Learning by Doing: The Habituation of Volunteers in a Virtual Citizen Science

Social computing systems depend on contributions from users. For many social computing systems, learning plays an important role in the quality of contributions and user's willingness to continue contributing. In crowdsourcing platforms such as Wikipedia, understanding how and when people learn to make contributions is a crucial step to understanding which users can be relied on to make valuable contributions and users who need additional guidance. Furthermore, a user's ability to learn may impact their willingness to continue contributing. Habituation is stated to be the simplest form of learning and since the 1960s has been used to describe how people interact with stimuli. In this paper, we use habituation as a theoretical lens to investigate the learning behaviors of users who contribute to virtual citizen science projects. We describe the relationship between habituation and performance and habituation and retention. We find that habituation is a useful lens to characterize learning and can help identify high performers and sustained contributors. Our findings shed light on the process of habituation for similar types of crowd work.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; *Empirical studies in HCI*.

Additional Key Words and Phrases: citizen science, habituation, human behaviors, learning, retention

ACM Reference Format:
. 2022. Learning by Doing: The Habituation of Volunteers in a Virtual Citizen Science. In . ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Over the last decade, advances in information and communication technologies (ICTs) have facilitated a rise in the number and variety of citizen science projects. Across many scientific disciplines including entomology, genomics, botany, ornithology, psychology, neuroscience, and socio-linguistics professional scientists rely on the "wisdom of the crowd" to help conduct research. Citizen science supports collaboration between amateurs and professional scientists and involves users in one of many stages of inquiry including question formulation, data collection, data analysis, and writing up results [3, 21]. Examples include, helping biochemists discover novel protein structures by folding protein cells in FoldIt [6] and observing the presence, variety, and quantity of birds to help ornithologists investigate bird migratory patterns [17].

Citizen science is a method for engaging the public in scientific research projects and can be conducted in-person, online, or a hybrid approach that involves both in-person and online interactions. Citizen science projects are often categorized by the type of collaboration and between volunteers and scientists - *contributory* projects where volunteers are engaged primarily in contributing data, *collaborative* projects, where in addition to contributing data, volunteers work with scientists develop research questions, analyze data, or disseminate findings, and *co-created* projects, where scientists and volunteers work together in most or all steps of the scientific process Bonney et al. [3]. Contributory projects, the focus of this research, often engage volunteers in activities developed by a science team. For example, the Galaxy Zoo is a project where amateurs help professional

CSCW '22, Date, Location

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

astronomers understand how galaxies are formed by reviewing and categorizing images. The images were captured by the Hubble Space Telescope [8] during the Sloan Digital Sky Survey (SDSS). The training supplied to volunteers instruct them to review each image and supply judgments (i.e., classifications) about the galaxy's shapes. Volunteers are presented questions in a decision-tree e.g., "Is the galaxy simply smooth and rounded, with no signs of a disk?" and "How rounded is it?" Volunteers choose their answers from a list of choices and each image receives judgments from multiple volunteers. When enough volunteers have classified a galaxy image, a consensus label is applied to the image and the date are relayed to astronomers who conduct additional analysis on the dataset.

Other contributory projects may ask volunteers to transcribe writing e.g., Old Weather or trace boundaries around phenomena that appear in images e.g., Floating Forest. Tasks in Galaxy Zoo, Floating Forest, and others involve some level of pattern recognition [11]. In Galaxy Zoo, volunteers need to develop mental models of what constitutes smoothness or roundness. For some contributors, these mental models along the patterns they describe may be non-existent and become established through training. In fact, many citizen science projects require some training before contributing. Training and auxiliary learning materials (e.g., field guides) are necessary to help volunteers develop skills for detecting and classifying phenomena in images [16]. In addition to training and auxiliary materials, volunteers may develop the necessary skills in the wild or in this context, while they are classifying images. Some patterns may be ambiguous and impossible to articulate in traditional learning materials. In that sense, volunteers can learn by doing and in the process create and update existing mental models around patterns. For training that involves "learning by doing" there are undoubtedly challenges in ensuring volunteers make quality judgements. Learning to recognize patterns and building competency in consistent identification of patterns requires time. For some projects, during the learning process real data are intertwined with those used for training (or gold standard data) which can be used to assess performance. This format means prior to learning, there may be patterns for which volunteers have a deficient mental model.

To better understand the implications of "learning by doing," we use the theory of habituation to describe the characteristics of learning in a citizen science project. Habituation describes the period at which responsiveness to stimuli decreases [2, 15, 18]. Habituation emerged from the field of psychology and has been described as one of the simplest forms of learning. In the context of citizen science, stimuli are the images being displayed to volunteers and responsiveness is the time a volunteer spends classifying the image. It is expected that as volunteers engage in learning, they enhance their mental models of patterns, decrease the amount of time reasoning about the characteristics of phenomena in an image and as an outcome of this habituation are more accurate in their judgements. In this paper, we (1) describe the characteristics of habituation (i.e., learning) and (2) examine important behaviors in citizen science projects - performance and retention. Accordingly, our research questions are:

RQ1: What are the characteristics of habituation to pattern recognition tasks?
RQ2a-b: How is habituation related to performance and retention?

2 THEORY: HABITUATION

Habituation describes the process by which responsiveness to stimuli decreases [2, 15, 18] resulting in changes behavioral engagements with stimuli. Habituation is a mode of behavior that is acquired through repetitive action or thought results in decreased responsiveness of individuals' attention after repeated exposure [4, 13, 23] and allows humans to identify and filter out irrelevant stimuli. Experimental and observational studies of habituation have demonstrated that humans elicit a host of behavioral responses to interactions with stimuli. The earliest writing on habitation theory can be traced to the discipline of psychology where scholars conducted experiments to understand reactivity to stimuli. Scholars recognize habituation as a basic form of learning since habituation can lead to a reduction in uncertainty and or conflict [14]. We also know that humans exhibit habituation to

a variety of objects including those containing visual and auditory stimuli. More recently, advances in brain imaging technologies such as functional magnetic resonance imaging (fMRI) have advanced our collective knowledge about habituation's physiological responses. These studies confirm existing research while with respect to behavioral responses while also finding that humans experience a decrease in neuronal activation in with repeated stimuli exposure [9, 24].

Thompson and Spencer [18] surveyed the extensive literature on habituation containing experimental and observational studies and developed a set of rules outlining the mechanisms of habituation. These rules were updated some decades later to account of new research on the topic [2, 15]. The rules describe the general tendencies of human responses to stimuli with respect to habituation. In general, the habituation rules posit that stimuli lose their effectiveness over time resulting in habituation and that response time tends to recover (revert to earlier observed times) after periods of non-stimuli exposure. The first rule for example states that exposure to stimuli should result in a decline in the magnitude of a response parameters to an asymptotic level. The rule further suggests that the decrease will be negative exponential function of the number of stimuli presented, however, revisions in Rankin et al. [15] suggest other relationships are possible e.g., linear. The other nine rules are articulated in Rankin et al. [15] and Thompson and Spencer [18].

Habituation plays an important function in explaining human interaction with technologies [12, 22]. A dearth of research exists in the field of computer supported cooperative work that examines how habituation materializes in settings like citizen science (or social computing systems more generally). A search of "habituation" in proceedings since 2015 yields six results. In most cases, habituation is used to explain behaviors and not treated as a lens to examine behaviors. A search in human-computer interaction (HCI) yields a handful of studies focused almost exclusively on using habituation to explain the limited effectiveness of privacy and security warnings [1, 5, 7, 19] These studies tend to coalesce around the finding that people tend to ignore warnings as a result of habituation. A handful of studies point to strategies that designers could implement to mitigate the effects of habituation on people's privacy behaviors. For instance, Bravo-Lillo et al. [5] reduced habituation to security warnings by changing the appearance of dialogue boxes on users' computers during software installation and Anderson et al. [1] used fMRI to study how habituation occurs in the brain and found that repeated exposure to warnings caused decreases in visual processing activity in the brain; however, when warnings varied, people tended to habituate later. Beyond privacy and security warnings, other studies have emerged e.g., social media [22] and video games [12], however, a robust body of theoretical discussions and empirical examples in the computing field is lacking.

To our knowledge, habituation theory has not been used as a theoretical lens with which to examine the behaviors of citizen science volunteers. Citizen science shares many qualities that make it a fertile ground for exploration in the computing field. Volunteers are repeatedly exposed to a stimulus i.e., images that contain phenomena of interests to science teams. The point of departure in the exploration of habituation in this context is that there are many phenomena spread out across time both within and between sessions.

Questions remain about when exactly habituation occurs and whether habituation is necessarily desirable in this context The nature of some citizen science tasks - having many stimuli with often ambiguous patterns makes it difficult to know the relationship between habituation (decreased engagement with stimuli) and performance. Additionally, little is known about how habituation affects a persons desire to contribute since tasks can be monotonous.

3 GRAVITY SPY

Gravity Spy is a virtual citizen project hosted on the Zooniverse platform. The project involves members of the public in astrophysics research. The search for gravitational waves is conducted by the LIGO Scientific Collaboration (LSC), a group of researchers who use powerful interferometers to detect the presence of gravitational waves. A challenge for LIGO scientists is the high sensitivity of the detectors, which is expected to look for gravitational

waves, yet in addition, brings about recording a huge amount of noise (called glitches). The glitches jumble or even take on the appearance of gravitational-wave signals, diminishing the viability of the pursuit.

Gravity Spy participants are tasked with categorizing glitches which help scientists discover and isolate glitches from the gravitational wave data stream. The noise appears as normalized energy (yellow spots) in images. Subjects are periodically uploaded to the Gravity Spy platform and queued up to volunteers in an interface displayed in (Figure 1). On the left side of the interface, volunteers are shown a subject that could contain a glitch. The subject can be viewed along several axes with time on the x-axis, noise frequency on the y-axis, and normalized energy on the second y-axis. Volunteers also have access to tools that allow them to explore the subject further, including an information button that displays additional metadata about the subject (timestamp, q-value) and a play button to shuffle through four images with contiguous times.

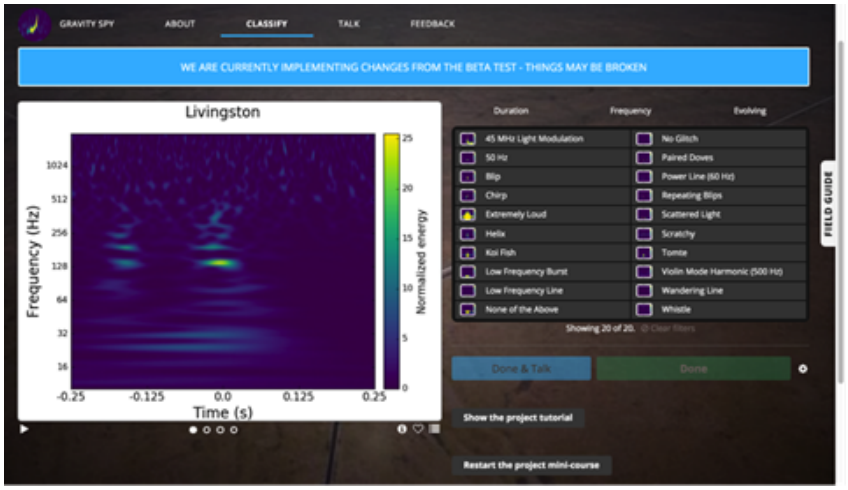


Fig. 1. The Gravity Spy classification interface.

A series of glitch options are displayed on the right side of the interface. Volunteers browse the different glitch class options and determine which category best matches the noise profile of the glitch displayed in the subject. While the classification task appears straightforward the characteristics of noise patterns represented in images can be quite confusing without having an expert eye. At the time of writing, there are 20 known glitch classes. While there are many glitch options, participation is scaffolded such that volunteers are gradually exposed to new glitch classes. Displayed in Table 1 are the glitch class options available in each workflow with new glitches in bold font. After making a judgment about the similarity between the subject and glitch class options, the volunteer can click "Done" (volunteers receive a new subject) to submit the classification or "Done & Talk" redirects to a Talk page where the volunteer can engage in discussions about the same image they just classified. The system records every answer supplied by a volunteer and a timestamp indicating the date and time the response was posted to the system.

3.1 Machine-learning guided training and stimuli

The stimuli used for this study are the images served to volunteers during the classification task. Gravity Spy uses a machine-learning guided training (MLGT) [10] regime that trains volunteers to recognizing the characteristics of glitch class options. Exposure to glitch class options is scaffolded in the MLGT using two features. First, volunteers are only introduced to a handful of glitches in each workflow. Each volunteer's performance on a

set of gold standard glitch class options is periodically assessed to determine whether their accuracy has met a threshold for performance which determine whether they are promoted to the next workflow. Second, volunteers are introduced to glitches that are similarly difficult. A computer vision algorithm assesses the likelihood of each image subject belonging to each of the twenty glitch class options. Displayed in (Figure 2) are glitches that Gravity Spy's computer vision algorithms have evaluated. The image shows the ML confidence score for the images likelihood of belonging to the blip glitch class. The first image subject received machine confidence of 98% while the last image received an ML confidence score of 34%. In the MLGT, only the first image would to be included in workflow 1. The other glitches might appear in more advanced workflows facilitating a gentle introduction into learning the morphological features of a blip. Third, some image subjects have been evaluated by experts (i.e., professional astrophysicists) who have assigned a gold label for the image. These gold standard data are used to assess the performance of volunteers at each workflow. As volunteers classify data, they are periodically shown gold standard classifications. The system informs them whether the expert agreed with their annotation and provides the correct response if the volunteer's answer was incorrect. Volunteers' responses to gold standard classifications are used in a promotion algorithm that determines whether a volunteer should be promoted to the next workflow or remain in the current one.

4 METHOD

In the section below, we describe our methodology - first revealing the data we collected and transformations then we describe our data analysis methods and models we used therein. Using the data from Gravity Spy, we can have a good understanding of how the different kinds of questions (e.g., fundamental, advance) given at different phases of participation (e.g., newcomer) influence the habituation process. Which can help people to better understand the habituation rules.

4.1 Data Collection

The data we retrieved for this research were accessed on December 01, 2020. The data come from two sources - a database dump containing records of classifications submitted to the Gravity Spy platform by volunteers and another database dump containing metadata associated with each glitch image in the Gravity Spy system. The initial dataset contained 5,505,865 classifications. Each record in the dataset represents a classification executed by a single volunteer. The system records metadata and contains metadata about the classification such as the username of the volunteer who classified the image, a timestamp indicating when the classification was submitted to the system.

Workflow (# New Glitches)	**Glitches available**
Workflow 1 (2)	**blip, whistle**
Workflow 2 (3)	blip, whistle, **koi fish, power line, violin mode**
Workflow 3 (4)	blip, whistle, koi fish, power line, violin mode, **chirp, low frequency burst, no glitch, scattered light**
Workflow 4 (10)	blip, whistle, koi fish, power line, violin mode, chirp, low frequency burst, no glitch, scattered light, **helix, 45Mhz light modulation, low frequency noise fluctuations, paired doves, 50hz, repeating blips, scratchy, tomte, wandering line, extremely loud**

Table 1. The glitches available to volunteers in each workflow. The glitches in bold are new in that workflow.

Fig. 2. An example of glitch images is shown to volunteers with varying machine confidence scores. A computer vision algorithm calculates the likelihood that each image belongs to one of twenty-one glitch class options. Image subjects with high machine learning confidence are shown in early workflows where novices begin participating while image subjects with low ml-confidence scores are shown in more advanced workflows with more learned volunteers.

We made minor modifications to the dataset to ensure the ecological validity of our results. First, since the project had undergone a beta testing phase beginning in April 2016 to the project launch date (October 12, 2016). During the beta testing phase participants were invited to test the interface and make classifications, we removed data collected during this period. Second, since users who participated during the beta testing phase have advanced knowledge of the interface and the glitch classes being presented (an opportunity to learn) we removed all classifications contributed by that population of users. Third, we removed the classification records of non-ordinary users i.e., science team members, researchers (such as ourselves), and software developers. Since these categories of users helped build the platform and have expert knowledge of the glitch classes, their habituation may be shrouded in knowledge attained prior to the launch date. Our final dataset contained 5,111,010 classification records spanning a period of 3 years and 317 days beginning October 12, 2016 and ending August 20, 2020.

4.2 Outcome variables

Our analysis relies on an assessment of two variables that we imputed from the data - a response parameter and accuracy.

Response time Habituation is a mental state, making it difficult to see, however, many studies consider response times to stimuli as an approach to measure habituation. Consistent with theory and other empirical studies, we operationalized habituation as the time a volunteer spends examining an image prior to submitting the classification to the system. The dataset contained two timestamps - page load time indicating when the image was rendered on the volunteer's screen and a time indicating when the classification was posted to the system.

Accuracy The second research question seeks to understand the relationship between habituation and accuracy as demonstrated by a volunteer's ability to correctly select the right glitch class option on gold standard classifications and their agreement with other volunteers. We assessed how well volunteers

performed the classification task by assessing two measures of accuracy: (1) performance on gold-standard labels for subjects derived by experts (N = 526,201) and a (2) combination of machine labels derived by a computer vision algorithm and consensus labels derived by the Gravity Spy volunteers. The gold labels were classified by expert astrophysicists who help manage Gravity Spy. The gold standard data are shown to volunteers at irregular intervals and their answers are supplied to a user algorithm that determines whether they can be promoted to the next workflow. They used a volunteer consensus score to determine accuracy for all other subjects. This approach is naive in that even though most volunteers could classify a subject as belonging to a particular class, their consensus rating could be incorrect. To ensure this approach would lead to acceptable answers, we assessed agreement with the expert labels.

5 DATA ANALYSIS

To answer RQ1, on the characteristics of habituation, used two complementary methods - trend and change-point detection. The trend analysis helps us to determine whether there was a significant trend in the slope of the line. We used the sens.slope() function which takes a numeric vector and computes a linear rate of change and corresponding confidence intervals according to the Theil–Sen estimator method. The results of the Sen's slope allow us to compare the rate at which users habituate to a particular glitch class. To measure the point at which users habituate, we chose to conduct a change point analysis using the Pettitt algorithm. As with most data originating from social processes, we find that parametric analytic methods are not sufficient. In many cases, violations of normality assumptions were detected. The Pettitt method is a rank-based non-parametric test that measures abrupt changes in the central tendency of time series (in this instance the response parameter). The algorithm uses the Mann-Whitney statistic for testing if two samples (before and after a change point) come from the same distribution. The change-point that maximizes the statistic is then chosen as a single change point. In the results, we report habituation for each glitch in the workflow in which the glitch was first introduced and describe the results in the context of each rule.

We used standard group tests to answer RQ2-a and RQ2-b. We wanted to determine whether habituation to the was associated with improved accuracy, indicating that volunteers learned to identify glitches in images. We conducted our analysis using standard group means-testing. As is protocol, we checked whether assumptions associated with each statistical test were violated. We conducted an F test to assess whether the equality of variance assumption holds, and we conducted a Shapiro test and examined the normal Q-Q plot for each outcome variable to determine whether the normality assumption holds. In instances where assumptions were not met, we conducted analyses using a non-parametric version. For RQ2-b, we compared the proportions of volunteers who remained contributors after the habituation period. Again, using accuracy as an outcome variable, we evaluated the performance trajectory prior to habituation with their long-term status (drop out or sustained) prior to habituation as a grouping variable.

6 RESULTS

6.1 What volunteers do in Gravity Spy

Reported in table 2 are the descriptive statistics for volunteer contributions to each workflow. Included in the first column is the average number of classifications volunteers execute prior to being promoted to that workflow - 57 for workflow 2, 154 for workflow 154, and 462 for workflow 3. As is typical of many online communities, few people remain active beyond the first session and in Gravity Spy, 48% stop contributing before being promoted and only 15% of volunteers reach the last workflow. On average, a volunteer contributes 74 (σ = 94.1, $\widetilde{x}$ = 42) classifications during their tenure in Gravity Spy. The data in table 2 also reveal that the largest portion of the work executed by volunteers is submitting in the first workflow (42%) attesting to the importance of early contributions.

Workflow (promoted classification)	**Volunteers**	**Classifications**	**Response** (σ,$\widetilde{x}$)	**Accuracy** (%)
Neutron Star Mountain (-)	18,504	629,462	87.4 (4219, 7)	96
Galactic Supernova (57)	9,659	281,738	81.2 (3153, 8)	95
Binary Neutron Star Merger (154)	5,847	330,341	113 (4990, 8)	88
Binary Neutron Star Merger (462)	2,826	243,269	189 (6451, 11)	89

Table 2. Contribution statistics and aggregated behaviors in each Gravity Spy workflow.

6.2 RQ1: The characteristics of habituation

The first rule of habituation postulates "repeated applications of the stimulus result in decreased response (habituation). The decrease is usually a negative exponential function of the number of stimulus presentations." To investigate this relationship in our data, we examined habituation during the first five hundred classifications (approximately the time it takes to be promoted to the final workflow). Based on the results in table 2, we find an increase in the amount of time volunteers spend providing judgements - 87.4 sec. (σ = 4,219, $\widetilde{x}$ = 7) in workflow 1, 81.2 sec. (σ = 3,153, $\widetilde{x}$ = 8) in workflow 2, 113 sec. (σ = 4,990, $\widetilde{x}$ = 8)in workflow 3, and 87.4 sec. (σ = 6451, $\widetilde{x}$ = 11) in workflow 4. Given the distribution of response values, the median may be a more useful measure of the changes to the response parameter.

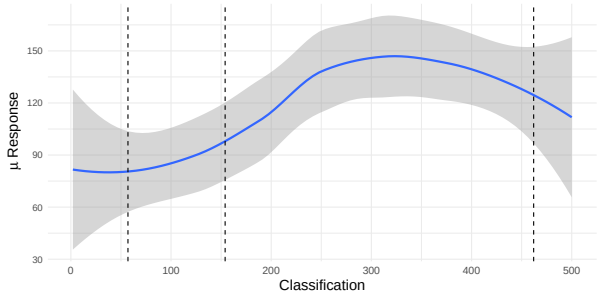


Fig. 3. Smoothed line the response parameter (time spend classifying) for the first five hundred classifications in Gravity Spy. The figure has vertical lines which indicate average classification when a volunteer is promoted.

We plotted the response parameter in Figure 3. The graph shows a smoothed line representing the response parameter over the first five hundred classifications. Each vertical line also represents the mean promoted classification. We excluded the first classification from this data since during the image display, volunteers are asked to complete a short training. Visual inspection of the graph shows an increasing trend for the response parameter from workflows 2 and 3. We also observe a decreasing parameter only after workflow 3. Over the observation, we find only a small increase in the response parameter (0.004). The results of the Sen's Slope estimation were not significant ($z = 0.22$, $p = 0.823$) and the results of the Pettitt test shows that a change in the response parameter occurred on the 192nd classification, however, this change-point was not significant. (U = 4376, $p = 0.794$).

To better understand the nuanced characteristics of habituation, we examined response time for each new glitch in a workflow. Figure 4 depicts the response parameter over glitch classes across fifty classifications in the workflow the glitch class was first introduced. The line helps determine the slope of the response indicating how quickly the volunteer response changes from its original trajectory. Depicted in Figure 4 is the log10 median response (for visualization purposes) time during analysis of the first fifty classifications and reported in Table 3

are the results of our analysis of the response parameter for each glitch class option. We took the response time for each glitch class and ran the Theil–Sen estimator over the series of response times. Our results are presented in Table and reveal that the response parameter for most glitch classes have a significant negative slope (except Air Compressor), indicating that with repeated exposure, volunteers respond more rapidly (or habituate) to glitch classes.

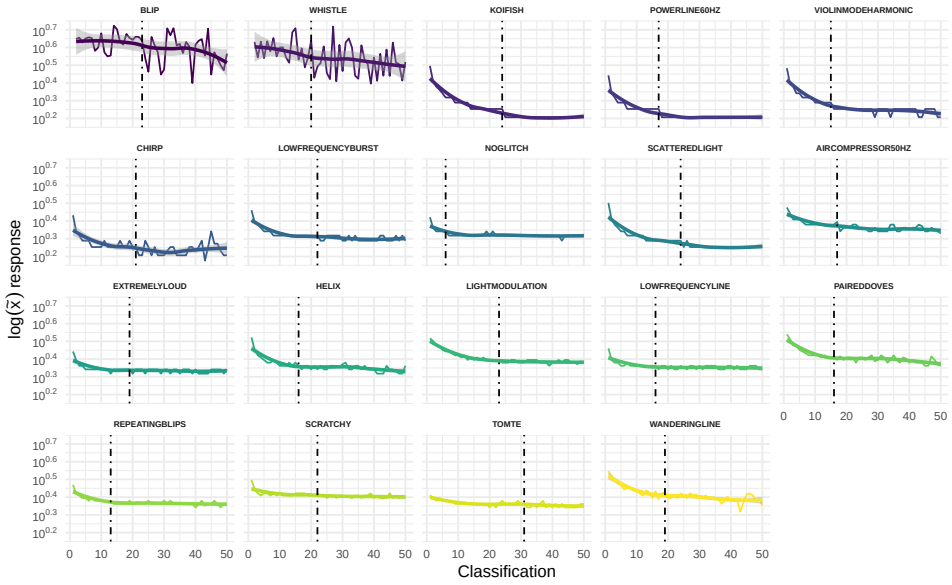


Fig. 4. Line charts showing the log10 median response time for the first fifty occasions that a glitch class option is shown to a volunteer. The dashed vertical line indicates the change point (or habituation) for each glitch class option.

Next, we used the Pettitt Change-point estimate to determine when the change in the central tendency of the response parameter occurs. These results are also shown in Table and again, are grouped by workflow. For most glitch classes, we observed a significant change-point. Habituation in each workflow for each glitch class option appears to happen within a range of exposure periods - between the 17th and 21st classification in workflow 1, the 15th and 24th in workflow 2, the 21st and 24th in workflow 3, and he 13th and 23rd in workflow 4.

6.3 RQ2-a: Habituation and performance

Turning to performance, we find that the accuracy decreases slightly (0.0006) as volunteers level-up – 95% in workflow 1, 95% in workflow 2, 88% in workflow 3, and 89% in workflow 4. Figure 5 shows accuracy across the first three hundred images. Slope estimation was performed using the Sen's Slope estimator and the results showed a significant downward trend (z = -26.66, $p < 0.001$). We also used the Pettitt test to detect when a significant change in the trend of accuracy occurs, we detected that a change in the response parameter around the 235th ($U = 59171$, $p < 0.001$).

Much like the response parameter, a closer examination of accuracy is necessary. Habituation does not necessarily indicate learning. Since habituation is said to trigger an automatic response, it could be that volunteers habituate, but fail to actual learn a glitch's pattern. We address the question of whether habituation leads to differences in accuracy before and after the habituation period we identified in the previous section and reported

	Theil–Sen estimator	**Pettitt Change-point**	
	Slope trend (z)	Estimate	U
	Neutron Star Mountain		
Blip	**-7.04***	**21**	**609***
Whistle	**-6.98** *	**17**	**561***
	Galactic Supernova		
Koi Fish	**-7.57***	**24**	**624***
Power line	**-6.39***	**17**	**561***
Violin mode	**-6.65***	**15**	**525***
	Binary Neutron Star Merger		
Chirp	**-3.46***	**21**	**366**
LFB	**-5.67***	**22**	**451***
No Glitch	**-4.08***	6	256
Scattered Light	**-7.51***	**24**	**612***
	Neutron Star Black Hole Merger		
Helix	**-6.05***	**16**	**459***
45Mhz light modulation	**-6.72***	**23**	**549***
Low-frequency noise	**-4.1***	16	357
Paired doves	**-6.46***	**16**	**457***
Repeating Blips	**-5.23***	**13**	**439***
Scratchy	**-4.66***	**22**	**426***
Tomte	**-4.97***	31	364
Wandering Line	**-6.47***	**19**	**495***
Extremely Loud	**-3.59***	19	258
Air Compressor	-5.58	**17**	**511***

Table 3. Resulting statistics for the Theil–Sen estimator indicating the slope of the response time and the Pettitt Change-point estimate indicating the classifications at which habituation occurs. Note: A Bonferroni correction was applied, and the new alpha is 0.002.

here: (Table 3). For each volunteer, we computed the average accuracy prior to the habituation period. Next, we computed the average accuracy for the same number of classifications after the habituation period. For example, volunteers habituate to the blip glitch class at the 21st classification. To conduct our comparison of

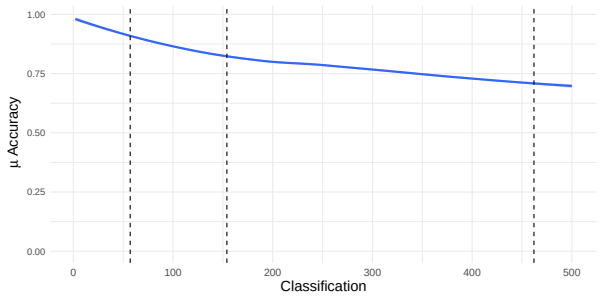


Fig. 5. Smoothed lines for mean accuracy. The figure includes vertical lines which indicate average classification when a volunteer is promoted.

accuracy, we computed the average accuracy for the 22nd through the 42nd classifications. Given that the accuracy has a non-normal distribution we conducted our comparison of the group means accuracy using a Wilcoxon signed-rank test. Additionally, since we conducted multiple tests, we increased the significance threshold with a Bonferroni correction which sets a new alpha to .002.

	Habituation (t)	Accuracy before (t)	Accuracy after (t)	$\vec{\Delta}$ t	Wilcoxon signed-rank test
Neutron Star Mountain					
Blip	**21**	**96 (100)**	**95 (98)**	**-2%**	**501809**
Whistle	**17**	**90 (96)**	**96 (95)**	**-1%**	**663612**
Galactic Supernova					
Koi Fish	**24**	85 (89)	86 (90)	+1%	59864
Power line	**17**	96 (100)	97 (99)	-1%	10779
Violin mode	**15**	**80 (88)**	**87 (91)**	**+3%**	**21552**
Binary Neutron Star Merger					
Chirp	**21**	88 (94)	93 (94)	+/-0	22
LFB	**22**	77 (81)	78 (82)	+1	126771
No Glitch	6	81 (91)	81 (88)	-3	250741
Scattered Light	**24**	**88 (96)**	**91 (96)**	-	**461683**
Neutron Star Black Hole Merger					
Helix	**16**	**76 (75)**	**84 (84)**	**+9**	**2547**
45Mhz light modulation	**23**	**63 (64)**	**68 (70)**	**+6**	**2457**
Low-frequency noise	16	**72 (78)**	**76 (79)**	**+1**	**39392**
Paired doves	**16**	45 (43)	53 (56)	+13	400
Repeating Blips	**13**	**65 (67)**	**70 (72)**	**+5**	**16972**
Scratchy	**22**	58 (59)	61 (62)	+3	16869
Tomte	31	77 (82)	79 (80)	-2	1189
Wandering Line	**19**	50 (46)	52 (53)	-2	357
Extremely Loud	19	80 (84)	79 (80)	-4	11263
Air Compressor	**17**	**55 (52)**	**66 (72)**	**+20**	**3066**

Table 4. Resulting statistics for the Wilcoxon signed-rank test where accuracy before and after the habituation period was compared. Note: A Bonferroni correction was applied and the new alpha is 0.002.

The results of our analysis are presented in Table 4 where the habituation period, accuracy on gold standard data prior and after the habitation period, the change in accuracy, and the Wilcoxon signed-rank test is reported. In workflow 1, we observed a decreased in accuracy for both glitch classes. A Wilcoxon signed-rank test indicated that there was a significant difference ($p < 0.002$) between the accuracy on gold data before habituation compared to the period after habituation. In workflow 2, accuracy in one glitch class was significantly different between periods - a 3% increase for Violin Mode ($Z = 21552$, $p < 0.001$). In workflow 3, we find no different in accuracy. Finally, glitch class options in workflow 4 we observed quite large increases in accuracy - Helix (+9%), 45Mhz light modulation (+6%), Low-frequency noise (+1%), Repeating Blips (+5%), and Air Compressor (+20%). Accuracy in workflow 4 are interesting since accuracy for many of the glitch class options had sub-optimal performance ranging from 52 - 78 percent.

6.4 RQ2-b: Habituation and retention

Next, we explored whether the effect of habituation on volunteers' performance. We pulled classification data for each glitch class prior to the habituation period. We then identified volunteers who leveled up and examined their performance prior to habituation. Reported in Table 5 is the number of volunteers that contributed to each workflow and the number of volunteers who did not level-up. In most cases, between 40-50 percent of volunteers are not promoted.

We compared pre-habituation performance between volunteers who were promoted and volunteers who dropped out prior to being promoted. We used the Wilcoxon rank sum test since the outcome variable accuracy violated several assumptions associated with t-tests and to ensure volunteers had seen enough gold standard classifications, we limited our population to volunteers who had seen at least 10 glitch classes. A Wilcoxon signed-rank test showed that their performance did reveal a statistically significant difference in performance ($Z = 36296000$, $p < .001$). The median accuracy rating was 100 for the dropout (N = 4138) and promoted (N = 5633) group. For workflow 2, we also observe that volunteers who were promoted had done better than those who dropped out ($Z = 12202000$, $p < .001$). Additionally, we find in workflow 3 that promoted volunteers also perform better prior to the habituation period ($Z = 6758000$, $p < .001$).

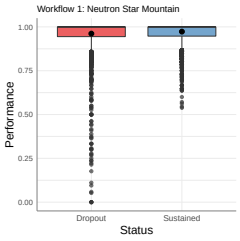
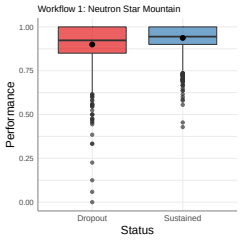
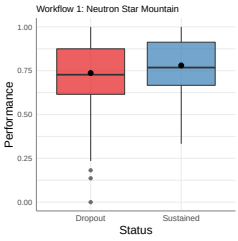
	Neutron Star Mountain	Galactic Supernova	Binary Neutron Star Merger
No. promoted	5,633	3,280	2,150
No. not promoted	4,138	1,271	1,196
			
Wilcoxon signed-rank test	Z = **36296000*****	Z = **12202000*****	Z = **6758000*****

Table 5. Performance on gold standard data prior to the habituation period for glitch classes introduced in that workflow. The number of contributors and dropouts exclude volunteers who did not meet the ten-classification threshold.

We conducted additional analysis to determine whether a volunteer's decision to remain an active contributor was related to their performance in specific glitch classes. We computed each volunteer's accuracy prior to habituation for each glitch class. Next, we determined whether their performance was above or below the average performance for that glitch class. We counted the number of volunteers and the number of occasions they performed below average in a workflow to determine their likelihood of dropping out or continuing to be a contributing member. The results of this analysis are displayed in Table6 and show for each workflow, the number of glitches where the volunteer performed below the average. The data are further grouped by the number of glitches where the volunteer performed below average and their contribution status. We used a test of proportions to determine if the number of volunteers dropping out at each glitch below average for each group (dropouts and retained) was different. The null hypothesis is that the two populations have the same proportion of volunteers performing below average, thus negating the rationale that competence in the classification task could be the reason volunteers drop out.

For workflow 1, the proportion of volunteers who performed below average and dropped out was almost identical to those who had also under-performed but continued in the project. The test of proportions revealed

that there was no difference in the likelihood of being retained given performance per-habitation in workflow 1. The estimated difference in proportions between-group dropouts, however, the difference was not significant ($\chi^2(1) = 3.69$, $p = 0.054$). Thus, accuracy in workflow 1 doesn't appear to be a significant distinguishing factor. In workflow 2 - below average performance for two and three glitch class options appeared to be good indicator of dropping out. The proportions between-group dropouts were significant for two ($\chi^2(1) = 7.88$, $p = 0.004$) and three ($\chi^2(1) = 8.77$, $p = 0.003$) glitch class options. In workflow 3, under-performance in one glitch class option was significant determining factor ($\chi^2(1) = 4.59$, $p = 0.03$) for being retained. Finally, in workflow 4, under-performance was significant for one ($\chi^2(1) = 87.74$, $p < 0.001$), two ($\chi^2(1) = 176.81$, $p < 0.001$), four ($\chi^2(1) = 17.46$, $p < 0.001$), five ($\chi^2(1) = 50.65$, $p < 0.001$) and six or more ($\chi^2(1) = 44.5$, $p < 0.001$) glitch class options.

	Neutron Star Mountain		Galactic Supernova		Binary Neutron Star Merger		Neutron Star Black Hole Merge	
	Drop-out	**Retained**	**Drop-out**	**Retained**	**Drop-out**	**Retained**	**Drop-out**	**Retained**
1	2,096 (19%)	2,049 (18%)	1,286 (24%)	1,935 (23.3%)	946 (20%)	856 (18%)	236 (7%)	10 (1%)
2	187 (2%)	123 (1%)	388 (7%)	514 (6.2%)	882 (18%)	876 (18%)	508 (14%)	487 (30%)
3	-	-	40 (0.7%)	31 (0.4%)	314 (6.4%)	291 (6%)	431 (12%)	225 (14%)
4	-	-	-	-	54 (1%)	36 (1%)	262 (7.3%)	69 (4.2%)
5	-	-	-	-	-	-	194 (5.4%)	19 (1%)
6+	-	-	-	-	-	-	202 (5.6%)	25 (1.5%)

Table 6. The number of volunteers who dropped out and performed below the average for the number of glitch class options introduced in that workflow.

7 DISCUSSION

In this section, we discuss our results in the context of theoretical implications for habituation and practitioners who manage virtual citizen science projects. To our knowledge, this is the first research that habituation as a theoretical lens to examine the contribution behaviors of people who participate in virtual citizen science. Our results also provide empirical evidence for the nuances of habituation in the context of a social computing system. In the sections below, we discuss our results in the context of habituation theory and implications for designing and managing citizen science projects.

7.1 The characteristics of habituation for pattern recognition tasks

Overall, we find evidence that accuracy changes and shares an inverse relationship to accuracy. We suspect this relationship forms because of the introduction of more challenging images being shown to volunteers - a feature of the training regime implemented in Gravity Spy. For the response parameter, our main outcome variable, habituation appeared to have little support when we examined the task without considering the type of pattern recognition. However, habituation is more nuanced in this context and our deeper examination shed more nuanced light on habituation. When tasks were disambiguated, visual inspection of the line charts in Figure 4 and the results of our statistical analysis reported in Table 3 revealed the rate and the classification when habituation occurs. Confirming, habituation rule 1 [18], in most instances, the rate of change was in the response parameter was a negative exponential function indicating that volunteers habituate relatively quickly. Based on the results of the Pettit change-point, for the pattern recognition task habitation occurs between the 15^{th} and 24^{th} exposure to a glitch class. These results suggest that system designers have a short window to engage participants.

We also find that while Gravity Spy exposes volunteers glitches that are more prevalent in the data stream during earlier workflows, glitch class options in the same workflow have varying rates of habituation. For instance,

in workflow 2, the average habituation to violin mode occurred during the 15th exposure, while habituation to the koi fish glitch occurred during the 24th exposure. Since volunteers' promotion to more advanced workflows depends on performance in the previous workflow, the habituation range suggest that some glitch class options may be misaligned in the current training regime and be too difficult to learn in a timely manner, delaying promotion. One recommendation would be to consider how best to support volunteers - aggregating glitches with similar rates of habituation. These findings related to the characteristics of habituation may help system designers and project organizers determine the most fortuitous placement of tasks in workflows to enhance the rate of learning for volunteers.

7.2 Habituation as a marker for learning and retention

With respect to RQ2a, we find evidence that habituation indicate learning. The literature on learning suggests that learners need to conduct tasks within their zone of proximal development. That is, it tasks appear too challenging, learners will become frustrated and leave. Alternatively, if tasks are too easy, learners will become bored and stop participating. The rationale emerges from Vygotsky's zone of proximal development [20], a theory describing what learners can do without help and what learners could do with minimal guidance. Based on a comparison of accuracy before and after habituation (see Table 4), we observed small increases in performance after the habituation period. It should be noted, however, that the median accuracy in workflow 1 is quite high (a 95-100 range). We can use knowledge about habituation to determine when people need additional guidance. Research by [10] suggests that for visual perception tasks that may be difficult to provide training for, system designers should make use of community-built resources. For instance, in Planet Hunters (a similar visual perception task), volunteers provide detailed comments pointing to the morphological features of transient planets [**?**]. We also find evidence that habituation is a useful marker for retention (RQ2a). Our results indicate that if volunteers are performing well prior to reaching habituation, they tend to be sustained in the project. These results suggest that projects may be able to predict retention based on the habituation behaviors of volunteers.

8 CONCLUSION

Our research sought goal in answering these questions is to understand how people learn and make suggestions for optimizing the introduction of pattern recognition tasks in citizen science. Understanding how habituation occurs can help project organizers optimize the design of workflows and task allocation algorithms. In applying habituation theory and considering the existing literature which seeks to limit habituation, we begin our discussion with the question, *should habituation be a goal for volunteers in a visual perception task?* Existing literature, which often suggests that habituation is a negative outcome, we argue that findings associated with can be used by platform designers to improve user experience and the performance of the system.

Limitations. As with any research, there are limitations. First, glitches are presented randomly to volunteers meaning that some images may be more challenging than others. This variability in the presentation of images means that some volunteers may be exposed to glitches that are more challenging early in their tenure. We expect these glitches to be distributed evenly among the groups we evaluated in our analysis. Second, our computation of response time approximates the time that volunteers spend examining a glitch. During volunteers' examination of the image, they could visit other web pages, read comments, or leave their screen for any number of reasons. Given the volume of records we examined, we do not expect that this will change our findings.

9 ACKNOWLEDGMENTS

Many thanks to the Zooniverse team for access to the data and our collaborators who helped us refine this research. We also thank the Gravity Spy volunteers who contribute their time and efforts to scientific research.

REFERENCES

[1] Bonnie Brinton Anderson, C Brock Kirwan, Jeffrey L Jenkins, David Eargle, Seth Howard, and Anthony Vance. 2015. How Polymorphic Warnings Reduce Habituation in the Brain *(the 33rd Annual ACM Conference)*. 2883 – 2892. https://doi.org/10.1145/2702123.2702322

[2] Daniel T Blumstein. 2016. Habituation and sensitization: new thoughts about old ideas. *Animal Behaviour* 120, C (10 2016), 255 – 262. https://doi.org/10.1016/j.anbehav.2016.05.012

[3] Rick Bonney, Caren B. Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V. Rosenberg, and Jennifer Shirk. 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience* 59, 11 (2009), 977–984. https://doi.org/10.1525/bio.2009.59.11.9

[4] Mark E Bouton. 2007. *Learning and Behavior: A Contemporary Synthesis*. Sunderland: Sinauer Associates.

[5] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W. Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. 2013. Your attention please: designing security-decision UIs to make genuine risks harder to ignore. *Proceedings of the Ninth Symposium on Usable Privacy and Security - SOUPS '13* (2013), 6. https://doi.org/10.1145/2501604.2501610

[6] V Curtis. 2015-11. Motivation to Participate in an Online Citizen Science Game: A Study of Foldit. *Science Communication* 37, 6 (2015-11), 723 – 746.

[7] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems* (2008).

[8] Lucy Fortson, Karen Masters, Robert Nichol, Kirk Borne, Edd Edmondson, Chris Lintott, Jordan Raddick, Kevin Schawinski, and John Wallin. 2011. Galaxy Zoo: Morphological Classification and Citizen Science. *arXiv* astro-ph.IM (04 2011). arXiv:1104.5513 arXiv.org

[9] Cécile Issard and Judit Gervain. 2017. Adult-like processing of time-compressed speech by newborns: A NIRS study. *Developmental Cognitive Neuroscience* 25 (2017), 176–184. https://doi.org/10.1016/j.dcn.2016.10.006

[10] Corey Jackson, Carsten Østerlund, Kevin Crowston, Mahboobeh Harandi, Sarah Allen, Sara Bahaadini, Scott Coughlin, Vicky Kalogera, Aggelos Katsaggelos, Shane Larson, Neda Rohani, Joshua Smith, Laura Trouille, and Michael Zevin. 2019. Teaching citizen scientists to categorize glitches using machine learning guided training. *Computers in Human Behavior* 105 (2019), 106198. https://doi.org/10.1016/j.chb.2019.106198

[11] Charlene Jennett, Laure Kloetzer, Daniel Schneider, Ioanna Iacovides, Anna L Cox, Margaret Gold, Brian Fuchs, Alexandra Eveleigh, Kathleen Mathieu, Zoya Ajani, and Yasmin Talsi. 2016. Motivations, learning and creativity in online citizen science. *WebSci 2013 Workshop: Creativity and Attention in the Age of the Web*. 15, 3 (Jan 2016).

[12] Ryan Lange. 2007. Video game habits: a reasoned action approach. *Proceedings of the 2007 conference on Future Play - Future Play '07* (2007), 213–216. https://doi.org/10.1145/1328202.1328242

[13] Catharine H. Rankin. 2009. Introduction to special issue of neurobiology of learning and memory on habituation. *Neurobiology of learning and memory* 92, 2 (2009). https://doi.org/10.1016/j.nlm.2008.09.010

[14] Catharine H. Rankin. 2009. Introduction to special issue of neurobiology of learning and memory on habituation. *Neurobiology of Learning and Memory* 92, 2 (2009), 125–126. https://doi.org/10.1016/j.nlm.2008.09.010

[15] Catharine H Rankin, Thomas Abrams, Robert J Barry, Seema Bhatnagar, David F Clayton, John Colombo, Gianluca Coppola, Mark A Geyer, David L Glanzman, Stephen Marsland, Frances K McSweeney, Donald A Wilson, Chun-Fang Wu, and Richard F Thompson. 2009. Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory* 92, 2 (09 2009), 135 – 138. https://doi.org/10.1016/j.nlm.2008.09.012

[16] Holly Rosser and Andrea Wiggins. 2018. Tutorial Designs and Task Types in Zooniverse *(Companion of the 2018 ACM Conference)*. 177–180. https://doi.org/10.1145/3272973.3274049

[17] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142, 10 (10 2009), 2282 – 2292. https://doi.org/10.1016/j.biocon.2009.05.006

[18] Richard F Thompson and William Aden Spencer. 1966. Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychological Review* 73, 1 (1966), 16 – 43.

[19] Anthony Vance, Brock Kirwan, Daniel Bjornn, Jeffrey Jenkins, and Bonnie Brinton Anderson. 2017. What Do We Really Know about How Habituation to Warnings Occurs Over Time? *(the 2017 CHI Conference)*. 2215 – 2227. https://doi.org/10.1145/3025453.3025896

[20] L S Vygotsky. 1980. *Mind in Society*. Harvard University Press.

[21] Andrea Wiggins and Kevin Crowston. 2012. Goals and Tasks: Two Typologies of Citizen Science Projects *(th Hawaii International Conference on System Sciences)*. 3426–3435. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9A6865CD5DF7F6C92AFC406CA35FFFEF?doi=10.1.1.224.1691&rep=rep1&type=pdf

[22] Donghee Yvette Wohn. 2012. The Role of Habit Strength in Social Network Game Play. *Communication Research Reports* 29, 1 (2012), 74–79. https://doi.org/10.1080/08824096.2011.639912

[23] Wendy Wood, Jeffrey M. Quinn, and Deborah A. Kashy. 2002. Habits in Everyday Life: Thought, Emotion, and Action. *Journal of Personality and Social Psychology* 83, 6 (2002), 1281–1297. https://doi.org/10.1037/0022-3514.83.6.1281

[24] Shuhei Yamaguchi, Laura A Hale, Mark D'Esposito, and Robert T Knight. 2004. Rapid Prefrontal-Hippocampal Habituation to Novel Events. *Journal of Neuroscience* 24, 23 (2004), 5356–5363. https://doi.org/10.1523/jneurosci.4587-03.2004