). They were able to fit the model for 62 individuals and obtain a good fit for 39 out of those 62, finding that "performing the behaviour more consistently was associated with better model fit." Their results showed that the median time to habit formation was 66 days, with a range of 18 to 254 days to habit formation depending in part on the complexity of the behavior (e.g. the relatively simple act of drinking a glass of water was quicker to form habits than a more complex physical activity).

Another study looked at the development of exercise habits by asking new gym members to complete surveys over the course of 12 weeks (42). They found that exercising at least four times per week for 6 weeks was the minimum requirement to establish an exercise habit, based on the time at which behavior appeared to reach an asymptote (i.e. not change significantly after that time period). The most recent observational study focused on the effect of circadian cortisol (modulated by time of day) on the development of a simple physical habit. (43) tracked 42 French students for 90 days as they did a stretching exercise behavior. Some students were assigned to do it in the morning (when cortisol levels are high) and some in the evening (when cortisol is low). The SRBAI was collected daily, and the speed of habit formation process was then modelled using learning curves by fitting a four-parameter logistic curve to SRBAI responses. The curve-fitting process was successful, converging for each participant (in contrast to the power function following (2), which the researchers also tried, finding that only 48% had a moderate fit as defined by $R^2 > 0.70$). Their results showed that the morning group achieved automaticity at an earlier time point (106 days) than the evening group (154 days), concluding that time of day influences the speed of habit formation.

Of these three quantitative studies, all showed that "habit typically develops asymptotically and idiosyncratically, potentially differing in rate across people, cues and behaviors" (44) (pg. 220).

### 1.3 Computational Neuroscience

What does habitual behavior look like in brain activity? This has been the driving question for much research in computational neuroscience. This research tends to focus on the neural basis of the two types of cognitive processing mentioned in the last section: "goal-directed" behavior, a more deliberate cognitive functioning, and habitual behavior. The existence of these respective decision making systems is now well-accepted and commonly modeled theoretically as model-free (MF) and model-based (MB) decision-making (45–47). MF learning transitions to habit learning with extensive experience.

When a new habit is being learned, inputs to the midbrain dopamine system drive dopaminergic neural activity which encodes reward prediction errors (RPEs). These RPEs serve as learning signals. Learning an accurate prediction of a stable reward results in smaller and smaller reward prediction errors over time. These signals are thought to modulate synaptic plasticity in the striatum which in turn serves as the "gate-keeper for tentative motor plan representations" (48). The striatum can be further segmented into two distinct areas: the dorsolateral striatum (DMS) and the dorsomedial striatum (DLS).

Instrumental behaviours which respond to reward values may start out as goal-directed actions largely controlled

by the associative striatum (DMS), which controls more goal-directed activity, when they are first being learned. But under certain conditions and with enough repetition, these behaviors may become habitual and no longer contingent on reward. Then cognitive control shifts to the sensorimotor striatum (DLS), which controls more stimulus-driven behaviors (49, 50). Functional MRI studies which are used to localize brain activity during decision making have confirmed that habitual processing tends to occur in the "sensorimotor loop," which connects the basal ganglia with the sensorimotor cortices and parts of the midbrain (13, 49). Brain scans have therefore been used to confirm that the brain has two independent sources of action control which govern behavior, and to help determine whether a behavior is habitual or goal-directed (12).

So what are the conditions necessary for a behavior to move from being goal-directed to being habitual? The animal literature suggests that habituation requires a behavior to be repeated many times – a process known as "overtraining" (13). Another variable which seems to contribute to habit formation is learning under stress. Lab studies have found that inducing stress (in animals, including humans) leads to quicker formation and reliance on habitual behavior (51). Finally, optogenetic studies using rodents found that the disruption of rodent infralimbic cortex (a region in the medial prefrontal cortex which has been shown to be necessary for the expression of habits) temporarily blocked habit responding (11).

One important test used to determine whether a behavior is habitual or not is a test of sensitivity to reward devaluation. The procedure originated in animal learning studies, with (9), who studied how lever pressing in rats could become habitual. When they analyzed habit, they described it as a behavior which becomes so automatic that even devaluation of the reward value of an outcome will not have a large effect on the execution of the habitual behavior. Specifically, they found that mildly poisoning a food pellet after a rodent has developed a highly-trained habit of lever-pressing for the pellets did not deter the rodent from continuing to press the lever. This phenomenon has been termed insensitivity to reward devaluation, and is a behavioral hallmark of habitual processing.

There is some evidence of insensitivity to reward devaluation in humans. (13) trained participants to learn that responses to two different fractal images were associated with two different snack rewards. After overtraining (choosing their preferred fractal many times in short succession), they were given one of the snacks to eat to satiety, which presumably devalued it. Subjects who had food devalued this way continued to choose the fractal associated with the devalued foods, indicating habit. This is evidence of human insensitivity to reward change similar to the animal experiments.

However, other researchers have not been able to replicate these findings (14). This raises the question of whether an experimental paradigm using rodents can be easily transferred to human behavior. Another concern which has been raised about the reward devaluation paradigm is that it implies that behavior which is not goal-directed is necessarily habitual. For example, the goal-independent behavior may not be context-sensitive (21) (pg. 23). However, there remains an interest in replicating this effect with humans with different paradigms and training protocols.

One of the best studies showing insensitivity to reward devaluation in humans is a psychology study. While it does not have neuroscience data, it is included here because it is a clear illustration of this reward devaluation test. (10)

found that people were more likely to overeat stale ("devalued") popcorn in a context which cued habitual behavior of eating popcorn (e.g. watching a movie in a cinema) but not when they were in an unfamiliar popcorn-eating context (e.g. watching a movie in a meeting room, or eating the popcorn with their non-dominant hand) which did not cue the habitual behavior. The effect captures a two-way interaction (cinema vs. meeting room or dominant vs. non-dominant hand and whether the popcorn received was stale or fresh) and is evident only among individuals classified as "high habit" (vs. medium or low habit) per self-reports on a 7-point scale used to assess habit strength for eating popcorn in movie theaters. The same study found that for low or medium habit individuals, or high habit individuals in novel contexts, like eating popcorn in a meeting room, behavior remained sensitive to reward value and decreased in frequency when the popcorn was stale (devalued).

There is some indirect evidence that the *reliability* of the reward history is associated with habit formation, as measured by insensitivity to reward devaluation (which is a central concept in (52)'s theory of neural autopilot). For example, (53) track the fraction of trials on which a model-free system or goal-directed system recommends the optimal choice. These fractions are then compared by an "arbitrator" system which weights actions recommended by the two systems according to their recommendation accuracy. Using fMRI they report evidence for neural circuitry consistent with this arbitration computation.

There is substantial evidence that habit formation is stronger after training on a random interval schedule compared to a random ratio schedule. (These "schedules" are terms from animal learning theory referring to how often, and based on what behavior, reinforcing rewards are delivered.) In ratio schedules reward is based on the animal's behavior– e.g., each lever press has an independent 1/30 chance of being rewarded in a so-called RR30 schedule. In interval schedules, reward is based on passage of time– e.g. every 2 seconds, there is a 10% chance of reward being delivered upon the next lever press, in a so-called RI20 schedule. In RI20, if the animal waits 20 seconds, then reward will be delivered with certainty upon the next lever press. In RR schedules there is no such guarantee.

Reward reliability in (52) is defined by the absolute value of reward prediction errors. Consider the simple case in which rewards, normalized to 1, are random at rate $p$. Then the expected reward reliability is $p(1-p) + \mid (0-p)(1-p) \mid = 2p(1-p)$. This expression has a minimum at $p^* = .5$ and is increasing for lower and higher reward rates. In most animal learning paradigms, the reward rate $p$ is well below .5, so that reward reliability is increasing in the reward rate $p$. In the experiments of (54), the reward rate is higher from the interval schedule than in the ratio schedule training; and habit formation is stronger after interval schedule training. This is a small bit of evidence consistent with a role for reward reliability. In addition, (55) reports slightly stronger habit formation when the reward rate is higher (RR15 compared to RR30 in experiments 2 and 1), also consistent with a role for reward reliability.

Habitual behavior that is automatic is accompanied by measurable psychological and biological features, including faster response times, limited attention during choice (50) and degraded declarative memory (explaining the basis for choice when asked, see (56)).[2] These attributes can be studied using a range of measurement tools, some of which

[2]Studying two patients with large MTL lesions, (57) found neurotypical-level performance in an overtrained discrimination task with

are more portable outside of a laboratory setting, including eye-tracking methods to measure attention.

## 1.4 Economics

Economic theories and empirical tests have generally used the term "habit" in one way: To describe history-dependent "adjacent complementarity" of goods or services [3]. The theories are motivated by strong evidence of empirical correlation between past and current consumption. These models therefore specify consumption utility as a function of actual immediate consumption *relative to a reference point* or 'consumption habit' (60–62).

This approach was never empirically microfounded in psychology or neuroscience but it is mentioned prominently in the earliest studies creating a foundation for intertemporal choice. (63) wrote: "One cannot claim a high degree of realism for [consumption insensitivity], because there is no clear reason why complementarity of goods could not extend over more than one time period".

In conventional microeconomic consumer theory, "complements" are pairs of goods X and Y which increase each other's marginal utilities when consumed together—that is, the marginal utility of X is greater if you have more Y. Familiar examples of complements include hot dogs and hot dog buns, hammers and nails, and computer hardware and software. Koopmans's point is that complementarity could extend to the same good consumed in adjacent periods (called "adjacement complementarity"). Rather than treating hot dogs and hot dog buns as complements, yesterday's hot dogs and today's hot dog consumption are considered as possible complements.

In one macro-finance specification (64), the crucial variables are current consumption $C_t$ and habit $X_t$. Utility depends on past aggregate consumptions $C_{t-1}, C_{t-2}\ldots$ through another equation. In that specification $U_t = \frac{(C_t - X_t)^{1-\gamma}-1}{1-\gamma}$ (and $X_t$ is related to previous consumption levels in a complicated way).

Such preference assumptions were used in macroeconomics and finance to explain facts which are puzzling in specifications in which utility depends only on consumption (65, 66). (64) motivate their specification with the following hypothesis[4]: "repetition of a stimulus diminishes the perception of the stimulus and responses to it" (pg. 208). This is indeed a property of sensory systems which are adaptive. However, these types of "repetition suppression" are very short-run (e.g., seconds to minutes or days). Whether the same kind of history-dependent adaptation works for, say, quarterly consumption by a household is an open question.

(68) derives a set of axioms relating the functional form of habitual history-sensitivity to underlying principles that are mathematically equivalent. The functional representation of utility is:

no declarative memory or conscious awareness. Thus, lesion patients could perform the task automatically. However, performance was completely degraded to random on a minor task variant. The two patients also learned the task about as quickly as four monkeys did.

[3] Another form of habit is the idea that the discount factor depends on consumption (58). They appeal to an intuitive concept of "habits of thrift" or luxurious spending hypothesized by (59) (pg. 337-338) (with no evidence) which link more income to less patience. This concept is theoretically interesting but appears to be empirically counterfactual, as much evidence suggests higher income is associated with more patience, rather than less patience.

[4] The phenomenon they are describing is similar to reward prediction learning or, in perception, is called "repetition suppression" (67). It would be useful to explore even a highly speculative link between these psychological foundations and the hypothesized micro-foundation for macroeconomics further..

$$U_h(c) = \sum_{t=0}^{\infty} \delta^t u \left( c_t - \sum_{k=1}^{\infty} \lambda_k h_k^{(t)} \right)$$

where $h_k$ is the habit consumption history k periods in the past and $\lambda_k$ is a decay factor which weighs more distant consumption history less.

A bolder extension of adjacent complementarity is called "rational addiction (RA)" (69). In this approach, current utilities depend on consumption history, due to adjacent complementarity, much as in the (68) formalization. But it is also coupled with self-awareness of the history-dependent structure and planning about the future. In this model, "rationally addicted" people understand that if they consume more X today, they will value X tomorrow more highly.

The key prediction of the RA model is that current consumption will depend on current prices and will *also* depend on expected future prices. For example, once they hear that a large cigarette tax increase will take place soon, rationally-addicted smokers might quit a habit abruptly - *before* the increase occurs. They'll quit right away because they prefer, today, to be an ex-smoker at time T when the tax goes up; otherwise, continuing to smoke at T will be too expensive.

Both the macro-finance and RA specifications are natural in economics because the primitives in economic analyses are stable preferences, Bayesian beliefs, and budget constraint. Habit can then enter into the theory in one of those three ways. The default approach is to define habit as current preference depending on past consumption.

Conventional economic theory with these ingredients does not have learning, RPE, reward reliability in it. There is also no implicit cost of mental effort. And there is no attempt to relate the history-dependent model to adaptive functionality or to neural implementation.

Most economic empirical studies using the RA approach treat the fact that history-dependent consumption could be present in a wide range of goods and activities as a provocative prediction. "People can be addicted not only to harmful goods like cigarettes, alcohol, and illegal drugs, but also to activities that may seem to be physically harmless, such as sports participation, shopping, listening to music, watching television, working, etc." (70) The RA approach does make the non-obvious prediction that current behavior depends on expectations of the future, in sharp contrast to the neuroeconomic habit model which is not forward-looking.

There are many studies of RA. There are two limits in these previous empirics: (1) Most of the early empirical evidence uses very coarse time scales (e.g., quarterly tax receipts to measure state-by-state cigarette consumption); and (2) estimates of the expected future price component are not very good. Expected future prices are usually proxied by past prices, and these proxies may not be independent of current consumption. Even very sophisticated tests on coarse quarterly data have very limited power to test whether there is actually forward-planned RA.

(71) demonstrate the kinds of biases that can lead to results consistent with RA even when the basic data-generating process has no actual adjacent complementarity mechanism. The central test of the forward-looking property of RA is whether current consumption is increasing in (expected) future consumption. Simulations show that when the consumption time series is highly auto correlated (as is typical), even if there is no history-sensitivity, the RA prediction can spuriously appear to hold. However, other diagnostic features of these tests (such as inferred discount factors reasonably close to 1) can also fail in both artificial and actual data sets.

An illustrative example of how history-sensitivity is used in empirical practice is (72). He derived a tractable way to test whether optimal consumption with habit can be rationalized nonparametrically, in the sense that one can find some set of inferred utilities, satisfying simple restrictions like GARP and extended to allow adjacent complementarity, which fits a data set on consumption. The logic of this exercise is that if no set of inferred utilities can "rationalize" the data, then the specification of stable utilities with adjacent complementarity is incorrect.

Crawford applied the method to data on quarterly smoking expenditures for 3,134 Spanish households. The best-fitting habit lag is two quarters. Most households' (91%) data can be rationalized using two lags (compared to only 24% with one lag), but the power of the two-lag test is not very high (only 20% of random-generated data would fail the test for optimization).

History-sensitivity is seen again and again in many types of data: It is established in internet use (73) and employment (74). In marketing it is attributed to inertia or brand loyalty (75–77)[5].

The boldest predictions of the RA theory seem to be just flat wrong. In theory, rational addicts should take advantage of volume discounts on addictive goods, because they will optimally self-ration the goods over time. There is no direct evidence of this pattern (e.g., alcoholics buying in bulk and self-rationing), although it could be that rational addicts are liquidity-constrained. Instead, (79) found in lab and field data that "vice" goods, such as cigarettes, are often purchased in smaller quantities, have higher quantity discounts, and have lower price elasticities than similar virtue goods, regardless of liquidity-constraint. There is also substantial evidence that restricting hours at which addictive goods are sold (typically alcohol) reduces consumption (80). This is inconsistent with rational forward-looking optimization by addicts, who should plan their shopping around reduced hours.

For the purposes of this paper, we also note that the economic RA model does not connect with what is known from psychology and cognitive neuroscience. The latter is loosely constrained by the philosophy that a good understanding of a behavior should have an explanation for adaptive functionality, algorithmic specificity, and neural implementation.

(81) introduced an economic model of a specialized idea of context-sensitivity, from clinical psychology and neuroscience, to explain cue-sensitivity of addiction. In the model, the presence of a state-dependent cue actually changes utility. If a cue value is $x^i$, and consumption activity is $a^i$ (=0 or 1), then the period-specific utility is assumed to be $u(a^i, x^i) = u(a^i - \lambda x^i) + (1 - a^i)\eta$ where $(1 - a^i)\eta$ is the expected utility of the next-best activity if the target activity is not done.

This is a simple economic translation of the evidence about biological addiction from opponent processes to maintain homeostatis, but it is not a biologically plausible general model for everyday habits. An implication of the Laibson specification is that mere presence of the cue creates negative utility (through unpleasant craving) if the good isn't consumed. In the PCS view, the presence of a cue is typically not pleasant or unpleasant; it just predicts behavior through a neural autopilot mechanism driven by reward reliability, rather than via unpleasant craving which addicts "self-medicate" to avoid.

[5]There is some evidence of what (78) calls "situations" (the same as our cues or states) influencing choices but it has not been an active area of research.

(82) create a more general model tailor-made to understand addictive habits. Preferences are influenced both by a numerical state, which catalogs consumption history, how frequently states trigger an involuntary "hot" craving state, and some other features. Their model is not as much a specific theory, as it is a modelling language to describe different kinds of addiction patterns and invite empirical estimation.

The Laibson homeostatic cues model and Bernheim-Rangel M-states model are two examples of state-sensitivity of preferences which go beyond the history-sensitivity in so much empirical work. In their models, the relevant state, on which preferences depend, is a cue or history variable. The idea is that what people subjectively value could depend on an environmental or contextual state (83). Nothing is new or surprising about that– umbrella preference goes up when it's raining. Historically, however, economists were reluctant to allow too broad a range of state-sensitivity of preferences for fear—probably legitimately—that doing so would lead to an erosion of falsifiability. Common examples in which state-sensitivity is central are examples like health, in which health quality (a physical state) clearly influences subjective value of leisure or work.

## 1.5 Political Science

Political scientists have studied habit in the domain of voting. Voting is interesting for our purposes because it is very infrequent— particularly compared to hand-washing or gym attendance, and to other activities studied in empirical applied psychology. It is similarly far from the animal learning-based concept of motor habituation and insensitivity to reward change from hundreds of rapid trials in short time spans, on the time scale of hours or days.

We do not know if the term "habit" should be associated with voting at all, because it seems so dissimilar in many dimensions to animal learning, exercise or eating habits, and most others reviewed in this section. And it also could be that what is called acquisition of a voting habit is better explained by a change in costs and benefits or other causal explanations (84).

We include a discussion of these studies because at least one voting study (85) did link to habit scales. We also think that interested readers who are unfamiliar with these studies can learn a little and judge for themselves whether these political behaviors should be considered habits and how they can be better studies.

Most of the studies show that voting behavior does exhibit some context-sensitivity. Researchers have mostly focused on how a disruption to total voting ("turnout") in one election affects subsequent turnout. The disruptions that are diagnostic are exogenous "natural experiments" which suggests possible causality, as if an experimental treatment changed voting for some people but not similar others. If skipping voting one time breaks one's "taste for voting" – reducing the likelihood of voting in future elections – then voting is considered habitual, in the history-dependent sense, in these articles. And as with many behaviors, past voting behavior predicts future behavior (86).

These studies are of three types.

- Observational studies seek to isolate the impact of an "as-if random" inducement to vote in one year, on voting turnout in subsequent election years (87, 88).

- Experimental studies apply a truly random assignment to inducement to vote and test whether it increases future voting (89, 90).
- "Quasi-experimental" causal identification studies use regression discontinuity designs which take advantage of strict voting eligibility requirements – e.g. to test whether two similar people born days apart (91) vote more in the future, if one got a lucky chance to vote before while the other person did not.

A challenge, as pointed out by (92), is that these designs often suffer from weak identification of short-run and long-run effects. For example, if an inducement to focus on the treated election is focused on encouraging people to "do their civic duty," this effect of social pressure may endure into the next election, independent of habit formation. Similarly, the early inducement may lead to increased interest in politics, which then causes the later turnout.

More recent work has acknowledged that behavior alone is not enough to label an action as habitual, citing the psychology literature on automaticity and context-sensitivity as inspiration for creating a self-report voting habit index akin to the SRHI. (85) developed a 7-item scale for "voter turnout habit" and validates it using UK and US voting data. He argues that the "cost" of voting (93) will be lower when voting becomes habitual.

Other papers have looked at the consistency of environmental context voting behavior by looking at voting rates following a change in home address or voting location address. This approach is a special case of our general focus on PCS except for a narrow range of context variables and a long time between behaviors (and unfortunately, also a change in cost).

For example, (94) found that the consolidation of voting precincts in Los Angeles country decreased overall turnout substantially (which was partially, but not fully, offset with an increase in absentee votes). This change is consistent with the hypothesis that removal of the environmental cue of the physical precinct deterred some individuals from voting. (95) found that both self-reported previous voting and not moving (situational consistency) were associated with voting. Research into other contextual cues, like time of day, which may be predictive of voting behavior has been more limited (85).

# 2 Dataset Descriptions

The purpose of this section is to provide additional detail on the two main datasets used in this paper, along with a full list of the context variables which were used to train the LASSO models.

## 2.1 Hand Washing Data

Hand-hygiene data came from Proventix, a company which uses RFID technology to monitor whether a healthcare provider sanitized their hands during a hospital shift. The initial dataset tracks 5,246 hospital healthcare workers across 30 different hospitals. The dataset spans about a year, with over 40 million data points, each corresponding to whether an individual did or did not wash their hands. Each data point has a timestamp, room, and hospital location.

We further infer several other attributes, such as time of day and individual-level variables such as whether the healthcare worker complied (washed their hands) in this room previously. A full list of the variables that are used follows in Section S2.3.1.

## 2.2 Gym Attendance Data

We obtain check-in data from a North American gym chain, containing information for 60,277 regular gym users across 560 gyms. The data spans fourteen years, from 2006 to 2019. There were initially over 12 million data points, each corresponding to one gym check-in. Each data point is accompanied by a timestamp, gym location, and other information about the gym (such as the number of amenities and wi-fi availability, which we do not use in this analysis).

We further infer several other attributes, such as the day of the week and individual-level variables such as the time since gym membership creation. A full list of the variables that are used follows in Section S2.3.2.

## 2.3 Description of Context Variables

### 2.3.1 Hand washing data

- **Time at work:** minutes elapsed since the start of a person's shift.
- **Rooms visited in shift:** number of rooms the caregiver had visited previously during the shift.
- **Compliance last opportunity:** an indicator variable of whether the caregiver washed her hands at the last opportunity.
- **Time since last opportunity (mins):** minutes elapsed since the last opportunity.
- **Time since last compliance (mins):** minutes elapsed since the last compliance.
- **Frequency of patient encounter:** percentage of time in patient rooms as a fraction of time worked. At any moment in the shift, this is defined as $\frac{\text{cumulative time spent in patient room}}{\text{cumulative time elapsed in shift}}$.
- **Entry indicator (0-1):** an indicator of whether the opportunity to wash is an entry (1) into a room (as opposed to an exit (0) from a room).
- **Previous unit compliance:** average compliance (%) across previous shifts in the current hospital unit.
- **Unit frequency:** % of previous shifts in the current hospital unit.
- **Previous day-of-week compliance:** average compliance (%) across previous shifts in the current day of week.
- **Day-of-week frequency:** % of previous shifts in current weekday (compared to other weekdays).
- **Previous room compliance:** average compliance (%) across previous shifts in the current room.

- **Room frequency:** % of time spent working in current room (compared to other rooms in the same hospital).
- **Room compliance of others:** average compliance rate (%) of other caregivers in the current room.
- **Compliance last shift:** compliance rate in the last shift before the current one.
- **Days since start:** number of days worked since the observed start date.
- **Time off:** hours elapsed between end of the last shift and the current shift.
- **Streak:** number of consecutive shifts with less than 36 hours apart.
- **Hour-slot fixed effects:** time of day is divided into four categories: 12am-6am, 6am-12pm, 12pm-6pm, and 6pm-12am.
- **Compliance within a room:** an indicator of whether the caregiver washed her hands in this room in the current opportunity (e.g. if she washed upon entry, this variable value for the exit opportunity is equal to 1).
- **Month of the year.**

#### 2.3.2 Gym attendance data

- **Streak:** number of consecutive days with gym visits prior to the current day.
- **Day-of-week streak:** number of consecutive corresponding day-of-the-week gym visits prior to the current day.
- **Time lag:** number of days since the last gym visit.
- **Attendance last 7 days:** number of gym visits during the last 7 days.
- **Month of the year.**
- **Day of the week.**

## 3 Analysis Details

The purpose of this section is to provide additional detail on our analysis methodology. Specifically, we provide a formal description of the LASSO models and include a discussion of the model output (predictability) vs. a traditional measure of habit (frequency). We then provide a formal description of the exponential model used to fit the behavioral data to identify speed of habit formation, and discuss model.

## 3.1 Individual LASSO Regressions

We apply LASSO logistic regressions at the individual level. LASSO is an acronym for "Least Absolute Shrinkage and Selection Operator". LASSO is good for our purposes because it can improve out-of-sample predictive accuracy by reducing variance without significantly increasing bias. It also useful for feature selection via shrinking many insignificant variable coefficients towards 0. This property winnows down a large set of variables to smaller subsets.

For each individual, we select about 15% of their time series data as a holdout ("test") set on which we will assess the performance of the model. For the remaining ("training") data, we train the model based on the following logit specification:

$$\mathbb{P}(Y_t = 1) = \frac{\exp(\beta_0 + \mathbf{S}_t\beta_1)}{1 + \exp(\beta_0 + \mathbf{S}_t\beta_1)},$$

where $t$ indexes time, $Y_t$ is the binary outcome variable indicating whether a habit was executed at time $t$, and $\mathbf{S}_t$ is a vector of state variables. The list of state variables $\mathbf{S}_t$ used in each dataset can be found above in Section 2.3. The LASSO includes a penalty term weighting the sum of absolute values of coefficients $\|\beta_1\|_1$ by the tuning parameter $\lambda$. The coefficients $\hat{\beta}_1$ are chosen to minimize the following loss function:

$$L(\beta \mid \lambda) = -\log\left[\prod_{Y_t=1} \frac{\exp(\beta_0 + \mathbf{S}_t\beta_1)}{1 + \exp(\beta_0 + \mathbf{S}_t\beta_1)} \prod_{Y_t=0} \frac{1}{1 + \exp(\beta_0 + \mathbf{S}_t\beta_1)}\right] + \lambda \|\beta_1\|_1 .$$

As is standard with machine learning applications, we use stratified 5-fold cross validation to pick the optimal $\lambda$. In particular, the holdout set and the 5 folds used in cross-validation are selected such that the proportions of observations with $Y_t = 1$ in each of them are the same.

To ensure reasonable performance of the LASSO model, we omit from the analytical sample people with too few observations (fewer than 365 for the gym data and fewer than 1000 for the hand-hygiene data), and with unbalanced habit execution rates ($< 5\%$ or $> 90\%$ for gym attendance, and $< 20\%$ or $> 80\%$ for hand-hygiene). Methods like LASSO are known to not estimate and predict well with small samples or unbalanced samples of binary outcomes.

We report the summary statistics of the full sets of LASSO coefficients in Tables S2 and S3 below.

**Table S2: Context Predictors of Gym Attendance**

*Summary statistics for the context cue variables for the individual LASSO models, sorted by variable importance (sometimes called "feature importance" by machine learning scholars). Importance is measured by averaging the absolute values of the standardized LASSO coefficients across individuals. The Q1, Median, and Q3 columns present the first, second, and third quartile coefficient values for the sample. The columns % zero, % positive, and % negative are the percentage of the individual LASSO models that had coefficients with zero, positive, and negative values, respectively. For more detailed descriptions of the context predictors, see S2.3.2.*

| | Variable importance | Q1 | Median | Q3 | % zero | % positive | % negative | General predictive effect |
|---|---|---|---|---|---|---|---|---|
| Time lag | 1.25 | −1.40 | -0.34 | -0.02 | 22 | 2 | 76 | 74 |
| (Time lag)$^2$ | 0.92 | 0.00 | 0.00 | 0.86 | 57 | 39 | 3 | 36 |
| Monday | 0.36 | 0.00 | 0.11 | 0.50 | 32 | 57 | 11 | 46 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tuesday | 0.35 | 0.00 | 0.10 | 0.49 | 33 | 56 | 11 | 45 |
| Wednesday | 0.34 | 0.00 | 0.06 | 0.46 | 35 | 54 | 12 | 42 |
| Attendance last 7 days | 0.34 | 0.09 | 0.29 | 0.47 | 9 | 82 | 8 | 74 |
| Thursday | 0.31 | 0.00 | 0.00 | 0.40 | 37 | 49 | 14 | 35 |
| Friday | 0.28 | 0.00 | 0.00 | 0.27 | 36 | 39 | 24 | 15 |
| Day-of-week streak | 0.23 | 0.00 | 0.11 | 0.30 | 25 | 69 | 7 | 62 |
| Streak | 0.22 | 0.00 | 0.00 | 0.14 | 36 | 40 | 24 | 16 |
| Saturday | 0.22 | -0.04 | 0.00 | 0.15 | 35 | 36 | 29 | 7 |
| $(\text{Streak})^2$ | 0.15 | -0.13 | 0.00 | 0.00 | 46 | 13 | 42 | 29 |
| $(\text{Day-of-week streak})^2$ | 0.13 | -0.16 | 0.00 | 0.00 | 48 | 9 | 43 | 34 |
| December | 0.11 | -0.05 | 0.00 | 0.00 | 47 | 16 | 38 | 22 |
| January | 0.10 | 0.00 | 0.00 | 0.05 | 45 | 39 | 16 | 23 |
| July | 0.09 | 0.00 | 0.00 | 0.01 | 48 | 27 | 26 | 1 |
| August | 0.09 | 0.00 | 0.00 | 0.00 | 48 | 27 | 25 | 2 |
| September | 0.09 | -0.01 | 0.00 | 0.00 | 49 | 25 | 27 | 2 |
| October | 0.09 | -0.01 | 0.00 | 0.00 | 49 | 22 | 29 | 7 |
| November | 0.09 | -0.02 | 0.00 | 0.00 | 49 | 21 | 31 | 10 |
| February | 0.08 | 0.00 | 0.00 | 0.02 | 48 | 31 | 21 | 10 |
| March | 0.08 | 0.00 | 0.00 | 0.01 | 49 | 29 | 22 | 7 |
| April | 0.08 | 0.00 | 0.00 | 0.01 | 49 | 29 | 22 | 7 |
| May | 0.08 | 0.00 | 0.00 | 0.00 | 49 | 26 | 25 | 1 |
| June | 0.08 | 0.00 | 0.00 | 0.01 | 48 | 29 | 23 | 6 |

**Table S3: Context Predictors of Hospital Hand Washing**

*Summary statistics for the context cue variables (including interactions) for the individual LASSO models, sorted by variable importance. Importance is measured by averaging the absolute values of the standardized LASSO coefficients across individuals. Q1, Median, and Q3 columns are the coefficients at the first (lowest), second, and third quartiles of the sample. % zero, % positive, and % negative are the percentage of individual LASSO models which had coefficients that had zero, positive, and negative values, respectively. For more detailed descriptions of the context predictors, see Supplement Materials S2.3.*

| | Variable importance | Q1 | Median | Q3 | % zero | % positive | % negative | General predictive effect |
|---|---|---|---|---|---|---|---|---|
| Compliance last shift | 0.77 | 0.66 | 0.70 | 0.92 | 0 | 100 | 0 | 100 |
| Entry indicator | 0.35 | -0.33 | -0.28 | -0.04 | 18 | 5 | 77 | 72 |
| Compliance last opp.×Entry indicator | 0.13 | 0.00 | 0.00 | 0.21 | 49 | 47 | 4 | 43 |
| Compliance last opp.×Time since last opp. | 0.12 | 0.00 | 0.00 | 0.00 | 54 | 1 | 45 | 44 |
| Compliance within a room | 0.12 | 0.00 | 0.01 | 0.14 | 33 | 51 | 16 | 35 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Time since last opp. | 0.09 | 0.00 | 0.00 | 0.00 | 61 | 24 | 15 | 9 |
| (Time since last opp.)$^2$ | 0.08 | 0.00 | 0.00 | 0.00 | 74 | 7 | 18 | 11 |
| Room compliance of others | 0.08 | 0.04 | 0.05 | 0.12 | 32 | 66 | 2 | 64 |
| Time at work | 0.08 | 0.00 | 0.00 | 0.00 | 54 | 4 | 42 | 38 |
| Compliance last opp.×(Time since last opp.)$^2$ | 0.07 | 0.00 | 0.00 | 0.00 | 74 | 20 | 5 | 15 |
| Prev. room compliance | 0.07 | 0.03 | 0.04 | 0.11 | 32 | 65 | 2 | 63 |
| Compliance last opp. | 0.05 | 0.00 | 0.00 | 0.07 | 47 | 45 | 7 | 38 |
| Time at work×6am-12pm | 0.05 | 0.00 | 0.00 | 0.00 | 78 | 10 | 12 | 2 |
| Time since last compliance | 0.05 | 0.00 | 0.00 | 0.00 | 64 | 9 | 27 | 18 |
| Time at work×12pm-6pm | 0.04 | 0.00 | 0.00 | 0.00 | 73 | 10 | 17 | 7 |
| (Time since last compliance)$^2$ | 0.03 | 0.00 | 0.00 | 0.00 | 75 | 17 | 8 | 9 |
| 12am-6am | 0.03 | 0.00 | 0.00 | 0.00 | 68 | 22 | 10 | 12 |
| Frequency of patient encounter | 0.03 | 0.00 | 0.00 | 0.01 | 58 | 31 | 12 | 19 |
| Time at work×Patient encounter | 0.03 | 0.00 | 0.00 | 0.00 | 64 | 8 | 28 | 20 |
| Days since start | 0.02 | 0.00 | 0.00 | 0.00 | 83 | 9 | 8 | 1 |
| 6am-12pm | 0.02 | 0.00 | 0.00 | 0.00 | 80 | 7 | 13 | 6 |
| 12pm-6pm | 0.02 | 0.00 | 0.00 | 0.00 | 77 | 12 | 11 | 1 |
| Room frequency | 0.02 | 0.00 | 0.00 | 0.00 | 63 | 19 | 19 | 0 |
| Time at work×6pm-12am | 0.02 | 0.00 | 0.00 | 0.00 | 82 | 10 | 7 | 3 |
| (Time off)$^2$ | 0.01 | 0.00 | 0.00 | 0.00 | 84 | 8 | 8 | 0 |
| October | 0.01 | 0.00 | 0.00 | 0.00 | 81 | 10 | 9 | 1 |
| November | 0.01 | 0.00 | 0.00 | 0.00 | 82 | 10 | 8 | 2 |
| December | 0.01 | 0.00 | 0.00 | 0.00 | 81 | 10 | 9 | 1 |
| March | 0.01 | 0.00 | 0.00 | 0.00 | 82 | 9 | 10 | 1 |
| April | 0.01 | 0.00 | 0.00 | 0.00 | 80 | 10 | 11 | 1 |
| May | 0.01 | 0.00 | 0.00 | 0.00 | 80 | 9 | 10 | 1 |
| June | 0.01 | 0.00 | 0.00 | 0.00 | 80 | 10 | 10 | 0 |
| July | 0.01 | 0.00 | 0.00 | 0.00 | 79 | 11 | 10 | 1 |
| August | 0.01 | 0.00 | 0.00 | 0.00 | 78 | 11 | 11 | 0 |
| September | 0.01 | 0.00 | 0.00 | 0.00 | 82 | 9 | 9 | 0 |
| Day-of-week frequency | 0.01 | 0.00 | 0.00 | 0.00 | 77 | 13 | 10 | 3 |
| Rooms visited in shift | 0.01 | 0.00 | 0.00 | 0.00 | 83 | 8 | 9 | 1 |
| 6pm-12am | 0.01 | 0.00 | 0.00 | 0.00 | 84 | 7 | 8 | 1 |
| Prev. day-of-week compliance | 0.01 | 0.00 | 0.00 | 0.00 | 79 | 8 | 13 | 5 |
| Prev. unit compliance | 0.01 | 0.00 | 0.00 | 0.00 | 78 | 10 | 13 | 3 |
| Streak | 0.01 | 0.00 | 0.00 | 0.00 | 78 | 8 | 15 | 7 |
| Time off | 0.01 | 0.00 | 0.00 | 0.00 | 85 | 8 | 7 | 1 |
| Unit frequency | 0.01 | 0.00 | 0.00 | 0.00 | 72 | 20 | 8 | 12 |
| February | 0.00 | 0.00 | 0.00 | 0.00 | 82 | 8 | 9 | 1 |

## 3.2 Model Selection Challenges in LASSO

LASSO is used to choose variables that are predictive (which we call "predictors"). It is well-known that the derived coefficients of those predictors in LASSO will be different from the estimates of coefficients for the same variables derived from OLS or logit regressions. This is because LASSO includes what is called a "regularization penalty"– the sum of squared residuals is added to a weight times the absolute magnitude of the coefficients.[6]

This difference between the OLS or logit estimate for a variable, and the LASSO coefficient for the same variable, is an inevitable consequence of LASSO trying to balance the "bias-variance tradeoff". That is, in LASSO the derived predictor coefficients are biased away from their true values– they are typically "shrunk" toward zero to reduce the regularization penalty for high values of $\hat{\beta}$. However, this bias is helpful for prediction because it reduces the variance of the coefficient estimates, which helps to reduce in-sample overfitting. This property in turn guards against a big drop from good in-sample fits to much poorer out-of-sample predictive fits.

The fact that LASSO is estimating variable coefficients with bias creates two challenges which are special to LASSO and related regularization methods. These challenges are called "model selection consistency" and "stability".

Model selection consistency is the answer to this question: If a variable W is truly part of the data-generating process, will LASSO choose W as a nonzero variable? If the answer is yes, the model selection is consistent (i.e., similar across OLS or logit and LASSO). There are conditions under which LASSO is consistent in this sense Zhao and Yu 96. However, these conditions are asymptotic so they do not tell us anything very informative about finite samples. Furthermore, there is no procedure for estimating standard errors of LASSO coefficients without very restrictive assumptions (although Bayesian LASSO can produce credible intervals, Park and Casella 97).

Without standard errors, it is difficult to know how closely the predictors that are selected by LASSO are to the true coefficients that are generating behavior. Some true variables may be mistakenly regularized to zero by LASSO. This consistency concern is why strive to be careful in the text by talking about "predictors" rather than "variables".

"Stability" refers to a different consequence of high collinearity between variables that is special for LASSO predictor estimation (98, 99) for discussion). High collinearity is always a problem but it has a special consequence in LASSO. If two variables $W$ and $Z$ are correlated highly enough, LASSO will typically choose only one of the $W$ or $Z$ variables to have a nonzero value. This is because the squared residuals are not reduced much differently by including both variables (because they are contributing shared variance in explaining the dependent variable), but the regularization penalty increases if both variables are included because the penalty would be a multiple of $|\hat{\beta}_W| + |\hat{\beta}_Z|$ rather than, for example, a penalty of $|\hat{\beta}_W|$ if only W is included and $\hat{\beta}_Z = 0$ so that the $Z$ LASSO coefficient is zero. As a result, if both $W$ and $Z$ are included in the feature set of candidate variables, then Which variable is chosen can vary arbitrarily, meaning that it changes depending on small changes in the sampled values of

[6]It is crucial in doing this that the variables be standardized in some way, typically by subtracting the variable sample mean and dividing by the variable sample standard deviation. Otherwise, variables which happen to be scaled to create large coefficient values are unfairly penalized.

$W$ and $Z$. Keep in mind, however, that stability in this sense is similar to what happens in OLS: In OLS or logit both $W$ and $Z$ variables will have estimated coefficients, but their standard errors are inflated and there can be other iases in the correlated parameter estimates.)

For readers who are unfamiliar with LASSO, the most important part of this discussion, by far, is what you have just read: *Because of model selection inconsistency, and stability, the set of nonzero LASSO variables and their predictor coefficients is* not *guaranteed to overlap with the true coefficient values.* This difference is especially crucial, as noted throughout the main text of our paper, when one is trying to make statistical judgments in comparing coefficient values. For example, suppose one person in our gym attendance sample has a positive effect of a Monday dummy variable on gym attendance, and another person has a negative effect on the same Monday dummy variable. There is no statistical procedure to test whether those effects are significantly different between the two people, because the LASSO coefficients do not have standard errors. Therefore, any statements about heterogeneity do not have the usual backing of significance testing. It is still a true statement, and one that may be of some exploratory or managerial value, that one predictor LASSO coefficient is positive and one is negative. But we cannot have any statistical confidence that the coefficient signs are different, or any measure of how close the coefficients are compared to their variances (standard errors).

Next we describe some procedures for trying to evaluate how badly the stability problem (arising from collinearity) might be undermining our conclusions in our specific data. These procedures were created by us in response to reader concerns, as an attempt to explore numerically whether consistency and stability are big or small problems. These analytical procedures to evaluate impact of stability may be less useful, or even misleading, for readers who use LASSO type methods, as we have, to learn about predictive coefficients. (Readers are therefore encouraged to learn more and be cautious in using predictor coefficient estimates, or explore Bayesian credible intervals.)

- Since high correlation of pairs of variables $W$ and $Z$ is the biggest threat to stability, we first computed a two-variable correlation matrix for each person. Then we picked out pairs of variables which had the highest correlations for a large number individuals. These are pairs of variables for which the median absolute value of correlation across individuals is at least 0.4. By using the median, our method finds variable pairs that are substantially correlated for at least half of the individuals. Moreover, we only look at pairs of variables where there is at least one "important" variable (which is defined by a large percentage of non-zero coefficients and an imbalance toward either negative or positive signs, as reported in Tables 2 & 4 in the main text).

- For each pair of such correlated variables, we separate all individuals into two groups based on a median split of the absolute value of correlation. The two groups are individuals with variable correlations above and below the median.

- We identified whether each of the two variables was included with a nonzero magnitude or was regularized to zero. This procedure creates four possible selection outcomes of nonzero and zero magnitudes for each of the two variables.

- For each pair of variables, we looked at the percentage of each of the four possible selection outcomes between

| | Best case scenario | Two distinct alternative best case scenarios (one of the two variables is more predictive) | | Worst case scenario |
|---|---|---|---|---|
| Coefficient combinations | Both variables or no variables selected | Variable 1 is more predictive and selected | Variable 2 is more predictive and selected | Variables are arbitrarily selected |
| $(\hat{\beta}_1 \neq 0, \hat{\beta}_2 \neq 0)$ | x% | 0 | 0 | 0 |
| $(\hat{\beta}_1 \neq 0, \hat{\beta}_2 = 0)$ | 0 | 100% | 0 | 50% |
| $(\hat{\beta}_1 = 0, \hat{\beta}_2 \neq 0)$ | 0 | 0 | 100% | 50% |
| $(\hat{\beta}_1 = 0, \hat{\beta}_2 = 0)$ | 100-x% | 0 | 0 | 0 |

**Table S4: Stylized LASSO model stability outcomes for two variables that have true positive values**

the low and high correlation groups. If there is high instability there will be similar numbers of non-zero coefficient estimates for both of the variables W and Z in a pair, and that pattern will be more pronounced for individuals with pairwise correlations larger than the median

Table S4 illustrates two extreme examples of how highly correlated variables, which both have true positive $\beta$ values, might or might not be arbitrarily selected. In the "best case scenario", it is never the case that one of the variables is selected (non-zero) and the other is regularized to zero. Either both variables are selected or neither are. In the "worst case scenario" only one variable is arbitrarily selected; in half the individualregressions it is the first variable that is selected while in the other half it is the second variable. First, we report the results for the 5 pairs of variables in the 2 datasets in Tables S5 & S6 for which between-variable correlations (averaged over individual LASSOs) are above .40. Note that the interquartile range for these correlations are around (.40,.60), for gym data, or (.20,.60), for two variable pairs in the handwashing data. This is somewhat reassuring because it means that only a quarter of the individuals have collinearity .60 or higher.

The first thing to look for is whether the variable selection combinations look more like the best case or worst case scenarios. The results are far from the worst case in which there are high percentage of both $(\hat{\beta}_1 \neq 0, \hat{\beta}_2 = 0)$ and $(\hat{\beta}_1 = 0, \hat{\beta}_2 \neq 0)$ classifications (as if LASSO is arbitrarily selecting only one of the two correlated variables). Instead, the results look like a mixture of the three best case scenarios. When only one of the two variables is selected, it is typically the same variable across individuals.

Second, if the LASSO is indeed selecting variables arbitrarily due to high multicollinearity, then we expect the percentage of $(\hat{\beta}_1 \neq 0, \hat{\beta}_2 \neq 0)$ outcomes will go down as we go from the low correlation to the high correlation groups (and consequently, there will be an increase in the other three of the four categories). However, the results from Tables S5 & S6 allay this concern, because we do not see such general patterns for the pairs. Indeed, when there is a change across correlation magnitude columns, it goes weakly in the opposite direction for four out of five pairs. For example, Table S6 the leftmost pair percentages go up strongly, from 14% to 80%. Furthermore, there is not a general convergence toward equal frequencies of outcomes with exactly one zero variable selected for each of the two variables in a pair, when correlations are higher. For example, in Table S5 the percentages for whom only

one variable is nonzero (rows 2 and 3 for the first two data columns) is 2% and 16% for lower correlation and 6% and 18% for higher correlations.

There is one odd result based on strength of collinearity $\rho$, in the first part of Table S6 for the two variables (Compliance last shift, Days since start). When the correlation is low ($|\rho| < .43$) the first variable is uniquely selected most often (i.e., $\hat{\beta}_2 = 0$). However, when the correlation is high ($|\rho| < .43$) both variables are selected most often. We do not have a good explanation for this, but also note that it is the only odd result for the five correlated variable pairs that are analyzed and reported.

While we did not exhaust all possible correlation structures between the variables, we believe the procedure laid above indicates that the LASSO stability problem associated with correlated observed variables, if exists, is not that severe in our data.

| | ($X_1$=Time lag, $X_2$=Attendance last 7 days) $(Q1, Q2, Q3) = (.40, .50, .59)$ | | ($X_1$=Streak, $X_2$=Attd. last 7 days) $(Q1, Q2, Q3) = (.46, .52, .57)$ | |
|---|---|---|---|---|
| | $\|\rho\| \leq 0.50$ | $\|\rho\| > 0.50$ | $\|\rho\| \leq 0.52$ | $\|\rho\| > 0.52$ |
| $(\hat{\beta}_1 \neq 0, \hat{\beta}_2 \neq 0)$ | 78% | 70% | 60% | 60% |
| $(\hat{\beta}_1 \neq 0, \hat{\beta}_2 = 0)$ | 2% | 6% | 3% | 6% |
| $(\hat{\beta}_1 = 0, \hat{\beta}_2 \neq 0)$ | 16% | 18% | 34% | 28% |
| $(\hat{\beta}_1 = 0, \hat{\beta}_2 = 0)$ | 4% | 6% | 3% | 6% |

**Table S5: Behavior of LASSO Coefficients Across Pairs of Correlated Variables - Gym Data.**
*For each pair of variables $X_1, X_2$, this table reports the proportions the pair being in one of the four selection categories, broken down by low & high correlation subgroups. We also report the interquartile range $(Q1, Q2, Q3)$ (the lowest 25% boundary, median, and highest 25% boundary) for the absolute value of correlation $|\rho|$.*

| | (Compl. last shift, Days since start) $(Q1, Q2, Q3) = (.21, .43, .65)$ | | (Compl. last shift, Rooms visited) $(Q1, Q2, Q3) = (.19, .41, .62)$ | | (Within ep. compl., Compl. last opp.) $(Q1, Q2, Q3) = (.50, .56, .61)$ | |
|---|---|---|---|---|---|---|
| | $\|\rho\| \leq 0.43$ | $\|\rho\| > 0.43$ | $\|\rho\| \leq 0.41$ | $\|\rho\| > 0.41$ | $\|\rho\| \leq 0.56$ | $\|\rho\| > 0.56$ |
| $(\hat{\beta}_1 \neq 0, \hat{\beta}_2 \neq 0)$ | 14% | 80% | 14% | 21% | 33% | 51% |
| $(\hat{\beta}_1 \neq 0, \hat{\beta}_2 = 0)$ | 86% | 20% | 86% | 79% | 25% | 8% |
| $(\hat{\beta}_1 = 0, \hat{\beta}_2 \neq 0)$ | 0% | 0% | 0% | 0% | 13% | 25% |
| $(\hat{\beta}_1 = 0, \hat{\beta}_2 = 0)$ | 0% | 0% | 0% | 0% | 29% | 16% |

**Table S6: Behavior of LASSO Coefficients Across Pairs of Correlated Variables - Hand-washing Data.**
*For each pair of variable, this table reports the proportions the pair being in one of the four selection categories, broken down by low & high correlation subgroups. We also report the interquartile range $(Q1, Q2, Q3)$ for the absolute value of correlation $|\rho|$*

Because LASSO is selecting predictors, and not creating unbiased estimates of coefficients, there can be large differences in which predictors are selected even for different subsamples of an entire sample. Any such differences are likely to be indicators of potential model selection inconsistency. In other words, except for sampling error, if a variable has a true positive coefficient in the data generating process, then it should be selected in most subsamples with a positive sign. If it is sometimes selected and sometimes has a predictor coefficient of zero, then it is possible it might have a zero coefficient in the full sample. That would be a textbook model of misleading model selection *in*consistency. This challenge is dramatically illustrated by an example in (99) (Figure 2 and pg. 96-98). They took the prices of 51,808 houses from the a 2011 American Housing Survey, and recorded 150 features for each house (such as base area, number of rooms, and census region information). In one analysis, they partitioned the entire sample into 10 subsamples of about 5,000 houses each. In each of the 10 subsamples they ran a separate LASSO. This created ten vectors of 150 feature predictive coefficients.

One would hope that these ten feature vectors would be highly correlated; but they were not. For each of the ten subsamples the derived sets of nonzero predictors were not highly overlapping (see their Figure 2).

As they wrote (pg. 96)

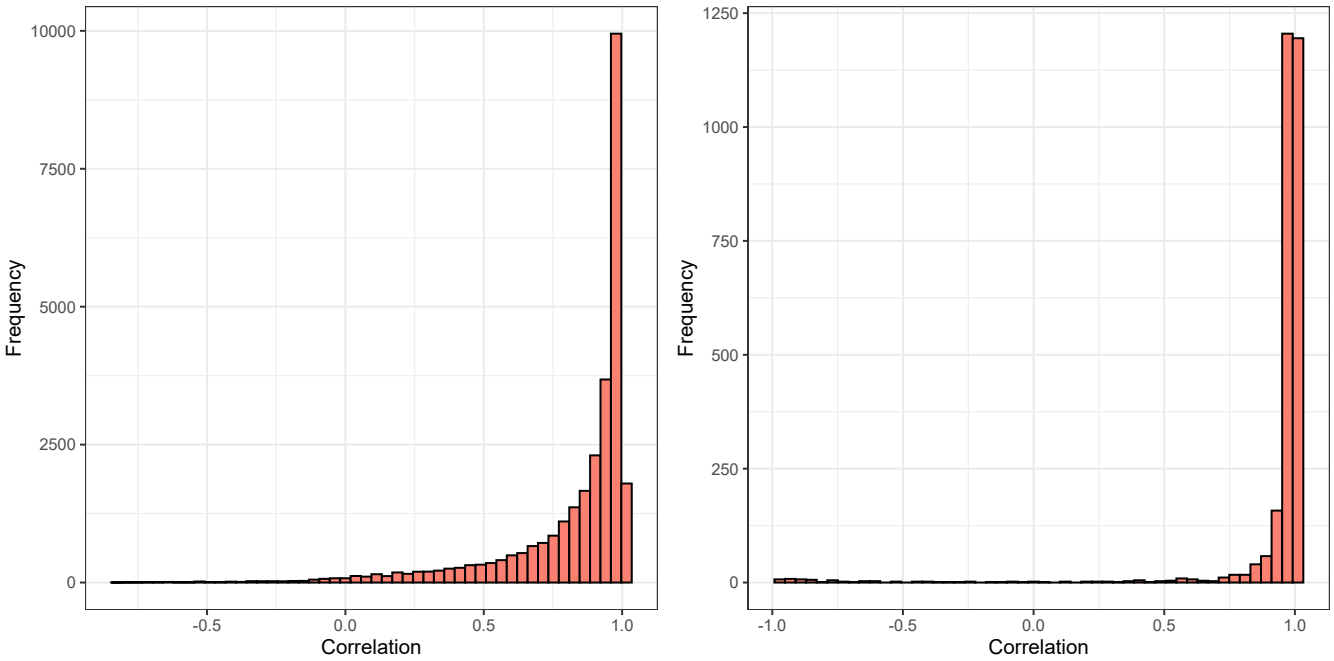
> Figure 2 documents a fundamental problem: a variable used in one partition [subsample] may be unused in another. In fact, there are few stable patterns overall.

They note that while the selected predictors vary, the overall predictability is about the same in each subsample. The stability problem in their analysis arises (pg. 96-97)

> ...because the variables are correlated with each other (say the number of rooms of a house and its square-footage), then such variables are substitutes in predicting house prices. Similar predictions can be produced using very different variables. Which variables are actually chosen depends on the specific finite sample.

We use an analogue to their subsample exercise to see whether highly variable predictor selection also occurs in our data. For each individual, their entire sample is split into two distinct subsamples (based on odd and even numbered data rows).

Two LASSOs are estimated, one for each of the two subsamples. Each LASSO produces a vector of predictor values. The overlap or correlation in predictor values between the two subsamples is measured in two ways. For each individual, the correlation between LASSO coefficients in the two subsamples is computed. Histograms of those correlations across individuals are shown in Figure S1. The mean correlation for the gym data is .824 and the quartiles are (Q1,Q2,Q3) = (.775, .933, .983). For the hand-washing data, the mean correlation is .933 and the quartiles are (Q1,Q2,Q3) = (.973, .990, .995). The correlations are generally very high. This means that the variation in subsamples observed in the housing data by (99) does not occur in these data.




**(a)** Gym data

**(b)** Hand-washing data

**Figure S1:** Distribution of correlations between LASSO coefficients of the odd-numbered and even-numbered dates (or shifts) subsamples.

The second way to measure overlap in predictor values between the two subsamples is to code the selected variables as 0 or non-zero and compute the percentage which are estimated into the same category in the two subsamples. This type of categorization is what (99) seem to be referring to as "few stable patterns overall" in their comparison of ten subsamples in their Figure 2.

These statistics are summarized in our analysis in Table S7. The table reports overlap subsamples (Odd, Even) and for each subsample with the full sample (where the overlap is naturally higher, because the subsample contains the full sample). The mean and median overlap are 63% and 68%. However, the upper quartile of overlap is 84%; so that for a substantial percentage of subjects the overlap appears to be quite high. Furthermore, the fact that most correlation coefficients are around .80-.95 (recall Figure S1) indicates that when coefficient categories do mismatch– a coefficient is zero in one subsample and nonzero in the other– the nonzero coefficient is likely to be small in magnitude. (Otherwise the mismatch would pull down the correlation a lot.) That is why the correlations of coefficient values between subsamples are so close to 1 even when the percentage of categorical overlap is not so high.

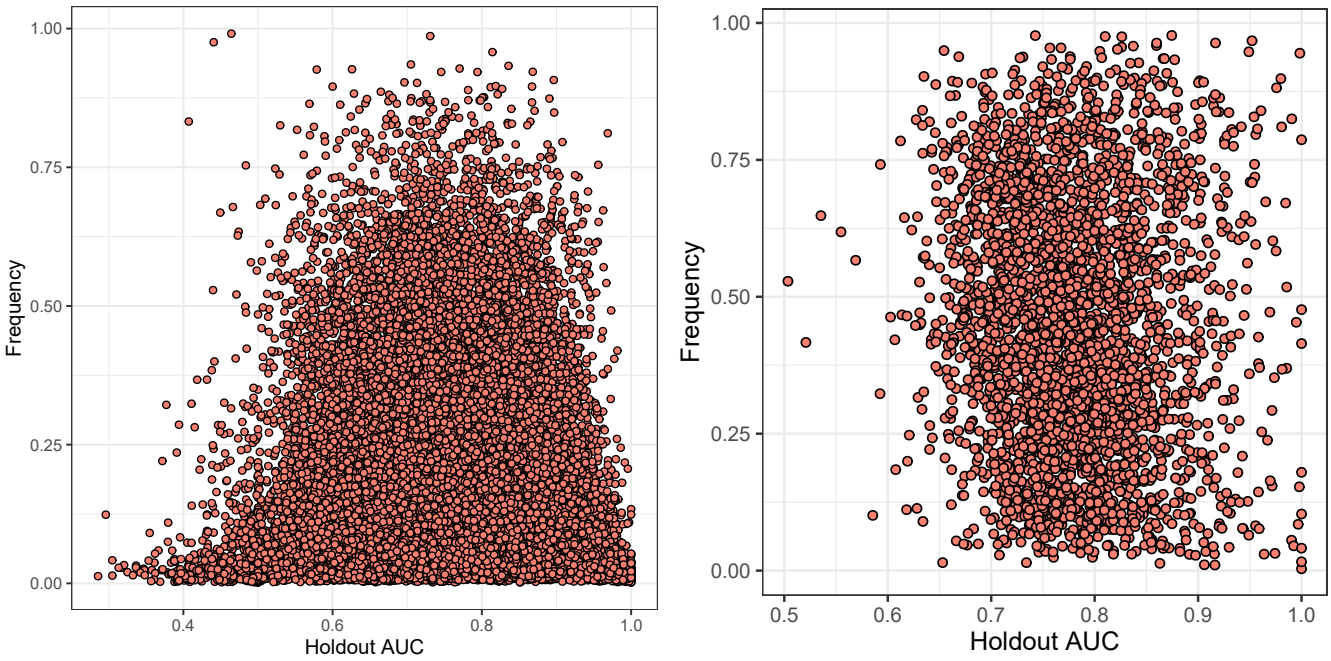
| | Gym-data | | | | Hand-washing data | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Q1 | Median | Q3 | Mean | Q1 | Median | Q3 |
| (Odd, Even) | 63% | 48% | 68% | 84% | 74% | 66% | 75% | 86% |
| (Even, Full) | 71% | 56% | 80% | 88% | 78% | 68% | 82% | 91% |
| (Odd, Full) | 71% | 56% | 80% | 88% | 79% | 70% | 82% | 91% |

**Table S7:** Percentage of matches between the coefficients of two subsamples across individuals.

Furthermore, the average number of data points for individuals with an above-median match rate (*i.e.*, $\geq 68\%$) is 947, while that for individuals with below median match is 759. Therefore, there is modest evidence that, as one might expect, with larger samples the match rate improves so that the subsample stability problem is smaller in magnitude. In general, if an analyst is considered about model selection consistency and stability, those potential problems are likely to be partly alleviated in larger samples.

### 3.3 AUC vs Frequency

Figure S2 plots the relationship between holdout AUC and frequency of behavioral execution for each individual in the two datasets. There is no visually evident relationship between the two, and the correlations are -0.11 and -0.06. This lack of correlation between frequency and predictability (AUC) highlights the importance of the latter as a novel measure of habit rather than relying only on behavioral frequency and other measures.




**(a)** Gym data ($r = -0.11$)

**(b)** Hand-washing data ($r = -0.06$)

**Figure S2: Predictability (AUC) versus Frequency of Behavior.**

*This figure plots the relationship between holdout AUC and outcome frequency for the two data analyzed in this paper.*

## 3.4 Speed of Habit Formation

As we discussed in the literature review section, even though the speed of habit formation is an interesting phenomenon of practical value, the evidence about the speed of habit formation is thin and far from conclusive. Most studies have typically relied on self-report measures and automaticity scales (2, 100). However, given that habitual behavior might, as part of its essence, be accompanied by degraded memory, self-report measures are not ideal. Specifically, if people do not have much awareness about how strongly their behaviors are cued by context, they may misattribute habits to volition (32–34). People may also misconstrue items on self-report scales (36) and reflecting on the characteristics of behavioral performance may lead them to report behavioral frequency rather than habit symptoms. We seek to avoid errors induced by self-report by taking advantage of our granular observational data and using the predictability measurement AUC as a proxy of the strength of habit.

Our approach is to *define* habit formation as increasing predictability of activity from contextual state variables. We fully understand there is more to habit formation than increased predictability (e.g., motor automaticity) but the concept of increased predictability can be explored in our data. We therefore look for habit formation in the specific form of increasing sequence of AUCs (predictability) over time.

To illustrate this method with the gym data, the data for an individual are partitioned into 2-week windows. A LASSO model is trained in every 4-week period and used to make predictions for the subsequent 2 weeks.[7] Described

[7] We are grateful to an anonymous referee for this suggested improvement over an inferior estimation method used in our first version.

in notation, suppose we observe a gym member $i$ for a total sample of $W_i$ weeks. We start by using weeks $1-4$ to train and weeks $5-6$ to test. This procedure gives a single test-period AUC. The next step uses data from weeks $3-6$ to train and weeks $7-8$ to test. The next step after that slides the window of both training and test ahead two weeks. It uses data from weeks $5-8$ to train and weeks $9-10$ to test, and so forth. In general, For every $w \in 3, 4, \ldots, W_{i-5}$, we define the "habit level" at week $2w$ as the AUC value obtained from predicting gym attendance in weeks $\{2w-1, 2w\}$ using a LASSO model trained on data from weeks $\{2w-5, 2w-4, 2w-3, 2w-2\}$. This process generates a sequence of $\left(\left\lfloor \frac{W_i}{2} \right\rfloor - 2\right)$ AUC values. For the hand-hygiene data, we partition the data into 2-shift windows, rather than 2-week gym data windows, and conduct the same type of analysis.

In preliminary work, it proved difficult to recover stable individual-level estimation because of constraints on how little data we have for each person. To improve estimation, we aggregate individuals into 10 deciles. The deciles are sorted by how the size of the sample of data we have for each person. People in Decile 1 have the least data and people in decile 10 have the most data. Note that because deciles were created to have approximately equal numbers of individuals, the higher-sampling rate deciles will have many more observations than the lower-sampling rate deciles. The decile method can also make evident whether there are any differences based on per-person sample sizes.

Denote $A_d(2t)$ to be the AUC corresponding to test period $\{2t-1, 2t\}$ for individuals in decile $d$, where $t$ is either week (gym data) or shift (hand-hygiene data). Note that the process described above only allows us to define $A_d(2t)$ for $t \geq 3$, given our timing notation. (The first observation is denoted $A_d(6)$ because it is testing AUC for weeks 5-6 based on estimation from weeks 1-4.) We assume that individuals have not developed any meaningful habit strength before they enter the sample, so that $A_d(2) = A_d(4) = 0.5$.

Following (2), we assume an asymptotic predictability curve of the form $A_d(t) = a_d - b_d e^{-c_d t}$. This shape of curve is highly plausible in this case because AUC naturally goes from .5 toward the maximum of 1 if predictability increases. The exponential structure implies that the rate of change is positive but asymptotes to zero. The parameters $a_d, a_d - b_d$ and $c_d$ represent the asymptotic level of AUC, the starting value $D_i(0)$, and the speed of adjustment. A higher value of $c_d$ represents faster adjustment.

We define the time to habituation $T_d^* = -\ln(a_d/20b_d)/c_d$ as the time it takes for $A_d(t)$ to reach 95% of its estimated asymptote $a_d$. We use nonlinear least squares to fit the empirical $A_d(2t)$ to each decile $d$'s AUC sequence and obtain the estimates $\hat{a}_d, \hat{b}_d, \hat{c}_d$. The results are reported in Table S8. All models have high quality of fit, as measured by $R^2$, indicating that individuals do become more contextually predictable over time in general. The estimated values of parameters $\hat{a}_d, \hat{b}_d, \hat{c}_d$, and time to habituation $T^*$ are rather consistent across all deciles. Specifically, it takes about 6 months for individuals to form a gym attendance habit (i.e., $T^* \approx 183$) .It only takes around 9 shifts for caregivers to learn to wash their hands habitually.

| | Gym data | | | | | Hand-hygiene data | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Decile | $a$ | $b$ | $c$ | $R^2$ | $T^*$ (days) | $a$ | $b$ | $c$ | $R^2$ | $T^*$ (shifts) |
| 1 | 0.77 | 0.35 | 0.018 | 0.94 | 126 | 0.86 | 0.75 | 0.32 | 0.85 | 9 |
| 2 | 0.78 | 0.37 | 0.019 | 0.92 | 121 | 0.85 | 0.68 | 0.27 | 0.89 | 10 |
| 3 | 0.78 | 0.37 | 0.017 | 0.93 | 131 | 0.84 | 0.72 | 0.32 | 0.85 | 9 |
| 4 | 0.79 | 0.32 | 0.012 | 0.90 | 180 | 0.83 | 0.66 | 0.30 | 0.87 | 9 |
| 5 | 0.81 | 0.35 | 0.011 | 0.91 | 187 | 0.85 | 0.67 | 0.27 | 0.88 | 10 |
| 6 | 0.81 | 0.33 | 0.010 | 0.89 | 209 | 0.83 | 0.68 | 0.31 | 0.86 | 9 |
| 7 | 0.82 | 0.32 | 0.009 | 0.84 | 226 | 0.83 | 0.69 | 0.31 | 0.86 | 9 |
| 8 | 0.84 | 0.34 | 0.010 | 0.83 | 220 | 0.83 | 0.70 | 0.32 | 0.85 | 9 |
| 9 | 0.83 | 0.36 | 0.012 | 0.80 | 187 | 0.82 | 0.67 | 0.31 | 0.83 | 9 |
| 10 | 0.80 | 0.34 | 0.012 | 0.77 | 172 | 0.85 | 0.70 | 0.29 | 0.82 | 10 |

**Table S8: Summary of Asymptotic Curve Fitting on AUC Sequences.**

*This table reports the estimated parameters, quality of fit ($R^2$), and estimated time to habituation for each decile in the two datasets. For each decile $d = 1, 2, \ldots, 10$, the time to habituation $T^*$ is defined as the time it takes for the asymptotic curve $A_d(t)$ to reach 95% of its asymptote.*

# 4 Field Tests of Insensitivity to Reward Devaluation

A hallmark of strong habits used in animal learning, psychology, and neuroscience is insensitivity to reward devaluation.[8] In animal studies, after a period of reward learning, food rewards are then devalued in various ways for a treatment group, and there is no devaluation change for an untreated control group.

After either devaluation or no-devaluation control, the animals enter a short "extinction period" experimental phase in which they can freely take action to gain rewards (e.g., by pressing a lever). If the animals are purely goal-directed, they will not work for further reward because the goal they pursue has been devalued– that is, if their actions are guided by the goal of reward-seeking, since actions no longer yield valuable reward (due to devaluation), they will not perform the actions. However, if animals have become completely habitual, they will work for reward at the same rate as the control group. These habitual animals are judged to have become insensitive to reward devaluation. The difference in the rate of responding to the devalued rewards for the control and treatment animals is an index of the degree of habit formation (typically from 0 to 1).

This section describes our novel approach for running a test of reward devaluation insensitivity in non-experimental field data. This is the first such test using field data. Having an exploratory test of how such insensitivity could be measured, and what a result parallel to the animal learning experiments would look like, is therefore useful (regardless

[8]There are also studies of insensitivity to reward contingency– e.g., the probability of reward contingent on a behavior such as a lever press is lowered, but the animal keeps responding at the baseline rate of presses per unit time.

of the results).

Unlike in animal experimental protocols, in the field data there is usually no explicit control over reward value. We therefore study two types of exogeneous changes that occur in the data and which are plausibly random (i.e., unrelated to a person's behavioral history), might change the value of the subjective reward of gym attendance and handwashing. The statistical tests are therefore tests of the joint hypothesis that reward is actually changed and that habitual people are less sensitive to the reward change. If the reward is not actually changing, then we will not find insensitivity in response to habit.

In two others cases we have access to data from controlled experiments specifically designed to increase subjective reward from an activity. Those interventions are various organizational interventions to promote handwashing at the hospitals, and a Stepup program to promote gym attendance . The crucial test is whether people who appear more habitual are less sensitive to these reward inducements.

Once reward-change variables are identified, testing for sensitivity to reward change uses two different approaches. The first approach is within-person: It uses our method for identifying the timing of when a person is in the different pre-habit and post-habit modes. As a possible source of reward change, we use unusual weather for gym attendance[9], and end-of-the-day room attendance for hand washing.

The second approach is between-person: It compares sensitivity to reward change for people who appear highly habitual, because they have a high AUC measure, to those who appear less habitual because they have a low AUC measure. The sources of reward change here do arise from explicit control– they are planned interventions used to encourage more gym attendance, and more hand washing.

## 4.1 Within-subject Field Tests of Insensitivity to Reward Devaluation Pre- and Post-habit

To test the impact of exogenous reward change on gym behavior in our truncated sample, unusual weather is used as an event which can change reward value of gym attendance. Weather is plausibly random (i.e., it does not depend on what gym goers did in the past) and may change the subjective reward value of going to the gym. To measure weather, we first mapped the ZIP codes of each gym to a latitude-longitude coordinate using data from 2013 collected by the US Census Bureau. As the average land area of a United States ZIP code is approximately 85 square miles, by modelling each ZIP code as a circle we find that the average radius of each ZIP code is approximately $r = 5$ miles (i.e., $\pi(r^2)$ is around 78.5 square miles. It is assumed that daily weather is similar within the artificial circle created by the radius.

The National Ocean and Atmospheric Administration (NOAA) of the Department of Commerce provides a detailed list of the weather stations across the country and their respective coordinates (101). It is known that one degree of latitude corresponds to a surface distance of approximately 69.1 miles, where surface distance is defined as the shortest path between two points at sea level. While the distance spanned by one degree of latitude remains

[9] A more ideal measure would be something like a large gym renovation (such that access to space and equipment was suddenly limited) or malfunctioning equipment, but neither of those are available in this dataset.

relatively constant regardless of location on the earth, the surface distance spanned by one degree of longitude varies considerably due to the inherent curvature of the earth. Since ZIP codes are small in relation to the size of the earth, we use the standard conversion formula $D = D_E cos(L)$, where $D$ is the surface distance spanned by one degree of longitude at a given location, $DE = 69.1$ miles is the distance spanned by one degree of longitude at the equator, and $L$ is the approximate latitude of the location in radians. Then, for each gym, we compiled a list of nearby weather stations falling within intervals of size 1/3-degree latitude and longitude centered at the gym.

For each date and gym combination for which weather data is made available, the highest temperature, lowest temperature, average temperature, precipitation, and snowfall were measured. Table S9 provides summary statistics on the main weather attributes and shows how these statistics differ on only those days when individuals attended the gym. We note that this restriction causes the means and standard deviations for both temperature attributes to increase slightly, and vice versa for precipitation and snowfall.

Unfortunately, not all weather stations provide measurements for all of the five aforementioned attributes. Hence, to obtain recordings for each date and gym and each weather attribute, we searched through the list of nearby weather stations once for each attribute, in order from closest to farthest, until a station with a measurement of that attribute was found.

The mean distances to each weather station remained relatively small (from a minimum of 0.01 miles to a maximum of 13.67 miles, means ranging between 3.05 for rain to 6.89 for average temperature). The slightly higher mean distance for average temperature measurements is due to the NOAA's classification of the other four as "core" elements of weather measurement, hence there are more stations that are equipped to regularly record them.

| Statistic | TMAX (F) | TMIN (F) | PRCP (mm) | SNOW (mm) |
|---|---|---|---|---|
| Mean | 72.11 | 51.50 | 18.19 | 0.64 |
| Median | 73 | 53 | 0 | 0 |
| St. Dev. | 15.24 | 14.05 | 81.13 | 9.85 |

**Table S9: Weather Summary Statistics.**

*This table reports summary statistics on the weather data used for the test of reward devaluation sensitivity in gym attendance. TMAX is maximum temperature (in Fahrenheit), TMIN is minimum temperature (in Fahrenheit), PRCP is percipitation (rain accumulation, in millimeters), and SNOW is snow (in millimeters).*

The next step is to use weather data to distinguish days with unusually good and bad weather. Obviously, the relationship between temperature and perception of weather varies greatly with geographic locations, seasons and individual tolerances. For instance, while a 60°F spring day in Illinois is felt as warm and pleasurable, a winter day with similar temperature in California would feel cold.

Therefore, to link temperature and weather quality, we focus on the change in temperature relative to an "expected" level of temperature instead of raw measurements. For a given day, we define average temperature as the average of TMAX and TMIN. For each individual, we look at the distribution of average temperature in the indi-

vidual's area over the observed period. Temperatures between the 25th to 75th quantiles are considered normal. For each day in the individual's time series, we say that it has an adverse temperature fluctuation if (i) its temperature falls outside the normal zone, (ii) the average temperature of the previous 3 days is normal, and (iii) the change in temperature compared to the previous 3 day average is at least half the standard deviation of the average temperature distribution. This means that a day in states with stable weather like California (average temperature SD = $9.1^o$F) only needs a small temperature change to be considered to have adverse temperature fluctuation, as compared to states with larger weather variations like Colorado (average temperature SD = $15.8^o$F) or Utah (average temperature SD = $15.3^o$F). Formally, for any individual $i$, denote $q^w_{1i}, q^w_{3i}, \sigma^w_i$ as the 25th quantiles, 75th quantiles, and standard deviation of the average temperature distribution. Let $d_{ti}$ be the average temperature of calendar day $t$ in individual $i$'s area and $\bar{D}_{ti} = (d_{t-1,i} + d_{t-2,i} + d_{t-3,i})/3$ be the average temperature of the previous 3 days. An adverse temmperature fluctuation is observed on day $t$ if (i) $d_{ti} < q^w_{1i}$ or $d_{ti} > q^w_{3i}$, (ii) $q^w_{1i} \leq \bar{D}_{ti} \leq q^w_{3i}$, and (iii) $|d_{ti} - \bar{D}_{ti}| > \sigma^w_i/2$.

Similarly, a positive temperature fluctuation is observed when there is a change in temperature of at least half standard deviation in the opposite directions. In other words, we must have (i) $q^w_{1i} \leq d_{ti} \leq q^w_{3i}$, (ii) $\bar{D}_{ti} < q^w_{1i}$ or $\bar{D}_{ti} > q^w_{3i}$, and (iii) $|d_{ti} - \bar{D}_{ti}| > \sigma^w_i/2$. A day is considered to have "unexpectedly bad" weather if it has either snowfall, persistent moderate rain to shower (defined as at least 5mm of rain per hour), or an adverse temperature fluctuation. Conversely, a day with "unexpectedly good" weather has a positive temperature fluctuation and no precipitation or snow. Note that the relation between unexpected weather could have either positive or negative effects on gym attendance. Bad weather can raise the cost of getting to the gym (reducing attendance) or lower the opportunity cost of being inside rather than exercising outside (increasing attendance). Therefore, the analysis examines the absolute value of the coefficients associated with both types of unexpected weather shocks.

We examine the relationship between habit formation and sensitivity to weather shocks as follows. We again partition the individuals into 10 deciles based on a their sample sizes. For each decile, we use the $T^*$ estimated from Section 3.3 as a cutoff for pre- and post-habituation periods. (Recall that using our method, this will generate a single $T^*_d$ for everyone in decile $d$. We then run 10 linear regressions, one for each decile, with the indicator for whether the individual goes to the gym on a given day as the dependent variable. In addition to all the context variables described in Section S2.1, the covariates also include individual fixed effects, a pair of weather dummy variables (unexpectedly good or unexpectedly bad) and their interactions with the indicator for post-habituation period, which is defined as the period after $T^*$. The key prediction is that the interactions with post-habituation should move in the oppposite direction of the weather dummy variables. For example, as we will see both unusually good and bad weather are associated with more gym attendance (regression coefficients are positive). These effects should be weaker in the post-habit period, so the habit x weather interaction terms should be negative if habit is reducing reward sensitivity.

Table S10 shows the estimated coefficients of weather shocks and their interactions with post-habit indicator. Weather variables are only significant at conventional levels (despite the large sample size) and the weather x post-habit interaction terms go in the opposite direction for bad weather but not for good weather. The expected result

from devaluation insensitivity is an interaction term that has the opposite sign and hence offsets the simple weather variable. This pattern is just not evident in the data.

In the case of hand washing, it is harder to find a plausibly exogeneous shock which would reliably be predicted to change the value of washing from one episode to the next. The ideal candidate would be rapid response events which occur at various times throughout the day. Any member of the hospital who is on the rapid response team would receive a notification about an urgent case requiring them to rush to the patient(s). Such situations are often a matter of life or death and require immediate attention, therefore potentially affecting the comparative value of hand washing behavior in the moment. Unfortunately, these events are only reported by individual hospitals and not available in the dataset we have access to.

Instead, we use an indicator for whether a worker is exiting their last room episode of the day as a proxy for reward devaluation. As with unexpected weather on gym behavior, the relation between exiting last episode and hand-washing can have either positive or negative effects. If the key driver for hand-washing is to not spread infection from one patient/episode to the next, then the last episode decreases the value of hand-washing. However, if the key driver for hand-washing is to keep oneself clean and free of infection after work at an infectious hospital, then the last room episode increases the value of hand-washing. We include this analysis mostly to have a parallel to analysis of gym attendance based on weather, but its effect is more ambiguous and could well be statistically weaker (especially since there is only one last episode of each day).

Table S11 reports the estimated coefficient of last episode indicator and its interaction with post-habit indicator (and the post-habit indicator as a control). The interaction term has the same sign as the last episode indicator, which is negative, so it does not show the offsetting effect of revaluation insensitivity seen in animal experiments. However, this result should be interpreted with caution, as unlike the gym data, we do not really observe the caregivers from the first dates of work. It could be that many of the caregivers had already developed a habit of hand hygiene to some extent prior to the start of the data (even though, as noted elsewhere (102), there was a sharp jump in handwashing compliance from before the machines were introduced to afterward, which could suggest starting a new type of habit a la the "habit discontinuity hypothesis").

| | *Dependent variable: Indicator for gym attendance* |
|---|---|
| Post-habit | $-0.014^{***}$ |
| | (0.001) |
| Good weather | $0.004^{**}$ |
| | (0.002) |
| Bad weather | $-0.002^{**}$ |
| | (0.001) |
| Good weather × Post-habit | $0.004^{***}$ |
| | (0.002) |
| Bad weather × Post-habit | $0.005^{***}$ |
| | (0.001) |
| Observations | 8,750,054 |
| $R^2$ | 0.237 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Table S10: Reward Devaluation Regression Analyses-Gym Data**

*This table reports the regression results with the indicator for whether the individuals go to the gym on a given day as the dependent variable. All regressions include the usual context variables described in Section S2.3, individual fixed effects, a pair of dummy variables for unusually good and unusually bad weather, and their interactions with post-habit indicator. Standard errors are clustered at the individual level and reported in parentheses.*

| | *Dependent variable: Indicator for hand sanitation* |
|---|---|
| Post-habit | $0.001^{**}$ |
| | (0.0003) |
| Last episode | $-0.010^{***}$ |
| | (0.003) |
| Last episode × Post-habit | $-0.013^{***}$ |
| | (0.002) |
| Observations | 18,514,432 |
| $R^2$ | 0.363 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Table S11: Reward Devaluation Regression Analyses-Hand-hygiene Data**

*This table reports the regression results with the indicator for whether the caregivers sanitize their hands at a given opportunity as the dependent variable. All regressions include the usual context variables described in Section S2.3, individual fixed effects, a dummy variable for last episode of the shift, and its interaction with post-habit indicator. Standard errors are clustered at the individual level and reported in parentheses.*

A challenge for the types of analyses we have just shown is that in animal and human learning experiments reward devaluation is carefully controlled– for example, by feeding animals food rewards until they are satiated. But here the reward revaluation is only crudely hypothesized based on unusual weather and the reduced value of having clean hands when leaving the hospital. Better reward change measures might detect stronger evidence of insensitivity, after habit formation, similar to the effects seen in experiments with animals (and some humans, (15)). This is an important challenge going forward for trying to find cognitive, neural, and behavioral hallmarks of habit that are reasonably well-established in tightly controlled laboratory experiments, in less well controlled (and measured) field data.

## 4.2 Between-person Predictability Reactions Toward Incentivised Intervention

A different approach to testing for a habitual reduction in reward sensitivity is to use each person's estimated value of AUC– a measure of statistical predictability– as a measure of habit formation. It is possible that different people with higher and lower AUC's have different responses to subjective reward changes. This "between-person" approach is novel to the empirical habit literature because AUC measures have never been used before as a correlate of habit. As a result, we also not expect that these results will necessarily be strong; they can be viewed as an illustration of a methodology that could be more useful in other settings and with other kinds of data availability.

In this section we will test for between-person effects of AUC using explicit interventions in both gyms and hospitals, designed to encourage more attendance and more hand washing. This question is interesting for the study of habits but it is also of large potential practical impact. A frontier in behavior change research is trying to figure out

not just which interventions work, but which interventions work for which kinds of people (or in which situations). The hypothesis here is akin to "personalized medicine"– i.e., there are predictable differences in which patients respond to which treatments, and a key challenge is to identify those differences and match them with treatment. One likely outcome is that a particular intervention can be predicted to work only on a subset of people. From a cost-benefit point of view, this outcome can make the difference between a worthwhile treatment when personalized, and a treatment which is not worthwhile because it wastes too much money and resources on people who are likely to be unresponsive.

To answer this question, we will be using an analytic sample of hand washing behaviors from only those individuals who experienced different interventions. Because attention is no longer restricted to people for whom pre- and post-habit timing could be estimated, the samples of individuals is larger.

**Interventions in Hand Washing**

The hospital applied multiple interventions at different time points to motivate more hand washing. (102) used these field manipulations to test whether the intervention had an effect. They confirmed positive intervention effects, though the effect sizes differed depending on which intervention was used. There were five kinds of interventions in total: Performance feedback, goal setting, leadership, competition, and incentive. We will be focusing on the two most common and most effective ones reported in (102): leadership and incentive. The interventions were introduced at the unit level.[10] A total of 27 units (leadership = 19, incentive = 25) applied at least one type of intervention [11].

We estimate the effects of interventions at the individual level using the LASSO logistic regression specified in section 3.1 with one additional variable: a post-intervention indicator. The hypothesis being tested is that caregiver who are more predictable (higher AUC) will be less sensitive to interventions. Following (102), the intervention indicator equals one on all days after the intervention was introduced, and equals zero on all days before the intervention.[12] The estimated coefficients of the intervention indicator from the LASSO model will be the first measurement of whether there is an intervention effect. It is called the Intervention Coefficient (IC) (see Table S12). We only include individuals who register before the intervention day and stay after it. In addition, we exclude outlying individuals whose IC is below or above the mean level minus or plus three times the sample standard deviations.[13] The remaining sample size for this analysis is reported in Table S12.

The AUC value of the LASSO model is used as an indicator of how habitual the person is. This LASSO model has the same specification as the one reported in section 3.1, but here is only based on periods from the time caregivers start to work until the day before intervention. In other words, we measured habit formation levels (AUC) from pre-intervention data only. Then, we regress these AUC values on the intervention effects measured by the

[10] A unit is a unique identifier for one department at one hospital site.

[11] One unit could have multiple interventions, but we ignore the interaction effect, as in the original paper. This means that we are presuming if there two interventions ongoing at the same time, their effects are additive.

[12] It is plausible that sensitivity to the interventions does not have a fixed value in the LASSO regression, as is specified by the 0-1 dummy variable we use. Further research could explore other nonlinear specifications reflecting either a slow response to intervention, or a reduction in response over time.

[13] This led to exclusion of 7 individuals in the leadership dataset and 28 individuals in the incentive dataset.

IC variable introduced above. If habituation, as measured by higher AUC, decreases intervention effects then these correlations should be negative. The results of the analysis are reported in Table S12 (the value of the coefficients and the corresponding p-values). We did not find a significant effect in either leadership or incentive intervention when the dependent variable is the pure intervention coefficient.

We also conducted another analysis to see whether the level of habit formation will affect whether or not the person has a positive effect on the intervention. A dummy variable was created which equals 1 if IC>0 and 0 otherwise. Correlations between this positive-only dummy and AUC are reported in Table S12). The effects for leadership and incentives interventions are both positive, but not strongly significant. A positive correlation is the opposite of what is expected if both higher AUC is associated with stronger habit formation, and stronger habit formation is associated with a lower response to reward change (from intervention).

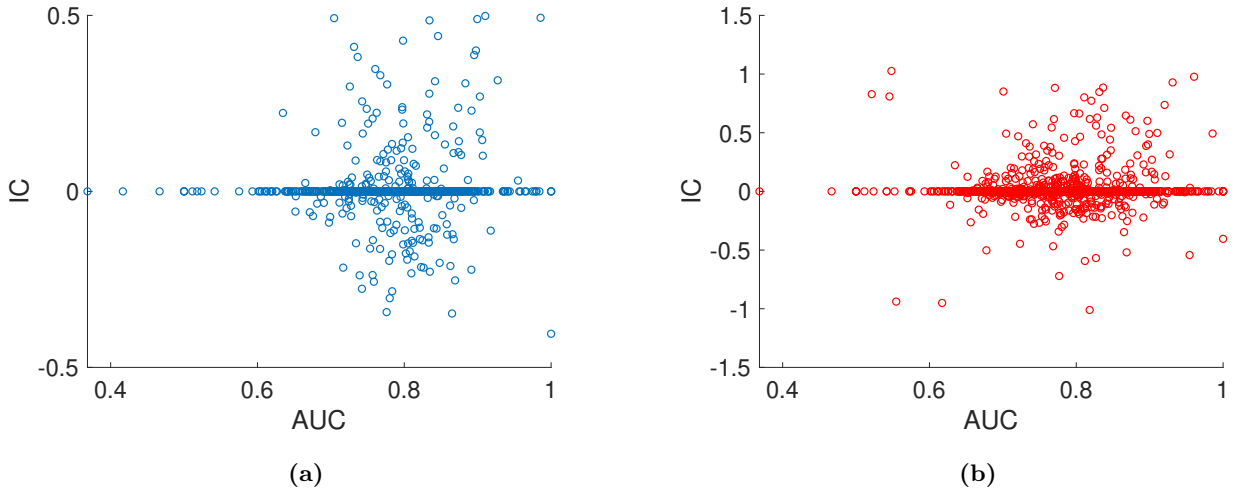



**Figure S3: Relationship Between AUC and IC for Hand Washing.**

*This figure shows the scatterplots of AUC and IC for (a) (left side) the hand-washing subsample who received incentive interventions and (b) (right side) the hand-washing subsample who received leadership interventions.*

| Intervention effect measurement | Coefficient | 2-sided p-value |
|---|---|---|
| Leadership ($N = 1,396$) | | |
| Intervention Coefficient (IC) | 0.08 | 0.12 |
| Indicator for positive IC | 0.32 | 0.74 |
| Incentive ($N = 878$) | | |
| Intervention Coefficient (IC) | -0.005 | 0.88 |
| Indicator for positive IC | 0.20 | 0.13 |

**Table S12: The Effect of Habit Level (AUC) on Individual-level Intervention Changes (IC) for Hand Washing**

*We regressed the intervention coefficient (IC) on the pre-intervention AUC value for two intervention datasets, leadership and incentive. Each observation represented one person. The coefficients of the transformations of variable IC for each dataset are reported in this table along with its corresponding 2-sided p-value.*

**Interventions in Gym Attendance Data**

(4) conducted a large field experiment "megastudy" (N = 61,293) with the 24 Hour Fitness gym chain to test the effectiveness of 54 different interventions aimed at promoting lasting exercise behavior. N=29,264 individuals in our sample happened to be among the participants of this megastudy. Again, our goal was to examine whether the interventions have different effects for individuals with different habit levels, as measured by AUC. However, for many individuals in our sample, we were unable to obtain full data between their first dates and the intervention period. In other words, there is a gap in data between the end of our sample and the start of the intervention. Therefore, we cannot conduct the same analysis we did for the hand washing data. Instead, we simply calculated the difference between weekly gym visits during the 4-week intervention period and weekly gym visits in the 4 weeks before the intervention period (mean = 0.20, SD = 1.08) as a measure for the intervention coefficient (IC).[14] (This is a cruder measure and is expected to generate weaker measures of true effects.) We then determine the relationships between individual-level AUC and IC measures much as was done above for hand-washing interventions.

The results are reported in Table S13. There is no significant effect of AUC on IC, but logistic regression using a 0-1 indicator for positive IC as the dependent variable shows a large and negative effect of AUC ($\beta$=-.48, $p = .91 * 10^{-5}$, summarized as $p < .001$ in the table). This result suggests that for more habitual higher-AUC individuals, the interventions are *less* likely to increase exercise behavior, consistent with the association of stronger habit formation with insensitivity to reward change.

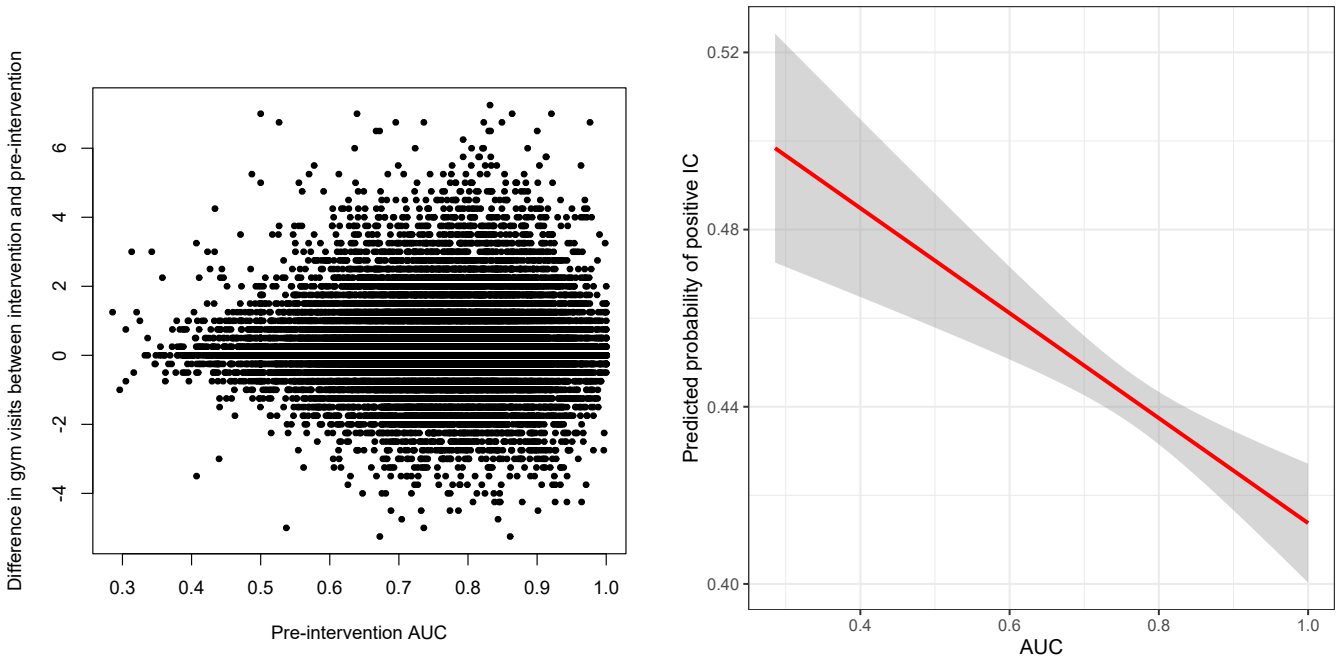
Keep in mind that the 24HF participants in the StepUp sample were self-selected volunteers. In general one must therefore consider how self-selection could create spurious incentive effects that would not apply to the full sample of participants. In this case, however, we are comparing lower and higher-AUC participants who all volunteered. The

[14]This way of measuring IC is different than the extended-LASSO method that was used for analyzing handwashing interventions. The reason is that the daily data are not available so the day-by-day LASSO estimation could not be conducted.

negative correlation could only be explained if the higher-AUC participants who volunteered had weaker self-selection motives than the lower-AUC participants. There is no reason to think this is the case, and no other data that can be brought to bear to evaluate this alternative explanation.

Because the negative IC-AUC correlation from StepUp is one of multiple similar comparison results, let's explore how robust it is likely to be. In the pre- vs. post-habit analyses there were three comparisons (bad and good weather gym attendance, and last-room handwashing ) to judge reward insensitivity. In the within-participant high-low AUC comparisons three interventions were compared (StepUp gym attendance, and Leadership and Incentives handwashing interventions), on two transforms of the IC variable (IC, I(IC>0) dummy variable). That creates a total of $3 + 3 * 2 = 9$ comparisons. While the post-habit gym weather variables are strong, suppose we take the large effect of positive IC on between-person gym AUC as the one of which we regard as interestingly significant and plausibly reproducible. (This is a conservative comparison, obviously, just to guide intuition about potential false positives). A Bonferonni correction for multiple comparisons would increase the p-value from the AUC x (IC>0) gym coefficient to $9 * (.91 * 10^{-5}) \approx 10^{-4}$, which is still extremely unlikely to be due to chance if there is no true effect.

Furthermore, the quality of interval validity of the intervention treatment is strongest for StepUp, since all participants received multiple correspondence and we have strong confidence they were aware of the treatment. We do not know (and lack detail about) whether all caregivers were equally aware of the Proventix interventions. Thus, in the main paper we describe the negative IC-AUC StepUp correlation as an example of what field evidence of reward insensitivity due to habit might look like.




**(a)** **(b)**

**Figure S4: Relationships Between AUC and IC for Gym Attendance**

*This figure shows the relationship between pre-intervention AUC (x-axis) and two measures of the IC Intervention Coefficient effect. (a) (left) The y-axis plots the IC as measured by the difference in average weekly gym visits between the intervention (StepUp) and pre-intervention periods. There is no association ($r = 1.3 \times 10^{-5}$, two-sided $p = 1.00$). (b) (right) The y-axis plots the predicted probability of the IC measure being positive (from a logit regression of 0-1 IC positivity against AUC). There is a small but highly significantly nonzero negative correlation between AUC and the probability of positive IC ($r = -0.026, p = 0.9 \times 10^{-5}$). Note that this $r = 0.026$ is a correlation coefficient and the value of $\hat{\beta} = -.48$ in is the logit regression coefficent.*

| Intervention effect measurement | Coefficient | 2-sided p-value |
|---|---|---|
| Megastudy ($N = 29,264$) | | |
| Intervention Coefficient (IC) | 0.0001 | 1.00 |
| Indicator for positive IC | -0.48 | <0.001 |

**Table S13: The Effect of Habit Level (AUC) on Individual-level Intervention Changes (IC) for Gym Attendance**

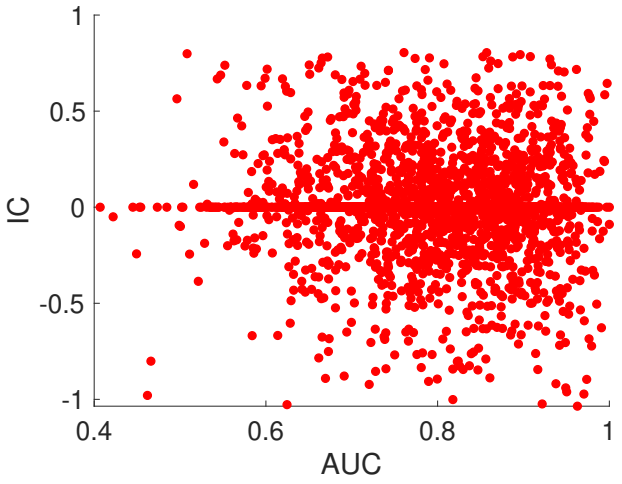
*We regressed the intervention coefficient (IC) on the pre-intervention AUC value for individuals in our gym data who volunteered for and participated in the megastudy. Each observation represented one person. The coefficients of the transformations of variable IC for each dataset are reported in this table along with their corresponding 2-sided p-values.*

### 4.3 Sensitivity Analyses

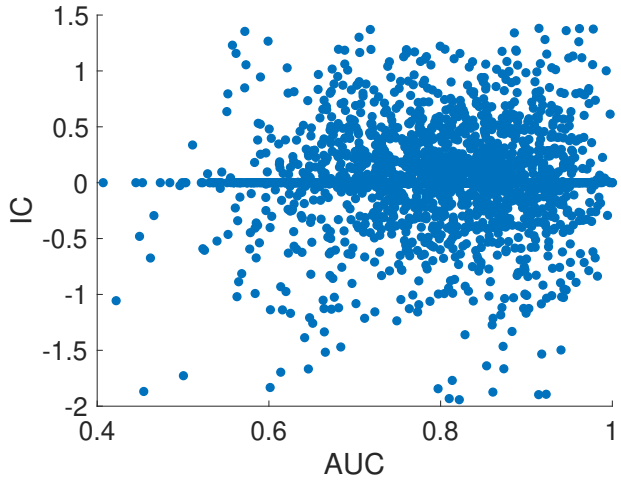
We conduct three additional sensitivity analyses to examine between-subject sensitivity towards exogenous reward revaluation variables. We repeat the analysis of the hand-washing data in section 4.2 for both datasets, but replacing the intervention indicator variable with the corresponding reward revaluation variables mentioned in section 4.1: indicator for good weather, indicator for bad weather in the gym data, and indicator for exiting the last room of the shift in the hand-washing data.

Then, we regress the AUC values on these additional sensitivity coefficients (to make notations consistent, we will still call them IC) estimated from the LASSO model (see scatter plots in Figure S5). We found a slight negative correlation between the bad weather impact and AUC level (IC = -.08, p-value = .07), which is consistent with the previous finding as the bad weather should hurt gym attendances so that more habitual individuals are more affected. We found a significant positive correlation on good weather impact (IC = 0.13, p-value = 0.02). Both findings point out that people with higher AUC levels are slightly more likely to be affected by weather changes, but in opposite directions.

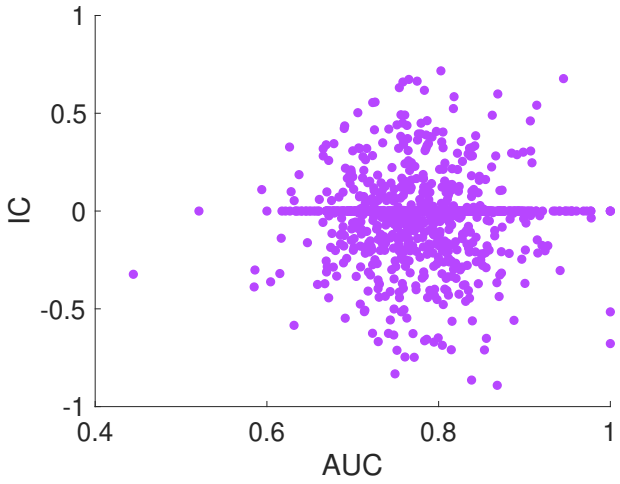
For the last shift impact on handwashing, there is no correlation of that IC with the level of AUC (r= .01, p-value = .91).




**(a)** Bad Weather Sensitivity vs AUC on Gym Data




**(b)** Good Weather Sensitivity vs AUC on Gym Data




**(c)** Whether Last Shift Sensitivity vs AUC on Hand Wash Data

**Figure S5: Relationship Between AUC and IC for Other Three Factors.**

*Similar to Figure S4, this panel of plots also shows the scatter plots of AUC and IC for (a) the gym attendance data where IC is the sensitivity to whether it is bad weather or not on the day (b) the gym attendance data where IC is the sensitivity to whether it is good weather or not. (c) the hand washing data where IC is the sensitivity to whether the current shift is the last shift or not.*

# 5 Additional Analyses: Demographic Predictors of AUC

## 5.1 Motivation

Given the rich individual-level data we work with, which includes a home zip code associated with each gym goer, it is possible to look for systematic categorical differences in the degree of predictability across sub-groups of gym

goers. In order to run this analysis, we link the individual-level AUC data from gym goers with Census information from the year 2019. The Census data was purchased online from Income by Zip Code.[15] Unfortunately, the data on hospital workers does not come with zip code information, so we are unable to use this technique to analyze demographic differences with respect to the predictability of handwashing behavior.

The census variables discussed below, along with demographic data captured by the gym chain at time of registration (gender and age), allow us to estimate the demographic and SES profile of each individual gym goer and investigate demographic differences in gym attendance predictability.

## 5.2 Variable List

1. **Income:** As a proxy for individual income, we use the average household income of the individual's ZCTA.[16]
2. **Rural/Urban:** As a proxy for how rural or urban an individual's environment, we use a continuous measure of population density for the individual's ZCTA.
3. **Children:** As a proxy for an individual's likelihood of having children, we compute the fraction of married and single households in the gym goer's ZCTA who have children (under the age of 18).
4. **Age:** We have age data on the gym goers in our sample because they were required to report this at the time of gym membership registration. In addition, we calculate relative age by taking the difference between the median age in an individual's ZCTA (median age comes directly from the Census dataset) and their self-reported age.
5. **Gender:** We have gender data on the gym goers in our sample because they were required to report this at the time of gym membership registration.

## 5.3 Correlation Matrix

We analyze the correlation matrix of continuous variables in the gym and Census ZCTA data to see if there are significant and/or surprising correlations between demographic variables, as well as with individual-level AUC and base rates of attendance.

Significant correlations which are worth noting include the expected positive correlation between auc.test and auc.train ($\rho$=0.661) and the negative correlation between auc.train and base rate of attendance ($\rho$=-0.237). The latter correlation underscores the fact that in our analyses, frequency and predictability are not positively correlated (and are slightly negatively correlated). Also notable are the positive correlations between income and median age of neighborhood ($\rho$=0.593) as well as propensity to be married with children ($\rho$=0.373) - which intuitively make sense given individuals tend to accumulate more income as they get older, and financial security gives people the

[15] The same data can be purchased at the following link under 'Income by Zip Code List + Demographics (All US)." https://www.incomebyzipcode.com/median-income-by-zip-code-list#pricing

[16] ZCTAs, or ZIP Code Tabulation Areas, are generalized areal representations of United States Postal Service (USPS) ZIP Code service areas. While the latter is a trademark of the U.S. Postal Service, the former is a trademark of the U.S. Census Bureau. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html

ability to financially support children. Population density is negatively correlated with the median age ($\rho$=-0.204) (younger people are more likely to live in urban areas) and with having children if one is married or single ($\rho$=-0.138, $\rho$=-0.229, respectively).

| | Attendance rate | Age | Holdout AUC | Training AUC | Income | Pop. density | Median age | Married children | Single children |
|---|---|---|---|---|---|---|---|---|---|
| Base rate | 1 | 0.002 | 0.073 | -0.237 | 0.015 | -0.010 | -0.007 | 0.006 | -0.004 |
| Age | 0.002 | 1 | -0.007 | 0.073 | 0.062 | -0.032 | 0.109 | 0.0005 | 0.076 |
| Holdout AUC | 0.073 | -0.007 | 1 | 0.661 | -0.002 | -0.011 | 0.018 | 0.007 | -0.004 |
| Training AUC | -0.237 | 0.073 | 0.661 | 1 | -0.020 | -0.020 | -0.010 | 0.011 | -0.008 |
| Income | 0.015 | 0.062 | -0.002 | -0.020 | 1 | 0.027 | 0.593 | 0.373 | 0.063 |
| Pop. density | -0.010 | -0.032 | -0.011 | -0.020 | 0.027 | 1 | -0.204 | -0.138 | -0.229 |
| Median age | -0.007 | 0.109 | 0.018 | -0.010 | 0.593 | -0.204 | 1 | -0.505 | -0.258 |
| Married w/ kids | 0.006 | 0.0005 | 0.007 | 0.011 | 0.373 | -0.138 | -0.505 | 1 | 0.151 |
| Single w/ kids | -0.004 | 0.076 | -0.004 | -0.008 | 0.063 | -0.229 | -0.258 | 0.151 | 1 |

**Table S14: Correlation Matrix of Continuous Variables.**

*This table shows a correlation matrix of variables in our data which take on continuous values.*

## 5.4 Regression Results

We further explore whether demographic and SES characteristics effect predictability using a regression framework. We truncated our sample to include only individuals for whom we have full information on gender, age, and zip code. We then ran a regression with AUC as the dependent variable and demographics characteristics described in Section 5.2 as explanatory variables. The results are reported in Table S15 below.

| | *Dependent variable: AUC* | | |
|---|---|---|---|
| | Coefficient | $t$-statistic | Effect size $d$ |
| log(avg. household income) | $-0.004^{**}$ | $t$ = -2.263 | $d$ = -0.027 |
| | (0.002) | | |
| log(population density) | $-0.005^{***}$ | $t$ = -6.786 | $d$ = -0.082 |
| | (0.001) | | |
| Fraction married with kids | $0.019^{**}$ | $t$ = 2.328 | $d$ = 0.028 |
| | (0.008) | | |
| Fraction single with kids | $-0.022^{***}$ | $t$ = -3.265 | $d$ = -0.039 |
| | (0.007) | | |
| Age | $0.001^{***}$ | $t$ = 9.409 | $d$ = 0.113 |
| | (0.0001) | | |
| Male | $-0.008^{***}$ | $t$ = -6.15 | $d$ = -0.074 |
| | (0.001) | | |
| Constant | $0.844^{***}$ | | |
| | (0.025) | | |
| Observations | 27,659 | | |
| $R^2$ | 0.007 | | |
| *Note: Standard errors in parentheses.* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

**Table S15: Demographic Predictors of AUC**

*Multiple regression results summarizing the predictive power that various demographic variables have on individual-level AUC. Average household income, population density (sq. mi.), and the fraction of married and single households with children under 18, were calculated using ZCTA Census data and the gym goer's zip code. Gender and age information came from the gym chain, based on gym goer self-report. The rightmost column reports the t-statistic and effect size (Cohen's d) for each variable.*

# 6 Human Subjects Protections

## 6.1 Gym Attendance

Before initiating this project, the California Institute of Technology Institutional Review Board and the University of Pennsylvania Institutional Review Board reviewed and approved this study. Because this study involved an analysis of de-identified, archival data, a waiver of informed consent was approved by the Institutional Review Boards per Federal Regulation HHS CFR 45.46.117(c) (2).

### 6.2 Hand Washing

Prior to initiating this project, the California Institute of Technology Institutional Review Board and the University of Pennsylvania Institutional Review Board reviewed and approved this study. Because this study analyzed de-identified archival data, a waiver of informed consent was approved by the Institutional Review Boards per Federal Regulation HHS CFR 45.46.117(c) (2).

## 7 Review of Habit Formation Studies

### 7.1 Summary of Previous Habit Formation Studies

| | Min | Q1 | Median | Q3 | Max | This Paper (Study 1) | This Paper (Study2) |
|---|---|---|---|---|---|---|---|
| Study sample size | 11 | 66 | 166 | 262 | 14,000 | 30,110 | 3,124 |
| Predictor variables | 1 | 1 | 1 | 2 | 3 | 22 | 34 |
| Study duration (days) | 1 | 30 | 84 | 240 | 1095 | 1,525 (median) | 98 (median) |

**Table S16:** Summary statistics for habit formation papers published following, and citing, (2). There were 43 papers reviewed (16, 26, 42, 43, 103–141), with a total of 47 studies between them. The number of predictor variables are only relevant for 5 studies. All the other studies looked at measures of self-report to judge habit formation, rather than the strength of specific predictor variables.

# References


1 Amanda Rebar, Benjamin Gardner, Ryan E. Rhodes, and B. Verplanken. The measurement of habit. In B. Verplanken, editor, *The Psychology of Habit*, pages 31–49. Springer, 2018.

2 Phillippa Lally, Cornelia H.M. van Jaarsveld, Henry W.W. Potts, and Jane Wardle. How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology*, 40(6):998–1009, 2010.

3 Gary Charness and Uri Gneezy. Incentives to exercise. *Econometrica*, 77(3):909–931, 2009.

4 Katherine L. Milkman, Dena Gromet, Hung Ho, Joseph S. Kay, Timothy W. Lee, Pepi Pandiloski, Yeji Park, Aneesh Rai, Max Bazerman, John Beshears, Lauri Bonacorsi, Colin Camerer, Edward Chang, Gretchen Chapman, Robert Cialdini, Hengchen Dai, Lauren Eskreis-Winkler, Ayelet Fishbach, James J. Gross, Samantha Horn, Alexa Hubbard, Steven J. Jones, Dean Karlan, Tim Kautz, Erika Kirgios, Joowon Klusowski, Ariella Kristal, Rahul Ladhania, George Loewenstein, Jens Ludwig, Barbara Mellers, Sendhil Mullainathan, Silvia Saccardo, Jann Spiess, Gaurav Suri, Joachim H. Talloen, Jamie Taxer, Yaacov Trope, Lyle Ungar, Kevin G. Volpp, Ashley Whillans, Jonathan Zinman, and Angela L. Duckworth. Megastudies improve the impact of applied behavioural science. *Nature*, 600:478–483, Dec 2021.

5 D. Acland and M.R. Levy. Naiveté, projection bias, and habit formation in gym attendance. *Management Science*, 61(1):146–160, 2015.

6 J. A. Ouellette and W. Wood. Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin*, 124(1):54–74, 1998.

7 Ann M. Graybiel. The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory*, 70(1-2):119–136, 1998.

8 Asaf Mazar and Wendy Wood. Illusory feelings, elusive habits: People overlook habits in explanations of behavior. *Psychological Science*, 33(4):563–578, Apr 2022. ISSN 0956-7976. doi: 10.1177/09567976211045345.

9 Christopher D. Adams and Anthony Dickinson. Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 33B:109–121, 1981.

10 David T. Neal, Wendy Wood, Mengju Wu, and David Kurlander. The pull of the past: When do habits persist despite conflict with motives? *Personality and Social Psychology Bulletin*, 37(11):1428–1437, 2011.

11 Kyle S. Smith, Arti Virkud, Karl Deisseroth, and Ann M. Graybiel. Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 109(46):18932–18937, 2012.

12 B. Balleine and J. O'Doherty. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 35(1):48–69, 2010.

13 E. Tricomi, B. W. Balleine, and J. P. O'Doherty. A specific role for posterior dorsolateral striatum in human habit learning. *The European Journal of Neuroscience*, 29(11):2225–2232, 2009.

14 S. de Wit, M. Kindt, S. L. Knot, A. Verhoeven, T. W. Robbins, J. Gasull-Camos, M. Evans, H. Mirza, and C. M. Gillan. Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of experimental psychology*, 147(7): 1043–1065, 2018.

15 E. Pool, Rani Gera, Aniek Fransen, Omar Perez, Anna Cremer, Mladena Aleksic, Sandy Tanwisuth, Stephanie Quail, Ahmet Ceceli, Dylan Manfredi, Gideon Nave, Elizabeth Tricomi, Bernard Balleine, Tom Schonberg, Lars Schwabe, and John O'Doherty. Determining the effects of training duration on the behavioral expression of habitual control in humans: a multi-laboratory investigation. *PsyArXiv: https://psyarxiv.com/z756h*, 2021.

16 I. Walker, G. Thomas, and B. Verplanken. Old habits die hard: Travel habit formation and decay during an office relocation. *Environment and Behavior*, 47(10):1089–1106, 2015.

17 Benjamin Gardner. A review and analysis of the use of 'habit' in understanding, predicting and influencing health-related behavior. *Healthy Psychology Review*, 9(3):277–295, 2015.

18 Wendy Wood and David Neal. The habitual consumer. *Journal of Consumer Psychology*, 19:579–592, 2009.

19 Mindy F. Ji and Wendy Wood. Purchase and consumption habits: Not necessarily what you intend. *Journal of Consumer Psychology*, 17(4):261–276, 2007.

20 Unna Danner, Nanne de Vries, and Henk Aarts. Habit vs. intention in the prediction of future behaviour: the role of frequency, context stability and mental accessibility of past behaviour. *British Journal of Social Psychology*, 47(2):245–265, 2008.

21 A. Mazar and Wendy Wood. Defining habit in psychology. In B. Verplanken, editor, *The Psychology of Habit*, pages 13–29. Springer, 2018.

22 Helen C. Fox, Kwang-Ik A. Hong, Kristen Siedlarz, and Rajita Sinha. Enhanced sensitivity to stress and drug/alcohol craving in abstinent cocaine-dependent individuals compared to social drinkers. *Neuropsychopharmacology*, 33(4):796–805, 2007.

23 Rajita Sinha. Modeling stress and drug craving in the laboratory: implications for addiction treatment development. *Addiction Biology*, 14(1):84–98, 2009.

24 Stuart G. Ferguson and Saul Shiffman. The relevance and treatment of cue-induced cravings in tobacco dependence. *Journal of Substance Abuse Treatment*, 36(3):235–243, 2009.

25 Benjamin Gardner, Charles Abraham, Phillippa Lally, and Gert-Jan de Bruijn. Towards parsimony in habit measurement: Testing the convergent and predictive validity of an automaticity subscale of the self-report habit index. *International Journal of Behavioral Nutrition and Physical Activity*, 9(102), 2012.

26 Sheina Orbell and Bas Verplanken. The automatic component of habit in health behavior: Habit as cue-contingent automaticity. *Health Psychology*, 29(4):374–383, 2010.

27 John A. Bargh. The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer and T.K. Srull, editors, *Handbook of social cognition: Basic processes; applications*, pages 1–40. Lawrence Erlbaum Associates, Inc., 1994.

28 Agnes Moors and Jan De Houwer. Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, 132(2):297–326, 2006.

29 Eric Garr and Andrew Delamater. Exploring the relationship between actions, habits, and automaticity in an action sequence task. *Learning and Memory*, 26:128–132, 2019.

30 B. Verplanken and S. Orbell. Reflections on past behavior: A self-report index of habit strength. *Journal of Applied Social Psychology*, 33(6):1313–1330, 2003.

31 Martin S. Hagger, Amanda L. Rebar, Barbara Mullan, Ottmar V. Lipp, and Nikos L. D. Chatzisarantis. The subjective experience of habit captured by self-report indexes may lead to inaccuracies in the measurement of habitual action. *Health Psychology Review*, 9(3):296–302, 2015. ISSN 1743-7202. doi: 10.1080/17437199.2014.959728.

32 Marieke A. Adriaanse, Floor M. Kroese, Marleen Gillebaart, and Denise T. D. De Ridder. Effortless inhibition: habit mediates the relation between self-control and unhealthy snack consumption. *Frontiers in Psychology*, 5:444, 2014.

33 Marleen Gillebaart and Marieke A. Adriaanse. Self-control predicts exercise behavior by force of habit, a conceptual replication of adriaanse et al. (2014). *Frontiers in Psychology*, 8:190, 2017.

34 Wendy Wood and Dennis Rünger. Psychology of habit. *Annual review of psychology*, 67:289–314, 2016.

35 Sheina Orbell and Bas Verplanken. The strength of habit. *Health Psychology Review*, 9(3):311–317, 2015. doi: 10.1080/17437199.2014.992031. URL `https://doi.org/10.1080/17437199.2014.992031`. PMID: 25559285.

36 B. Gardner and V. Tang. Reflecting on non-reflective action: an exploratory think-aloud study of self-report habit measures. *British journal of health psychology*, 19(2):258–273, 2014.

37 Thomas D. Wilcockson, David A. Ellis, and Heather Shaw. Determining typical smartphone usage: What data do we need? *Cyberpsychology, Behavior, and Social Networking*, 21(6):395–398, 2018.

38 D. Parry, B. Davidson, C. Sewall, J. Fisher, H. Mieczkowski, and D. Quintana. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour (forthcoming)*, 2021.

39 N. Harrington. Commentary: Why it doesn't pay to ask consumers about habitual behaviors. *Journal of the Association for Consumer Research: The Habit-Driven Consumer*, 2(3), 2017.

40 Sebastian Potthoff, Nicola McCleary, Falko F. Sniehotta, and Justin Presseau. Creating and breaking habit in healthcare professional behaviours to improve healthcare and health. In B. Verplanken, editor, *The Psychology of Habit*, pages 247–265. Springer, 2018.

41 Ryan E. Rhodes and Amanda L. Rebar. Physical activity habit: Complexities and controversies. In B. Verplanken, editor, *The Psychology of Habit*, pages 91–109. Springer, 2018.

42 Navin Kaushal and Ryan Rhodes. Exercise habit formation in new gym members: a longitudinal study. *Journal of Behavioral Medicine*, 38:652–663, 2015.

43 M. Fournier, F. d'Arripe Longueville, C. Rovere, C. S. Easthope, L. Schwabe, J. El Methni, and R Radel. Effects of circadian cortisol on the development of a health habit. *Health Psychology*, 36(11):1059–1064, 2017.

44 Benjamin Gardner and Phillippa Lally. Modelling habit formation and its determinants. In B. Verplanken, editor, *The Psychology of Habit*, pages 207–229. Springer, 2018.

45 Jan Gläscher, Nathaniel Daw, Peter Dayan, and O'Doherty John P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, 2010.

46 Peter Dayan and Kent Berridge. Model-based and model-free pavlovian reward learning: Revaluation, revision and revelation. *Cognitive Affective Behavioral Neuroscience*, 14(2):473–492, 2014.

47 N. Daw, S. Gershman, B. Seymour, P. Dayan, and R. Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

48 Wolfgang M Pauli, Jeffrey Cockburn, Eva R Pool, Omar D Pérez, and John P O'Doherty. Computational approaches to habits in a model-free world. *Current Opinion in Behavioral Sciences*, 20:104–109, 2018.

49 Henry H. Yin and Barbara J. Knowlton. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7:464–476, 2006.

50 B.J. Knowlton and T.K. Patterson. Habit formation and the striatum. In R.E. Clark and S. Martin, editors, *Behavioral Neuroscience of Learning and Memory. Current Topics in Behavioral Neurosciences, vol 37*. Springer, 2016.

51 Lars Schwabe and Oliver Wolf. Stress prompts habit behavior in humans. *The Journal of Neuroscience*, 29(22):7191–7198, 2009.

52 C. Camerer, P. Landry, and R. Webb. The neuroeconomics of habit. In A. Kirman and M. Teschi, editors, *The State of Mind in Economics*. In press, 2021.

53 S. Lee, Shinsuke Shimojo, and John O'Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699, 2014.

54 Anthony Dickinson, D. J. Nicholas, and Christopher D. Adams. The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, 35(1):35–51, 1983.

55 Aniek Fransen. Experimental considerations on habit formation in humans. Master's thesis, Maastricht University, 2019.

56 C. A. Seger and B. J. Spiering. A critical review of habit learning and the basal ganglia. *Frontiers in Systems Neuroscience*, 5(66), 2011.

57 Peter Bayley, Jennifer Franscino, and Larry R. Squire. Robust habit learning in the absence of awareness and independent of the medial temporal lobe. *Nature*, 436(7050):550–553, 2005.

58 S. Shi and L. Epstein. Habits and time preference. *International Economic Review*, 34(1):61–84, 1993.

59 Irving Fisher. *The theory of interest : as determined by impatience to spend income and opportunity to invest it*. New York: Macmillan Co., 1930.

60 J.S. Duesenberry. *Income, Saving and the Theory of Consumption Behavior*. Cambridge, Mass.: Harvard University Press, 1949.

61 Harl E. Ryder and Geoffrey M. Heal. Optimal growth with intertemporally dependent preferences. *The Review of Economic Studies*, 40(1):1–31, 1973.

62 Angus Deaton. *Understanding Consumption*. Oxford University Press, 1992.

63 Tjalling C. Koopmans. Stationary ordinal utility and impatience. *Econometrica*, 28(2):287–309, 1960.

64 John Y. Campbell and John H. Cochrane. By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, 107(2):205–251, 1999.

65 Suresh M. Sundaresan. Intertemporally dependent preferences and the volatility of consumption and wealth. *The Review of Financial Studies*, 2(1):73–89, 1989.

66 George M. Constantinides. Habit formation: A resolution of the equity premium puzzle. *Journal of Political Economy*, 98(3): 519–543, 1990.

67 B. D. Gonsalves, I. Kahn, T. Curran, K. A. Norman, and A. D. Wagner. Memory strength and repetition suppression: multimodal imaging of medial temporal cortical contributions to recognition. *Neuron*, 47(5):751–761, 2005.

68 Kareen Rozen. Foundations of intrinsic habit formation. *Econometrica*, 78(4):1341–1373, 2010.

69 Gary S. Becker and Kevin M. Murphy. A theory of rational addiction. *Journal of Political Economy*, 96(4):675–700, 1988.

70 Kaili Shen and David E. Giles. Rational exuberance at the mall: addiction to carrying a credit card balance. *Applied Economics*, 38(5):587–592, 2006.

71 M. C. Auld and P. Grootendorst. An empirical analysis of milk addiction. *Journal of health economics*, 23(6):1117–1133, 2004.

72 Ian Crawford. Habits revealed. *The Review of Economic Studies*, 77(4):1382–1402, 2010.

73 Hyeokkoo Eric Kwon, Hyunji So, Sang Han, and Wonseok Oh. Excessive dependence on mobile social apps: A rational addiction perspective. *Information Systems Research*, 27, 2016.

74 James Heckman. Heterogeneity and state dependence. In Sherwin Rosen, editor, *Studies in Labor Markets*, pages 91–140. National Bureau of Economic Research, Inc., 1981.

75 A. Kuehn. Consumer brand choice as a learning process. *Journal of Advertising Research*, 2:10–17, 1962.

76 M. P. Keane. Modeling heterogeneity and state dependence in consumer choice behavior. *Review of Economics and Statistics*, 15 (3):310–327, 1997.

77 Jean-Pierre Dubé, Günter J. Hitsch, and Peter E. Rossi. State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, 41(3):417–445, 2010.

78 Russell W. Belk. Situational variables and consumer behavior. *Journal of Consumer Research*, 2(3):157–164, 1975.

79 Klaus Wertenbroch. Consumption self-control by rationing purchase quantities of virtue and vice. *Marketing Science*, 17(4):317–337, 1998.

80 J. C. Middleton, R. A. Hahn, J. L. Kuzara, R. Elder, R. Brewer, S. Chattopadhyay, J. Fielding, T. S. Naimi, T. Toomey, and B. Lawrence. Effectiveness of policies maintaining or restricting days of alcohol sales on excessive alcohol consumption and related harms. *American journal of preventive medicine*, 39(6):575–589, 2010.

81 D. Laibson. A cue-theory of consumption. *Quarterly Journal of Economics*, 116(1):81–119, 2001.

82 B.D. Bernheim and A. Rangel. Addiction and cue-triggered decision processes. *American Economic Review*, 94:1558–1590, 2004.

83 E. Karni. State-dependent preferences. In Palgrave Macmillan, editor, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, London, 2008.

84 Kevin Volpp and George Loewenstein. What is a habit? diverse mechanisms that can produce sustained behavior change. *Organizational Behavior and Human Decision Processes*, 161:36–38, 2020.

85 Matthew Cravens. Measuring the strength of voter turnout habits. *Electoral Studies*, 64:102–117, 2020.

86 Richard A. Brody and Paul M. Sniderman. From life space to polling place: The relevance of personal concerns for voting behavior. *British Journal of Political Science*, 7(3):337–360, 1977.

87 Mark N. Franklin and Sara B. Hobolt. The legacy of lethargy: How elections to the european parliament depress turnout. *Electoral Studies*, 30(2):67–76, 2011.

88 Kevin Denny and Orla Doyle. Does voting history matter? analysing persistence in turnout. *American Journal of Political Science*, 53(1):17–35, 2009.

89 Donald P. Green and Alan S. Gerber. The downstream benefits of experimentation. *Political Analysis*, 10(4):394–402, 2002.

90 Lisa G. Bedolla and Melissa R. Michelson. *Mobilizing Inclusion: Transforming the Electorate Through Get-Out-the-Vote Campaigns*. Yale University Press, 2012.

91 Marc Meredith. Persistence in political participation. *Quarterly Journal of Political Science*, 4(3):187–209, 2009.

92 A. Coppock and D. P. Green. Is voting habit forming? new evidence from experiments and regression discontinuities. *American Journal of Political Science*, 60(4):1044–1062, 2016.

93 Anthony Downs. *An Economic Theory of Democracy*. New York: Harper Row, 1957.

94 Henry E. Brady and John E. McNulty. Turning out to vote: The costs of finding and getting to the polling place. *The American Political Science Review*, 105(1):115–134, 2011.

95 J. H. Aldrich, J. M. Montgomery, and W. Wood. Turnout as a habit. *Political Behavior*, 33(4):535–563, 2011.

96 Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

97 Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

98 Haohan Wang, Benjamin J Lengerich, Bryon Aragam, and Eric P Xing. Precision lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics*, 35(7):1181–1187, 2019.

99 Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31 (2):87–106, 2017.

100 K. Ersche, T-V. Lim, L. Ward, T. Robbins, and J. Stochl. Creature of habit: A self-report measure of habitual routines and automatic tendencies in everyday life. *Personality and Individual Differences*, 116:73–85, 2017.

101 NOAA. Us local climatological data: Global-hourly file access. us department of commerce. 2019. URL `https://www.ncei.noaa.gov/data/global-hourly/archive/`.

102 B. R. Staats, H. Dai, D. Hofmann, and K. L. Milkman. Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. *Management Science*, 63(5):1563–1585, 2016.

103 G. Judah, B. Gardner, and R. Aunger. Forming a flossing habit: an exploratory study of the psychological determinants of habit formation. *British journal of health psychology*, 18(2):338–353, 2013.

104 B. Gardner and P. Lally. Does intrinsic motivation strengthen physical activity habit? modeling relationships between self-determination, past behaviour, and habit strength. *Journal of behavioral medicine*, 36(5):488–497, 2013.

105 L. A. Phillips and B. Gardner. Habitual exercise instigation (vs. execution) predicts healthy adults' exercise frequency. *Health Psychology*, 35(1):69, 2016.

106 G. Nyberg, E. Sundblom, Å. Norman, B. Bohman, J. Hagberg, and L. S. Elinder. Effectiveness of a universal parental support programme to promote healthy dietary habits and physical activity and to prevent overweight and obesity in 6-year-old children: the healthy school start study, a cluster-randomised controlled trial. *PloS one*, 10(2):e0116876, 2015.

107 B. Gardner, K. Sheals, J. Wardle, and L. McGowan. Putting habit into practice, and practice into habit: a process evaluation and exploration of the acceptability of a habit-based dietary behaviour change intervention. *International Journal of Behavioral Nutrition and Physical Activity*, 11(1):1–13, 2014.

108 Lena Fleig, Megan M. McAllister, Peggy Chen, Julie Iverson, Kate Milne, Heather A. McKay, Lindy Clemson, and Maureen C. Ashe. Health behaviour change theory meets falls prevention: Feasibility of a habit-based balance and strength exercise intervention for older adults. *Psychology of Sport and Exercise*, 22:114–122, 2016.

109 G. J. de Bruijn, R. E. Rhodes, and L. van Osch. Does action planning moderate the intention-habit interaction in the exercise domain? a three-way interaction analysis investigation. *Journal of behavioral medicine*, 35(5):509–519, 2012.

110 R. Matei, Thuné-Boyle, I., M. Hamer, S. Iliffe, K. R. Fox, B. J. Jefferis, and B. Gardner. Acceptability of a theory-based sedentary behaviour reduction intervention for older adults ('on your feet to earn your seat'). *BMC Public Health*, 15(1):1–16, 2015.

111 V. Storm, J. Dörenkämper, D. A. Reinwand, J. Wienert, H. De Vries, and S. Lippke. Effectiveness of a web-based computer-tailored multiple-lifestyle intervention for people interested in reducing their cardiovascular risk: a randomized controlled trial. *Journal of medical Internet research*, 18(4):e5147, 2016.

112 S. Potthoff, J. Presseau, F. F. Sniehotta, M. Johnston, M. Elovainio, and L. Avery. Planning to be routine: habit as a mediator of the planning-behaviour relationship in healthcare professionals. *Implementation Science*, 12(1):1–10, 2017.

113 L. Fleig, S. Pomp, L. Parschau, M. Barz, D. Lange, R. Schwarzer, and S. Lippke. From intentions via planning and behavior to physical exercise habits. *Psychology of Sport and Exercise*, 14(5):632–639, 2013.

114 A. U. Wiedemann, B. Gardner, N. Knoll, and S. Burkert. Intrinsic rewards, fruit and vegetable consumption, and habit strength: A three-wave study testing the associative-cybernetic model. *Applied psychology: Health and well-being*, 6(1):119–134, 2014.

115 N. Kliemann, V. Vickerstaff, H. Croker, F. Johnson, I. Nazareth, and R. J. Beeken. The role of self-regulatory skills and automaticity on the effectiveness of a brief weight loss habit-based intervention: secondary analysis of the 10 top tips randomised trial. *International Journal of Behavioral Nutrition and Physical Activity*, 14(1):1–11, 2017.

116 G. Cleo, P. Glasziou, E. Beller, E. Isenring, and R. Thomas. Habit-based interventions for weight loss maintenance in adults with overweight and obesity: a randomized controlled trial. *International Journal of Obesity*, 43(2):374–383, 2019.

117 I. Pfeffer and T. Strobach. Behavioural automaticity moderates and mediates the relationship of trait self-control and physical activity behaviour. *Psychology  Health*, 33(7):925–940, 2018.

118 M. Piao, H. Ryu, H. Lee, and J. Kim. Use of the healthy lifestyle coaching chatbot app to promote stair-climbing habits among office workers: exploratory randomized controlled trial. *JMIR mHealth and uHealth*, 8(5):e15085, 2020.

119 P. Karppinen, H. Oinas-Kukkonen, T. Alahäivälä, T. Jokelainen, A. M. Teeriniemi, T. Salonurmi, and M. J. Savolainen. Opportunities and challenges of behavior change support systems for enhancing habit formation: A qualitative study. *Journal of biomedical informatics*, 84:82–92, 2018.

120 G. J. de Bruijn, B. Gardner, L. van Osch, and F. F. Sniehotta. Predicting automaticity in exercise behaviour: the role of perceived behavioural control, affect, intention, action planning, and behaviour. *International Journal of Behavioral Medicine*, 21(5):767–774, 2014.

121 S. Maltagliati, A. Rebar, L. Fessler, C. Forestier, P. Sarrazin, A. Chalabaev, D. Sander, H. Sivaramakrishnan, D. Orsholits, M. P. Boisgontier, N. Ntoumanis, B. Gardner, and B. Cheval. Evolution of physical activity habits after a context change: The case of covid-19 lockdown. *British journal of health psychology*, 26(4):1135–1154, 2021.

122 P. Haggar, L. Whitmarsh, and S. M. Skippon. Habit discontinuity and student travel mode choice. *Transportation research part F: traffic psychology and behaviour*, 64:1–13, 2019.

123 J. Keller, D. Kwasnicka, P. Klaiber, L. Sichert, P. Lally, and L. Fleig. Habit formation following routine-based versus time-based cue planning: A randomized controlled trial. *British Journal of Health Psychology*, 26(3):807–824, 2021.

124 A. Van der Weiden, J. Benjamins, M. Gillebaart, J. F. Ybema, and D. De Ridder. How to form good habits? a longitudinal field study on the role of self-control in habit formation. *Frontiers in Psychology*, 11:560, 2020.

125 A. Schnauber-Stockmann and T. K. Naab. The process of forming a mobile media habit: Results of a longitudinal study in a real-world setting. *Media Psychology*, 22(5):714–742, 2019.

126 T. Bartle, B. Mullan, E. Novoradovskaya, V. Allom, and P. Hasking. The role of choice in eating behaviours. *British Food Journal*, 121(11):2696–2707, 2019.

127 L. A. Phillips, M. Johnson, and K. R. More. Experimental test of a planning intervention for forming a 'higher order' health-habit. *Psychology  health*, 34(11):1328–1346, 2019.

128 Diego Garaialde, Christopher P. Bowers, Charlie Pinder, Priyal Shah, Shashwat Parashar, Leigh Clark, and Benjamin R. Cowan. Quantifying the impact of making and breaking interface habits. *International Journal of Human-Computer Studies*, 142:102461, 2020. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2020.102461. URL `https://www.sciencedirect.com/science/article/pii/S107158192030063X`.

129 J. Jaakkola, R. Virtanen, T. Vasankari, M. Salminen, and K. E. Airaksinen. Self-detection of atrial fibrillation in an aged population: three-year follow-up of the lietoaf intervention study. *BMC geriatrics*, 17(1):1–8, 2017.

130 R. J. van Bree, C. Bolman, A. N. Mudde, M. M. van Stralen, D. A. Peels, H. de Vries, and L. Lechner. Modeling longitudinal relationships between habit and physical activity: two cross-lagged panel design studies in older adults. *Journal of Aging and Physical Activity*, 25(3):464–473, 2017.

131 M. Stojanovic, A. Grund, and S. Fries. App-based habit building reduces motivational impairments during studying–an event sampling study. *Frontiers in Psychology*, 11:167, 2020.

132 R. J. van Bree, A. N. Mudde, C. Bolman, M. M. van Stralen, D. A. Peels, H. de Vries, and L. Lechner. Are action planning and physical activity mediators of the intention-habit relationship? *Psychology of Sport and Exercise*, 27:243–251, 2016.

133 S. Weyland, E. Finne, J. Krell-Roesch, and D. Jekauc. (how) does affect influence the formation of habits in exercise? *Frontiers in psychology*, 11:2866, 2020.

134 S. Di Maio, J. Keller, D. H. Hohl, R. Schwarzer, and N. Knoll. Habits and self-efficacy moderate the effects of intentions and planning on physical activity. *British Journal of Health Psychology*, 26(1):50–66, 2021.

135 H. Fritz, W. Tarraf, A. Brody, and P. Levy. Feasibility of a behavioral automaticity intervention among african americans at risk for metabolic syndrome. *BMC public health*, 19(1):1–12, 2019.

136 J. P. Maher, A. L. Rebar, and G. F. Dunton. The influence of context stability on physical activity and sedentary behaviour habit and behaviour: An ecological momentary assessment study. *British Journal of Health Psychology*, 26(3):861–881, 2021.

137 K. Byrka and K. Kaminska. Doing laundry with biodegradable soap nuts: Can rare and novel behaviors break bad habitual patterns? *Journal of Environmental Psychology*, 79:101730, 2022.

138 C. Gravert and L. O. Collentine. When nudges aren't enough: Norms, incentives and habit formation in public transport usage. *Journal of Economic Behavior  Organization*, 190:1–14, 2021.

139 M. Stojanovic, S. Fries, and A. Grund. Self-efficacy in habit building: How general and habit-specific self-efficacy influence behavioral automatization and motivational interference. *Frontiers in Psychology*, 12:643753, 2021.

140 J. W. Davis. Physical activity habit formation through a technology-based program. *Journal of the American Association of Nurse Practitioners*, 32(7):540–546, 2020.

141 P. Banca, D. McNamee, T. Piercy, Q. Luo, and T. W. Robbins. A mobile phone app for the generation and characterization of motor habits. *Frontiers in Psychology*, 10:2850, 2020.

142 T. Kirchner, J. Cantrell, A. Anesetti-Rothermel, O. Ganz, D. Vallone, and D. Abrams. Geospatial exposure to point-of-sale tobacco: Real-time craving and smoking cessation outcomes. *American Journal of Preventative Medicine*, 45(4), 2013.

143 B. Balleine and A. Dezfouli. Hierarchical action control: Adaptive collaboration between actions and habits. *Frontiers in Psychology*, 10(1):2735, 2019.

144 S. Fleetwood. A definition of habit for socio-economics. *Review of Social Economy*, 79(2):131–165, 2021.

145 H. Dai, K.L. Milkman, and J. Riis. The fresh start effect: Temporal landmarks motivate aspirational behavior. *Management Science*, 60(10):2563–2582, 2014.

146 H. Dai, K. L. Milkman, D. A. Hofmann, and 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. *Journal of Applied Psychology*, 100(3):846–862, 2015.

147 W. Wood and D. Neal. A new look at habits and the habit-goal interface. *Psychological Review*, 114(4):843–863, 2007.

148 D. Neal, W. Wood, and J. Quinn. Habits—a repeat performance. *Current Directions in Psychological Science*, 15(4):198–202, 2006.

149 W. Wood, J. M. Quinn, and D. A. Kashy. Habits in everyday life: Thought, emotion, and action. *Journal of Personality and Social Psychology*, 83(6):1281–1297, 2002.

150 Wendy Wood, Melissa G. Witt, and Leona Tam. Changing circumstances, disrupting habits. *Journal of Personality and Social Psychology*, 88(6):918–933, 2005.

151 Bas Verplanken, Oddgeir Friborg, Catharina E. Wang, David Trafimow, and Kristin Woolf. Mental habits: Metacognitive reflection on negative self-thinking. *Journal of Personality and Social Psychology*, 92(3):526–541, 2007.

152 B. Verplanken. Introduction. In B. Verplanken, editor, *The Psychology of Habit*, pages 1–10. Springer, 2018.

153 Barbara Mullan and Elizaveta Novoradovskaya. Habit mechanisms and behavioural complexity. In B. Verplanken, editor, *The Psychology of Habit*, pages 71–90. Springer, 2018.

154 O. D. Perez and A. Dickinson. A theory of actions and habits: The interaction of rate correlation and contiguity systems in free-operant behavior. *Psychological Review*, 127(6):945–971, 2020.

155 M. R. Steinfeld and M. E. Bouton. Context and renewal of habits and goal-directed actions after extinction. *Journal of Experimental Psychology: Animal Learning and Cognition*, 46(4):408–421, 2020.