NeuroCrypt: A Multimodal Coercion-Resistant Implicit Memory Authentication System

Ritul Satish*, Niranjan Rajesh*, Sristi Bafna*, Arup Mondal, Khushi Mohta, Argha Chakraborthy, Aditi Jain, and Debayan Gupta+

Computer Science Department, Ashoka University, India

Abstract

Authentication systems impose high cognitive loads on users (remembering multiple complex passcodes) and are susceptible to coercion attacks. In 2012, Bojinov et al. attempted to suppress explicit retention of passcodes by leveraging implicit memory to store passcodes through a Serial Interception Sequence Learning task. We aim to replicate and improve this by adding specific stimuli to strengthen sequence learning and retention (based on cognitive psychology literature). NeuroCrypt *is a coercion-resistant authentication system based on statistical patterns in human interaction with carefully generated visual and auditory stimuli. Through experiments across 70 participants, we observe that visual stimuli greatly aid implicit sequence learning and maintain attention; inclusion of auditory stimuli help participants in long-term retention of their passcodes. We also determine the practicality of our system through a systematic analysis of authentication threshold values. Overall, we find that integration of stimuli is a promising approach to a more effective implicit-learning authentication system.*

1 Introduction

Today's prevalent user authentication systems are knowledge-based – users are required to recollect a pre-set password from memory to validate their identity. Passwords, however, are not without their faults. One major flaw is the high cognitive load they impose on users [1]. Today, the average user is expected to memorize passwords for around a hundred online accounts [2], on top of increasingly complex character constraints [3]. This cognitive burden has led to most users reusing or creating similar passwords using a combination of predictable words, numbers, and symbols [4]. Moreover, this burden is usually so large that users have tended to circumvent them by simply writing down the passwords somewhere – meaning that a physical search is often sufficient to crack digital protections. Beyond the cognitive burden and the risk it carries, passwords are also prone to a wide range of attacks. One such example is a coercion or "rubber hose" attack. This class of attacks involves an adversary bypassing the encryption of a system by forcing the user to reveal the authentication secret through torture or extortion. The vulnerability of traditional passwords to such attacks, combined with their growing cognitive burden in general, compels us to rethink how and where such secrets are stored. In place of the traditional option – the explicit memory of a user – we turn to their *implicit memory* for a solution.

Tasks that utilize the implicit memory of a subject do not require them to consciously recollect previous experiences of task performance [5]. We use our implicit memory for various everyday tasks like brushing our teeth, riding a bicycle and tying our shoelaces. These are often motor actions that are carried out without explicit recollection or mental effort and are attributed to the cerebellum and basal ganglia in the brain [6]. In an authentication system incorporating implicit memory, we are able to plant a secret key into the user's mind without them having conscious access to that key. This not only reduces the cognitive burden on the user but promises to eliminate the threat of coercion attacks. Prior research that involves leveraging implicit memory to remember passcodes has found varying degrees of success [7, 8, 9, 10]. Such systems typically require users to be trained in tasks that are then tested at authentication. An untrained user, on average, performs worse than trained users which allows for distinguishing valid and invalid users. Despite their ingenuity, these systems typically suffer from low levels of implicit secret learning which may also decay over short durations. These weaknesses must be overcome if implicit memory-based authentication can be considered a practical solution in the real world.

Consequently, we present NeuroCrypt, an extension to the Serial Interception Sequence Learning (SISL) Task [11, 10]. We model our system after the popular arcade game "Guitar Hero", incorporating carefully selected sensory stimuli from cognitive psychology and human-computer interaction literature. The goal is to enhance implicit passcode learning and retention. First, we replicate the SISL system by Bojinov et al. [10] and then design variants of the system with visual and auditory stimuli. We then analyze the effects of the stimuli by investigating NeuroCrypt's efficacy, security and practicality as an authentication system.

Our work primarily makes four contributions:

1. We present an implicit memory-based authentication system design that integrates visual and auditory stimuli for enhanced passcode learning and retention.

2. We conduct a user study to investigate the extent of implicit passcode learning and retention that takes place in our proposed system. We make the data collected in our study and the code for analysis freely available (Section A). We find that visual stimuli induces greater extents of implicit learning while auditory stimuli aid implicit retention .

* Equal Contribution; + Correspondence: debayan.gupta@ashoka.edu.in

3. We systematically assess the security and practicality of our system based on statistical patterns in user data through authentication threshold values (σ_{ATV}). To the best of our knowledge, we are the first to employ statistical authentication thresholding to non-biometric systems.
4. We find that previously reported results [10] under identical conditions for implicit sequence learning is unachievable.

Our work highlights that drawing on informed insights from the fields of cognitive psychology and human-computer interaction is a promising approach to the development of practical and secure real-world authentication systems.

2 Background & Related Work

Our work is inspired from and built on prior work from the fields of computer security, cognitive psychology and human-computer interaction. In this section, we present the relevant background for our proposed authentication scheme.

2.1 Rubber Hose Attacks

Rubber hose cryptanalysis [12, 13, 10] is a euphemism for the forceful extraction of cryptographic secrets such as passwords or encrypted documents from a user. This could be achieved through coercion or torture of both physical (like the titular rubber-hose based torture method mentioned by Ranum in [12]) and psychological natures. These attempts at extracting authentication secrets bypass any encryption or security mechanisms designed in authentication systems, no matter its complexity and efficacy. This glaring flaw of such systems invite an exploration of authentication secrets that do not reside in retrievable human memory. Thus, this class of attacks serves as a motivation for our proposed authentication system.

2.2 Cognitive Load Theory and Multimodal Processing

The cognitive load theory in cognitive psychology builds upon the information processing model which consists of three memory modules: sensory memory, working memory and learning memory [14]. Sensory memory filters through sensory input which is then passed on to the working memory where it is processed or removed. When performing a task repeatedly, the process of 'learning' occurs when information is transferred from the working memory to the long term memory [15]. The working memory is constrained by the amount of information it can process at a given time [16]. If the working memory's cognitive load exceeds this constraining threshold, learning is impeded. Today's authentication systems present a significant load on the working and long-term memory of users. To facilitate learning while minimizing cognitive load, we turn to the working memory's ability to separately process information from different modalities.

In cognitive psychology and human-computer interaction, a modality is defined as a single independent channel of sensory input/output of a human. Examples of sensory modalities are vision (visual information), audition (auditory information) and tactition (haptic information). When a system involves multiple modalities, it is said to be multimodal. A multimodal approach to learning can be significantly effective as the working memory processes information from different modalities separately, ensuring reduced cognitive load and unimpeded learning [16]. The distribution of sensory information is also carried on to implicit learning as reflected in studies that used a multimodal approach to a Serial Reaction Time task [17]. Thus, we employ the visual and auditory modalities to improve implicit learning in our authentication system.

2.3 Implicit Memory

In cognitive psychology, implicit learning and implicit memory refer to the unconscious effect that prior information processing may exert on subsequent behavior. Implicit memory is one of the two main types of long term memory in humans along with explicit (declarative) memory. It is hypothesized that the implicit memory system is primarily related to the basal ganglia along with the cerebellum [6]. Some examples of activities that employ implicit memory include typing on a computer keyboard, brushing teeth and riding a bicycle. It is considered to be one of the most important and complex cognitive processes for the acquisition of most motor, perceptual and cognitive skills [18]. Implicit memory shows higher resistance to memory deficiencies and is more robust compared to explicit memory [19]. Several tasks have been designed to showcase implicit learning in humans like the Serial Reaction Time task (SRT) [17], Artificial Grammar Learning task (AGL) [20] and Serial Interception Sequence Learning task (SISL) [11, 21]. The evidence of learning in these tasks is proof of the implicit memory system. With intelligently designed systems, implicit learning can be leveraged for required information to be embedded in the human brain.

2.4 Implicit Memory-based Authentication Schemes

Denning et al. [8] proposed an implicit memory and priming effect-based user authentication. Participants were initially presented with original images and then the degraded counterparts for authentication through a familiarisation task. A minor priming effect was present in many images. However effective, the reliability and viability of the scheme depends highly on the identification and creation of images with a sufficiently strong priming effect. Due to this crucial dependency, it would be improbable for the system to perform efficiently in case of a large number of users.

Castelluccia et al. [22] developed an authentication system through a Mooney images-based implicit learning approach. Mooney images are low information, two tone representations of the original images. During the priming session, the participants are presented with a Mooney image, the original image, and a label to describe the object in the images. For the authentication, the participants are presented with the primed and unprimed Mooney images in a pseudo-random order and are instructed to label the same. The shift from decoy distorted

images to Mooney images for authentication displayed better retention in participants, as it triggered brain processes for implicit memory and recollection.

Bojinov et al. [10] developed an implicit learning-based authentication system which attempts to solve the problem of coercion attacks. The proposed scheme required participants to intercept cues falling from the top of the screen at varying speeds in six different columns before they reached the sink at the bottom. Unbeknownst to the participants, their sequence of falling cues involved repetitions of their uniquely assigned 30-character passcode during the training session. During authentication, their trained passcode is presented alongside two other untrained 30-character sequences. Successful authentication required the user to perform better on their trained passcode than on the untrained sequences. The authors were able to establish that implicit sequence learning could be used as an authentication mechanism. However, they also discovered that the performance advantage participants displayed on their trained passcodes were forgotten over time which necessitates expensive re-training sessions to ensure consistent, long-term success of the authentication systems. Additionally, the authors fail to provide insights into optimal advantage values that should be utilized in their authentication system to maximize true user entries and minimize chance entries to the system. In our work, we aim to address these shortcomings.

3 Proposed Authentication System

3.1 Overview

The authentication system proposed in this paper is an extension of the Serial Interception Sequence Learning (SISL) task from [10]. Thus, most of our design choices for our baseline system is identical to [10]. The SISL task is designed similarly to the popular arcade game Guitar Hero. The subject is expected to intercept falling circular cues by pressing keys on the keyboard corresponding to the columns through which they fall. A hit is registered when the user presses the correct key as the corresponding cue reaches the end line at the bottom of the screen. A miss is registered if the key is not pressed on time, if an incorrect key is pressed, or if more than one key is pressed at the same time. We measure performance by the hitrate, which is the ratio of successful hits to total cues seen by the user. The speed of the falling cues is regulated by a difficulty-modulating algorithm to maintain a hitrate of 70% in order to control for skill disparity across players. In addition to the baseline control game that is identical to the one presented in [10], we also designed two additional games that include visual and auditory stimuli as outlined in 3.4. Screenshot of the the base game is presented in Fig 1.

In our system, like in [10], the authentication secret takes the form of a passcode which is a 30-item sequence of keys that correspond to the cue columns in the game. Each user is assigned a unique passcode that they learn during the training phase of the games. During authentication, the user is presented with their unique passcode and two untrained sequences and respective user performances are measured. Authentication is successfully achieved if the user's passcode performance is sufficiently above their performance on the untrained sequences. The training and authentication phases are outlined in greater detail in 3.3.

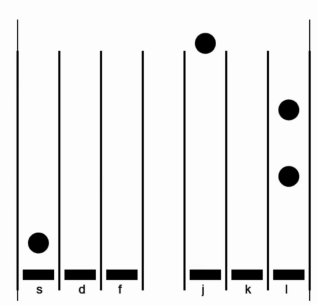


Figure 1: **Screenshot of the Control game**. The visual game involved flashing cues and thicker column separators. The audio game looks identical to the above but includes additional auditory stimuli.

3.2 Passcode Sequence Generation

The passcode consists of six keys corresponding to the six columns in the game as in Figure 1, represented by the set $S = \{s, d, f, j, k, l\}$. These sequences are designed to not stand out to the user to suppress explicit recollection despite repeated exposure during training. The passcode sequences consist of unique non-repeating bi-grams (pairs of characters that do not repeat consecutively) to increase the difficulty of recognising patterns. This algorithm results in a passcode with 38 bits of entropy which is substantially greater than typical passwords. We define Σ as the set of all passcodes, its cardinality being 2^{38} [10]. A detailed description of the passcode generation process is provided in Algorithm 1 in the appendix.

3.3 The Phases of the Authentication System

The NeuroCrypt authentication system is composed of two phases – the training phase and the authentication phase. Both phases have different game sequences that were shown to participants as explained below and visualized in Figure 2.

3.3.1 Training Phase

The training phase of our system is analogous to the process of setting your own password. During the phase, users learn a 30-character passcode that is uniquely generated (as mentioned in 3.2) for them. In our training design, we maximize user exposure to their passcode while minimizing the extent of their explicit 'recoverability'. This sequence is randomly embedded three times within an 18-character random 'noise' sequence that has no consecutive repeats in order to maintain homogeneity with the passcode's pattern. This creates a 108-character sequence referred to as a training 'sub-block'. Each sub-block is repeated five times to form a training 'block' which contains 540 characters. A full training session consists of 7 training blocks (3780 characters) separated by 20-second

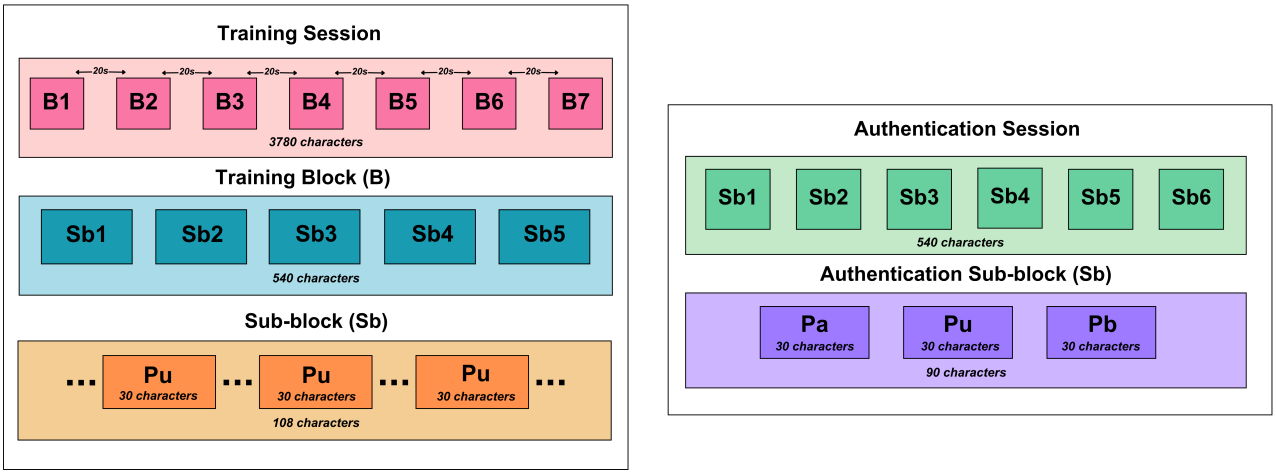


Figure 2: **Visualization of the phases of NeuroCrypt** The game sequences for the training and authentication sessions are displayed above. The training session is made up of 7 training blocks with 20 second intervals in between. A training block consists of 5 training sub-blocks which include the user's passcode repeated 3 times in between 18 characters of random noise. In the authentication session, the main sequence consists of 6 authentication sub-blocks where each sub-block presents the user's passcode P_u and two untrained passcodes P_a and P_b.

pauses. The user encounters their 30-character passcode 105 ($3\times5\times7$) times embedded irregularly between 630 ($18\times5\times7$) noise characters. An entire training session typically takes about 30-45 minutes to complete.

3.3.2 *Authentication Phase*

In the authentication phase, users are tested on their recollection of the passcodes they were trained on during the training phase. To achieve this, we present their passcode along with two randomly generated, untrained sequences. The same three sequences are used for all authentication sessions of the user. We present the authentication game sequence as 6 authentication sub-blocks. In each sub-block, we show a random permutation of the three sequences. The user thus encounters their passcode 6 times and untrained sequences 12 times during the entire authentication session. This is a substantially smaller proportion of the game sequence as compared to the training phase as the primary goal of the authentication phase is not to induce passcode learning but to verify it. Authentication is considered successful if the user's hitrate on the trained passcode sequence is higher than the average hitrate on the two untrained sequences and an authentication threshold value. Analysis and explanation of this value can be found in 6.1. The entire authentication session took around 5-10 minutes to complete.

3.4 *Modalities Employed*

NeuroCrypt's primary novelty over [10] is in the addition of visual and auditory stimuli that we hypothesize to improve implicit passcode learning and retention. We design two additional games that test the effectiveness of stimuli from each modality.

3.4.1 *Visual Game:*

The visual modality has been favored across literature for implicit learning based tasks. Popular techniques employed for the same include contextual cueing [23, 24], spatial arrangement [25] and color contrast [26] for visual distinction to improve visual perception. The following were features added to the visual game to investigate the effects of visual stimuli on implicit learning:

- The color contrast difference between the cues and the background is maximized by making the cue black (#000000) and the background of the game screen white (#FFFFFF).
- Solid visual separators between the columns that distinctly map each hand to a set of columns and provide spatial clarity for faster cue recognition.
- Flashing of the visual cues that improve perceptibility of cues while maintaining visual attention [27].

3.4.2 *Audio Game:*

The Auditory modality has also received much attention in work related to sequence learning due to the Auditory Scaffolding Hypothesis [28]. This hypothesis suggests that sound plays a significant role in perceiving and interpreting sequential information. The following features were added to the audio game to explore the benefits of auditory stimuli on implicit learning:

- A musical note is played on a key press by the participant to help provide the 'scaffolding' for higher learning of the sequence. To suppress the effect of

the musical notes on explicit learning, the notes followed an unorthodox arrangement. 3 unique notes were mapped to 2 keys each, one key on each side ({s,d,f} on the left and {j,k,l} on the right). The musical notes are selected with the aim of minimizing perceptual grouping based on timbre and pitch [29].

- A mild white noise is played in the background while the participants play the game. It has been proven that background white noise improves performance in inattentive participants and reduces performance in attentive participants for memory tasks [30]. For this reason, the white noise volume is modulated with an inverse relation to participant performance over the course of a game.

4 Experiment

The goal of our experiment is to test the efficacy of NeuroCrypt. To achieve this, we conducted user studies for each modality. We split our participants into three groups: 'Control' (n=22), 'Visual' (n=24) and 'Audio' (n=24) and collected performance and qualitative data over the course of the experiment.

4.1 Design

The participants first took part in the training session (as in 3.3.1) where they were repeatedly exposed to their own passcode sequence separated by random character. Participants in each modality group encountered corresponding stimuli as mentioned in 3.4 whereas the control group encountered the exact same baseline game from [10]. The participants typically took 30-45 minutes to complete training.

After the training session, participants were asked to fill out a post-game survey which collected qualitative data. This included demographic details like age, linguistic background, musical background and gaming background. We also collected 'familiarity ratings' for 5 different videos (randomly ordered) of 30-item long game sequences on a scale of 1 to 5 to test whether participants were able to recognize their assigned passcodes. One of the five videos was the participant's assigned passcode that they learned during the training session while the other four were randomly generated, unseen sequences. We also collected qualitative feedback for the game where participants were free to write about their experience during the game. This data was analyzed in 6.2.

Finally, we conducted two follow-up sessions: Auth One and Auth Two (as outlined in 3.3.2), where participants were called back one and two weeks after their training session respectively. In these authentication sessions, we studied the implicit retention of the trained sequences in the trained participants in relation to untrained sequences. We report the results for the experiment's training and authentication components in 5.

4.2 Participants

We recruited 88 participants from our host university for our experiment. Participants who did not attend or complete all three game sessions and the post-game survey were removed resulting in a total of N = 70 participants for the study. Our surveyed population consisted of university students between the ages of 18–25.

To prevent bias and preparation of any form from the participants, we did not reveal the aim or the details of the experiment during recruitment. The participants were informed that there would be a total of three sessions, each a week apart. They were to perform the experiment in person and had to report to the venue on the same day and time for three weeks. The experiment was conducted in the same lecture hall at the university. To provide incentive, participants were offered monetary compensation. Research staff briefed the participants on how to play the game at the beginning of each session and were present throughout to resolve any doubts or logistical issues. As an added performance incentive, participants were encouraged to beat a 'high score' to gain a bonus monetary reward.

4.3 Method

The NeuroCrypt control, visual and audio games were developed as web applications from scratch using the library PixiJS. Participant performance data was collected using Google Firebase that was integrated into our application. All participants were asked to bring their own personal laptops (and earphones in the case of the audio game) to a university lecture hall and were sent web links to play the game.

4.4 Ethics

Our study was reviewed and approved by our Institutional Review Board (IRB). Participants had to consent to take part in and could drop out of the study at any time. They were given monetary compensation at the end of the first session as well as the third to convey our appreciation for their effort and time.

5 Results

In this section, we report and comment on our findings from our experiment outlined in the previous section. We first discuss the passcode learning that took place in the training session and then the corresponding retention observed in the authentication sessions across all participants in the study.

5.1 Learning During Training

We first investigate whether passcode learning takes place during the training session of each experiment. We observe that there is a general, non-monotonic increase in the performance of participants across the training session for all three modality groups as shown in Figure 3a. Over the course of the 40-minute training session, participant performance in the

games increases as expected. The magnitude of increase is higher in the audio group than the visual and control groups. Additionally, Figure 3a also shows an unexpected decrease in the average performance of participants around the halfway point of the training session in the control and audio groups. The decrease in performance could be attributed to faltering attention over the course of the lengthy training session. On the other hand, the absence of a performance drop in the visual condition suggests that the flashing cues and visual separators may be contributing factors to holding participant attention.

Figure 3b displays the levels of passcode sequence learning over the course of the training sessions. The learning is characterized by the performance advantage metric which we define as the performance difference between their assigned passcode sequence and noise sequences. Recall that a single training block consists of 540 cues where 450 cues belong to the 6 occurrences of their passcode and 90 cues are random noise. We notice a varying level of increases in performance advantage across modality conditions. The most prominent increase in passcode performance advantage during training was found in the visual group, suggesting that visual training yielded the greatest extent of sequence learning.

All in all, each of the conditions showed a significant increase in passcode performance from the beginning of the training session till the end as shown in Fig 4b (t-tests revealed significant performance increases $p < 0.005$ relationships for each group).

5.2 *Authentication performance and retention over time*

Following the training session, we called the same participants back twice at one week intervals to systematically test their passcode performance advantages on their trained passcode sequence and two untrained passcode sequences in an authentication block. As seen in Figure 4a, we observe that the performance on the trained passcode sequence is higher than on the untrained sequences across all conditions in the authentication sessions. The statistically significant differences between trained and untrained sequences (control: $t = 1.71$, $p = 0.09$; visual: $t = 2.34$, $p = 0.02$; audio: $t = 2.55$, $p = 0.01$) imply that validating user identity based on performance advantage on trained sequence is possible. The statistical tests also tell us that the performance difference is more prominent in the audio and visual games than the control, and hints that authentication may be more robust in the visual and auditory variants of our system.

We also examined the retention of trained sequence advantage over the two week interval across different conditions. Previous work [10] has shown degradation in performance advantage when measured 1 and 2 weeks after training. Figure 4b shows that our experiment's authentication sessions display a degraded performance for the control and visual groups after the time jump. However, the audio group displays a slight but steady increase in passcode sequence performance which hints at the effect of audio stimuli in long-term implicit sequence retention. This trend between the modality groups is consistent in Figure 4c with the way performance advantage changes across sessions. In the audio condition, the mean passcode performance advantage across participants rises steadily over time whereas the visual and control groups see a decrease. It is also worth noting that the visual group exhibits a steeper decline in performance advantage than the control group. Interestingly, none of the performance advantages between the training and authentication sessions were found to be statistically significant (most significant was the increase in audio performance advantage with $t = 1.66$ and $p = 0.10$). This set of results suggests that the chosen auditory stimuli are involved in robust implicit sequence learning over longer periods of time in relation to visual stimuli and the baseline conditions.

In order to compare learned performance advantages between modality groups, we performed t-tests at confidence thresholds of $p = 0.05$ in each session but did not find any statistically significant relationships. Further, variability in users' performance advantages can be observed in Figure 5 in the appendix.

6 Security Analysis

We conducted some basic follow-up analysis to validate efficacy of NeuroCrypt as a secure authentication system. Unlike in traditional security systems, we are unable to rigorously and conclusively analyse user and adversarial success rates due to the probabilistic nature of our system. Thus, our security analysis centers upon the behavioural patterns of the participants (as potential users) of our system. In this section, we systematically analyze and find optimal Authentication Threshold Values, discuss possible explicit recollection of the passcodes that could be leveraged by adversaries and finally evaluate our system within the basic threat model framework.

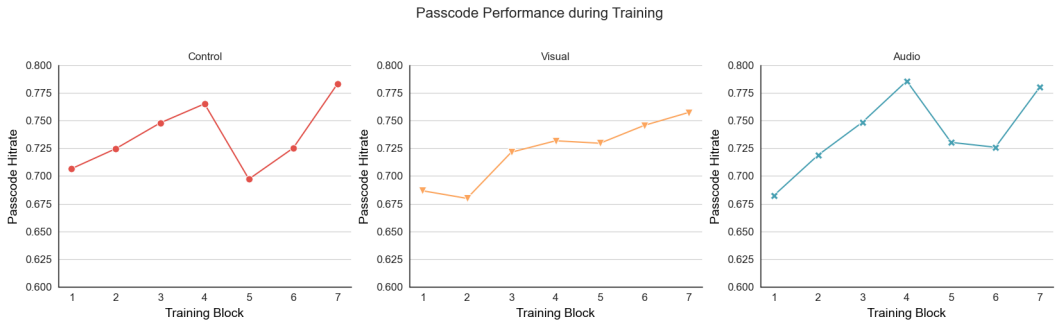
6.1 *Authentication Threshold Values (σ_{ATV})*

Recall that in NeuroCrypt's authentication phase, a user is successfully authenticated when their performance on their unique passcode is sufficiently greater than the average of other sequences of the same length. In this subsection, we aim to systematically investigate this 'sufficiency' threshold.

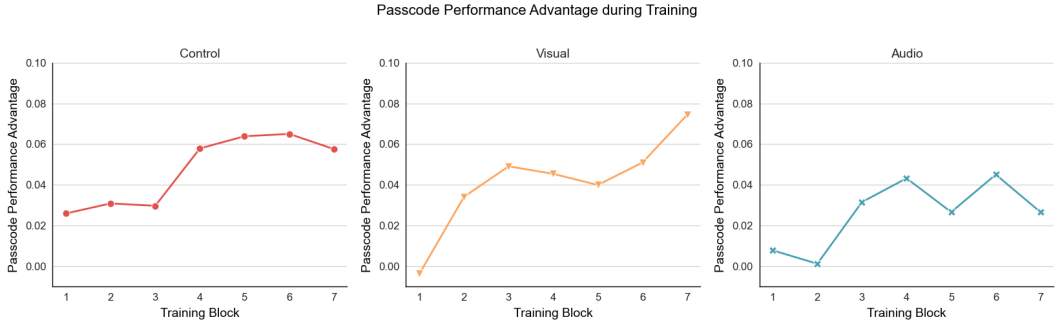
Thresholds play a significant role in current state-of-the-art biometric authentication systems. They determine the level of confidence required for a match between the user's performance during authentication and previous authentication success rates. Literature on the same has primarily looked at using such statistical patterns in user behaviors for facial recognition-based systems [31]. To our knowledge, our work is the first to employ authentication thresholding on the basis of collected user behavior data for a non-biometric authentication system.

During an authentication block, the user encounters their own passcode (P_u) 6 times alongside 12 instances of two other, untrained sequences (P_a, P_b). Formally, authentication is successful if the following inequality is satisfied (where $H(x)$ is the hitrate of input sequence x):

$$H(P_u) > Avg(H(P_a), H(P_b)) + \sigma_{ATV}$$



(a)



(b)

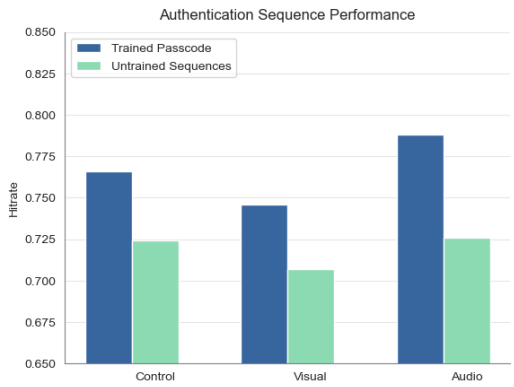
Figure 3: **Performance during Training** (a) The increase in average passcode performance over the course of the training session across all participants in each modality group. Performance is measured as the hitrate – the ratio of successful passcode cue hits over all shown passcode cues. (b) The average block-wise performance advantage of participants respectively across all participants for each modality group. The performance advantage is defined as the difference between passcode and non-passcode performance.

We define σ_{ATV} as the Authentication Threshold Value that necessitates the user to perform a certain level greater on their own passcode compared to untrained sequences. We set out to find a σ_{ATV} such that true user authentication success rate is maximized while false authentication (on sequences other than the assigned user passcode) is minimized. In other words, an ideal threshold would be low enough for all true users to be authenticated into the system but also high enough to eliminate any adversaries or untrained users from being let in.

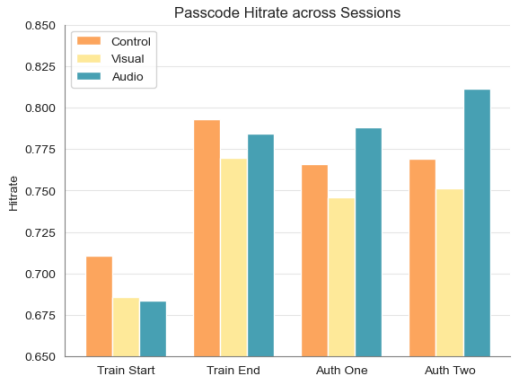
To achieve this, we calculated the proportion of participants that would be successfully authenticated at a range of σ_{ATV}'s as reported in Table 1 (columns 2-6). These columns show the true user authentication success rates for each modality for one and two weeks after training. Our system aims to maximize this measure and we find that the visual and audio games are able to successfully authenticate a greater proportion of users at low σ_{ATV} values than the control game in Auth One. However, only the audio game is able to maintain high success rates in Auth Two – reinforcing the retention effect of the auditory stimuli.

As our system authenticates on the basis of statistical patterns in user behavior, false authentications are likely. To investigate this, we also calculated the proportion of participants that are able to successfully authenticate on sequences that they were not trained on as reported in Table 1 (columns 7-9) across the same σ_{ATV} values. This is done by treating the untrained sequence in Auth One as their 'psuedo-passcode' and the untrained sequence in Auth Two as the noise sequences. Despite being untrained on their pseudo-passcodes, a proportion of participants were incorrectly authenticated into the system, especially at lower threshold levels. The control and audio game consistently suppress false authentication at a greater rate than the visual game.

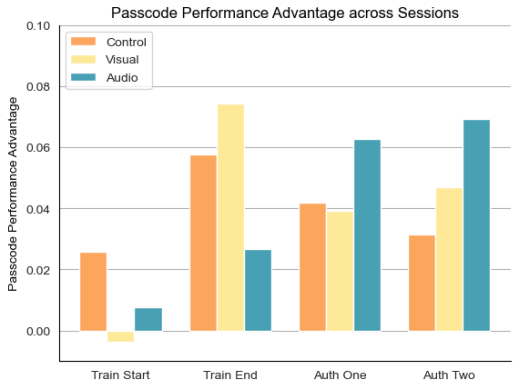
As mentioned earlier, striking a balance between ensuring maximal true user authentication and minimal false authentication in NeuroCrypt involves maximizing the difference between the two. Given the collected participant data, we find this difference to be greatest at σ_{ATV}=0.025 for the control and visual games and σ_{ATV}=0.05 for the audio game (from Table 2 and Figure 6 in Appendix). We concede that although maximizing the difference between true and false authentication success rates offer a technically sound manner of tackling the problem of security efficacy, it is not in any means the most practical option since at σ_{ATV}=0.05, the audio game only successfully authenticates less than half its users (in Auth One). Alternate approaches may look at fixing low σ_{ATV} values to ensure large-scale successful authentication while



(a)



(b)



(c)

Figure 4: **Performance and retention over time** (a) displays the average passcode and untrained sequence performance in both authentication sessions across all participants in each modality group. (b) and (c) display the average passcode performance and advantage over noise or untrained sequences in each modality group across participants in different stages of the experiment. Train Start and Train End refer to the first and last block in the training session whereas Auth One and Auth Two refer to the two authentication sessions that took place after the training session.

protecting against false authentication attempts by disallowing or punishing multiple sign-in attempts.

6.2 *Explicit Memory of Passcodes*

Recall that explicit memory is the phenomenon that we are trying to suppress in our authentication system to avoid rubber hose attacks and minimize cognitive burden on the user.

In [10], explicit recognition of the passcode was assessed by presenting participants with five different animations of the game sequence (of which, one was their assigned passcode) after the training and testing sessions and asking them how familiar each looked on a scale of 0 to 10. Participants displayed moderately higher explicit familiarity for their passcode sequence on average but there was low positive correlation between familiarity rating for the passcode and user performance on the SISL task. Participants were also tested on substrings of their passcode and displayed a statistically insignificant difference between their performance on the trained and untrained substrings.

We replicated the first task of testing explicit familiarity and chose not to perform the computationally expensive substring task. Participants were asked to rate five sequences (of which, one was their passcode sequence) as familiar on a scale of 1-5, after the training session. Our results on this task were similar to [10]. The proportion of participants who rated their passcode sequence more familiar on average was 41.7%, 50.0% and 45.8% for the audio, visual and control games ($p = 0.37, 0.07$ and 0.07) respectively as seen in Figure 7 in the appendix . We computed the Mann-Whitney U test statistic and observed a statistically insignificant difference between explicit familiarity ratings for trained and untrained sequences for all three cohorts. Additionally, there was low positive correlation between user performance advantage on our SISL task during training and their familiarity rating of the passcode sequence (Pearson's $r = 0.17$).

Although we chose to replicate the familiarity rating task from [10], we do not believe it to be a true measure of explicit recollection of the sequence. Participants showing a familiarity preference on their trained passcode does not necessarily imply the ability to explicitly recall the same sequence. Instead, it might simply be a result of the repeated exposure during the training session and possibly a correlate of any implicit sequence learning that has taken place. Explicit recollection is a generative task. Ideal methods to assess the explicit recollection would involve free recall tasks: requiring users to reproduce the entire 30-character passcode, and partial sequence reconstruction tasks: requiring reconstruction of substrings from their trained passcode. Accuracy metrics for these tests would be much more reliable measures of explicit recollection than the familiarity rating task or the substring recognition task. It is also worth noting that in the context of testing for explicit recollection for the purpose of coercion attack protection, these tests may not be reliable as users may behave extremely differently in a duress setting. Nevertheless, further discussion on this is beyond the scope of this paper, so

σ_{ATV}	Auth1 Passcode Success (%)			Auth2 Passcode Success (%)			Untrained Seq Success (%)		
	Control	Visual	Audio	Control	Visual	Audio	Control	Visual	Audio
0.000	65.22	70.83	82.61	69.57	66.67	91.30	34.78	50.00	34.78
0.025	52.17	62.50	60.87	60.87	66.67	69.57	13.04	37.50	17.39
0.050	30.43	45.83	47.83	34.78	45.83	65.22	13.04	29.17	4.35
0.075	30.43	33.33	26.09	13.04	37.50	30.43	4.35	12.50	4.35
0.100	21.74	25.00	21.74	13.04	25.00	26.09	4.35	12.50	0.00

Table 1: Authentication success rates of participants for varying σ_{ATV} values for their unique passcodes from the Auth sessions and success rates of participants on random untrained sequences.

we leave such explicit passcode reconstruction tasks for future work.

6.3 Basic Threat Model

While a standard cryptographic threat model is inappropriate in this case, it may nevertheless be useful to discuss some specific scenarios. As part of our security analysis, we highlight three cases that could compromise our authentication system and evaluate the success rate of an adversary. Following convention, each scenario involves three entities – Alice, the user; Bob, the authentication server; and Oscar, the adversary.

Case 1: "Impersonation" – Oscar attempts to authenticate as Alice

In this case, we assume that Oscar does not have access to Alice or any prior knowledge of how the passcode is constructed. In order to be authenticated successfully, Oscar would have to perform sufficiently better at Alice's passcode than the other sequences presented to him. Without knowledge of the authentication block and passcode structure, the probability of him being authenticated is equal to the untrained sequence probabilities from Table 1 and are directly related to the system's preset σ_{ATV}. Oscar cannot leverage any other information without access to NeuroCrypt's sequence design or Alice's sequence knowledge in order to beat the untrained sequence success rates. A high σ_{ATV} will ensure very low adversary success rates. It is important to note that the multiple attempts might see Oscar's entry into the system. This problem could be circumvented by safeguarding against multiple authentication attempts. The optimal number of permitted attempts for a given σ_{ATV} would require further analysis.

Case 2: "Coercion" – Oscar coerces Alice to reveal her passcode knowledge

Given that Oscar knows that a 30-character passcode is required to authenticate as Alice into NeuroCrypt, his success rate hinges on Alice's ability to explicitly reconstruct her passcode sequence. In the most basic threat model, we can assume that Alice is unable to explicitly recollect the full sequence. In this case, the adversary's success rate can be formulated as $\frac{q}{|\Sigma|}$ where q is the number of iterations of coercive extraction that Oscar attempts (yielding a different passcode each time) and $|\Sigma|$ is the set of all passcodes as mentioned in 3.2. We can reasonably upper bound q as each extracted sequence would have to be trialled in the system. Since each authentication session takes a minimum of 5 minutes, we can assume that Oscar can try up to 10^5 sequences which would take him about a year of non-stop authentication attempts. This leaves us with the success rate of 2^{-21}, an extremely low adversarial success rate. Indeed, in this basic threat model, we do not combine this probability with the random untrained sequence success rate. This adversarial success rate can be further minimized by adding safeguards to prevent multiple authentication attempts by a single user in a given time frame and checks to ensure the presence and liveliness (absence of duress) of the users [32].

Case 3: "Spying" – Oscar gains a transcript of the game sequence through shoulder surfing or eavesdropping

In general, most authentication systems are vulnerable to eavesdropping. In traditional authentication methods like entering a password, gaining a transcript of the authentication action of a user guarantees success for the adversary. However, in NeuroCrypt, although adversarial success is still possible, the process will be more convoluted. Let us assume that Oscar records a video of Alice playing the Authentication Session. First, the three authentication sequences will need to be identified and then user performance on each of the sequences will need to be measured. Given that the Alice herself authenticates successfully, Oscar is only able to succeed if he extracts the sequence that the user consistently performs better on. Even without knowledge of the authentication threshold, Oscar has a pretty high success rate of passcode extraction. To make passcode extraction more difficult, we could alter the system's design such that the authentication system has sequences of varying length. The passcode could take any length between 25-35 characters and so could the untrained sequences. This makes it difficult to isolate the different sequences in the session and, in turn, makes passcode extraction also more difficult.

7 Discussion

We conduct a comprehensive user study to test the efficacy and practicality NeuroCrypt. We find that, although the difference in learned performance advantages are not significant between modalities, they deliver noteworthy insights about the inclusion of independent sensory stimuli to the authentication system. Firstly, we find that the system variants with the additional sensory stimuli are more preferable for authentication as they exhibit greater statistical significance between passcode performance and non-passcode performance as in Figure 4a.

Consistent with literature, we note that the visual stimuli does not display a drop in performance during the long training session unlike with auditory or no stimuli (Figure 3a), which suggests that the visual stimuli play a role in maintaining participant attention. In a previous psychophysics study [27], flashing stimuli were reported to be more memorable which was attributed to the classic idea that moving objects tend to capture and hold attention more readily than static objects [33]. Additionally, visual stimuli seem to induce the greatest extent of sequence learning during the training session, which may well be a by-product of the increased attentiveness of participants. It is also possible that the optimal spatial arrangement and distinctiveness of the cues from non-cues allowed players to learn their passcodes more effectively.

We also discover, from participant behavior, that the auditory stimuli elicit greater retention over time than in the control and visual groups (Figure 4c). The Auditory Scaffolding Hypothesis [28] suggests that dual exposure to auditory sounds that are temporal and sequential in nature bootstrap the neural mechanisms necessary to robustly represent other temporal and sequential information (like our passcodes). This robust scaffolding that auditory stimuli provide may be responsible for participants in the audio game to fare better in retaining their learned sequence advantages over the two week periods.

Moving on, our control group was an exact replication of the experiment conducted by Bojinov et al. [10]. Despite an identical design, we were unable to reproduce the results reported in that work. The authors reported a performance advantage of upto 14% (Figure 4 in [10]) on average during training (well above 10% with error bars) while the maximum we were able to achieve was around 7.5%. This observation also stayed consistent with our pilot studies (n=40), where neither the control group nor the sensory stimuli groups witnessed such performance advantages.

To assess the practicality of our system in a real authentication setting, we conducted analysis on participant authentication data. By probing the statistical patterns of their game responses, we are able to find optimal authentication threshold values required to confidently validate user identity. We find that the audio variant has consistently higher true user success rates and lower false success rates across threshold values than its visual and control counterparts. We find that a threshold difference of 0.025 for control and visual as well as 0.05 for the audio systems maximize the difference between true and false authentication success.

Finally, we are able to assess participants' explicit familiarity with their own passcodes. Although passcodes were, on average, rated higher than other sequences, the differences was not statistically reliable and did not correlate with their performance advantages. Additionally, we evaluate our system's defense against possible threat vectors in detail. We are able to establish that, on average, our system is more secure than traditional authentication schemes especially during coercion and eavesdropping attacks.

7.1 *Limitations*

Our study has a few limitations that may affect the validity of our results and insights. Firstly, there was little diversity among the participants in the study as all of them were university students from ages of 18-25. They possessed similar levels of education, high familiarity with operating a keyboard and competence with working on laptops with distractions in the background. Specifically, young adults are linked to superior (implicit and explicit) learning capabilities [34]. Our participant base does not effectively represent the intended user base of NeuroCrypt, thus the validity of our observations remains incomplete.

Secondly, many participants complained about the training session being far too long in the post-game survey. This could have affected their attention during training and, in turn, their implicit learning. Measures to shorten the training or enrollment phase of NeuroCrypt might yield more reliable experimental results.

Finally, the experiment across all groups was conducted on personal laptops in lecture halls. Although efforts were made to eliminate technical and logistical difficulties for participants (like standardizing hardware, maintaining silence in the room, etc.) during their training gameplay, some participants did experience distractions. A more controlled, individual experiment setup would further validate our findings.

7.2 *Future Work*

A natural and interesting direction to take the study of sensory stimuli in implicit memory authentication systems is to add new types of stimuli from the same or different sensory modalities. With the aid of haptic feedback, the tactile modality can also be integrated into the authentication system. Additionally, a multimodal combination of audio, vision (and tactile) will also be an interesting extension as their joint relationship on learning and retention may be effective. We are also interested in finding the rate of passcode advantage degradation to estimate the optimal frequency of re-training sessions for the system to be effective in a real-world setting. On the side of security and usability, systematically studying users' ability to explicitly reconstruct passcodes and the cognitive load faced by the user would be greatly helpful in consolidating plans of integrating an implicit memory-based authentication system into the modern security ecosystem. Finally, addressing our limitations by recruiting a more diverse participant pool, shortening training sessions and controlled, individual data collection is also of great interest.

8 Conclusion

In this study, we explored NeuroCrypt, an implicit memory-based authentication system as an alternative to traditional knowledge-based authentication systems. Specifically, we aimed to overcome the threat of rubber hose attacks and minimize cognitive load on the user. By integrating visual and auditory stimuli into the existing Serial Interception Sequence

Learning (SISL) task, we aimed to enhance both the implicit learning and retention of passcodes. Our participant study involving 70 users demonstrated that the inclusion of sensory stimuli, particularly auditory cues, significantly improved the long-term retention of passcodes, supporting the robustness of implicit memory in authentication.

While visual stimuli maintained user attention during training, auditory cues proved more effective in supporting long-term recall, which is important for deploying NeuroCrypt as a real-world authentication option. Additionally, we introduce the framework of authentication thresholds from user data that can be used to define the parameters of real-world variants of NeuroCrypt. These thresholds will be set to maximize true authentication attempts while minimizing random, false attempts.

Our work highlights the potential of implicit memory systems to reduce the cognitive burden on users while safeguarding against coercion attacks. Although NeuroCrypt demonstrates significant promise in including sensory stimuli to authentication systems, future work should investigate the integration of additional sensory modalities, refine training processes, and test the system with a more diverse participant base. Ultimately, NeuroCrypt presents a step forward in the development of secure authentication mechanisms that leverage the untapped potential of implicit memory.

References

[1] Angela Sasse Anne Adams. Users are not the enemy. *Communications of the ACM*, 42(12):41–46, 1999.

[2] Digital Guardian. Uncovering password habits: Are users' password security habits improving? (infographic). https://digitalguardian.com/blog/uncovering-password-habits-are-users-password-security-habits-improving-infographic, Dec 2018.

[3] Robert Morris and Ken Thompson. Password security: a case history. *Commun. ACM*, 22(11):594–597, nov 1979. ISSN 0001-0782. doi: 10.1145/359168.359172. URL https://doi.org/10.1145/359168.359172.

[4] Harris Poll survey Google. Online security survey google / harris poll. Available at http://services.google.com/fh/files/blogs/google_security_infographic.pdf (Feb 2019).

[5] Henry L Roediger. Implicit memory: Retention without remembering. *American psychologist*, 45(9):1043, 1990.

[6] Paul J Reber. The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, 51(10):2026–2042, 2013.

[7] Daphna Weinshall and Scott Kirkpatrick. Passwords you'll never forget, but can't recall. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1399–1402, 2004.

[8] Tamara Denning, Kevin Bowers, Marten Van Dijk, and Ari Juels. Exploring implicit memory for painless password recovery. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2615–2618, 2011.

[9] Claude Castelluccia, Markus Dürmuth, Maximilian Golla, and Fatma Deniz. Towards implicit visual memory-based authentication. In *Network and Distributed System Security Symposium (NDSS)*, 2017.

[10] Hristo Bojinov, Daniel Sanchez, Paul Reber, Dan Boneh, and Patrick Lincoln. Neuroscience meets cryptography: Crypto primitives secure against rubber hose attacks. *Communications of the ACM*, 57(5):110–118, 2014.

[11] Paul J Reber, Daniel J Sanchez, and Eric W Gobel. Models of sequential learning. In *Proceedings of the eighth international conference on complex systems. NECSI (This Volume) https://www. researchgate. net/publication/228756767*, 2011.

[12] Marcus J. Ranum. Cryptography and the Law. https://groups.google.com/g/sci.crypt/c/W1VUQlC99LM, October 1990.

[13] Kevin Edward Fu. *Group sharing and random access in cryptographic storage file systems*. PhD thesis, Massachusetts Institute of Technology, 1999.

[14] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier, 1968.

[15] Nelson Cowan. Working memory underpins cognitive development, learning, and education. *Educational psychology review*, 26:197–223, 2014.

[16] Natalie Ruiz. *Cognitive load measurement in multimodal interfaces.* PhD thesis, University of New South Wales, Sydney, Australia, 2011.

[17] Philipp Taesler, Julia Jablonowski, Qiufang Fu, and Michael Rose. Modeling implicit learning in a cross-modal audio-visual serial reaction time task. *Cognitive Systems Research*, 54:154–164, 2019.

[18] H.E. Schendan. Implicit memory. In V.S. Ramachandran, editor, *Encyclopedia of Human Behavior (Second Edition)*, pages 400–409. Academic Press, San Diego, second edition edition, 2012. ISBN 978-0-08-096180-4. doi: https://doi.org/10.1016/B978-0-12-375000-6.00200-7. URL https://www.sciencedirect.com/science/article/pii/B9780123750006002007.

[19] Saul McLeod. Implicit vs. explicit memory, 2023. URL https://www.simplypsychology.org/implicit-versus-explicit-memory.html. Accessed: 2024-09-12.

[20] Emmanuel M Pothos. Theories of artificial grammar learning. *Psychological bulletin*, 133(2):227, 2007.

[21] Werner Sævland and Elisabeth Norman. Studying different tasks of implicit learning across multiple test sessions conducted on the web. *Frontiers in psychology*, 7:808, 2016.

[22] Claude Castelluccia, Markus Dürmuth, Maximilian Golla, and Fatma Deniz. Towards implicit visual memory-based authentication. In *Network and Distributed System Security Symposium (NDSS)*, 2017.

[23] Marvin M Chun and Yuhong Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, 36(1):28–71, 1998.

[24] Annabelle Goujon, Andre Didierjean, and Simon Thorpe. Investigating implicit statistical learning mechanisms through contextual cueing. *Trends in cognitive sciences*, 19(9):524–533, 2015.

[25] Jascha Russeler, Thomas F Münte, and Frank Rösler. Influence of stimulus distance in implicit learning of spatial and nonspatial event sequences. *Perceptual and Motor Skills*, 95(3):973–987, 2002.

[26] Mariam Adawiah Dzulkifli and Muhammad Faiz Mustafar. The influence of colour on memory performance: A review. *The Malaysian journal of medical sciences: MJMS*, 20(2):3, 2013.

[27] Harold Pashler and Christine R Harris. Spontaneous allocation of visual attention: Dominant role of uniqueness. *Psychonomic Bulletin & Review*, 8(4):747–752, 2001.

[28] Christopher M Conway, David B Pisoni, and William G Kronenberger. The importance of sound for cognitive sequencing abilities: The auditory scaffolding hypothesis. *Current directions in psychological science*, 18(5): 275–279, 2009.

[29] Jennifer C Romano Bergstrom, James H Howard Jr, and Darlene V Howard. Enhanced implicit sequence learning in college-age video game players and musicians. *Applied Cognitive Psychology*, 26(1):91–96, 2012.

[30] Göran BW Söderlund, Sverker Sikström, Jan M Loftesnes, and Edmund J Sonuga-Barke. The effects of background white noise on memory performance in inattentive school children. *Behavioral and brain functions*, 6:1–10, 2010.

[31] Willem Verheyen. Adaptive thresholding for fair and robust biometric authentication. In *Proceedings of the 24th International Middleware Conference: Demos, Posters and Doctoral Symposium*, pages 7–8, 2023.

[32] Jeremy Clark and Urs Hengartner. Panic passwords: Authenticating under duress. *HotSec*, 8:8, 2008.

[33] Walter Bowers Pillsbury. *Attention*. S. Sonnenschein & Company, Limited, 1908.

[34] Richard Midford and Kim Kirsner. Implicit and explicit learning in aged and young adults. *Aging, Neuropsychology, and Cognition*, 12(4):359–387, 2005.

Appendix

A Data and Code Availability

We are uploading all data and code used in this study to this **OSF Repository** for reproducability purposes. The following can be found in the repository:

- Code to deploy the NeuroCrypt web application
- Cleaned data collected in our User Study (n=70)
- Code used to analyze and generate the plots in this paper
- Video snippets of each NeuroCrypt variant (control, visual and audio)

B Passcode Sequence Generation Algorithm

Algorithm 1: Passcode Sequence Generation Algorithm

```
Procedure generatePassSeq():
    passSeq = [];
    possibleItems = [s, d, f, j, k, l];
    foreach i ∈ {1, ··· , 30} do
        if i = 1 then
            randIndex :− generateRandomInt(0, 5) randItem :− possibleItems[randIndex]
            passSeq.append(randItem)
        end
        else
            Set randItem = 0;
            while (randItem == 0 OR passSeq[−1] == randItem) do
                randIndex :− generateRandom(0, 5)
                randItem :− possibleItems[randIndex]
            end
            passSeq.append(randItem)
        end
    end
    return passSeq;
```

C Variability of trained passcode advantages across participants

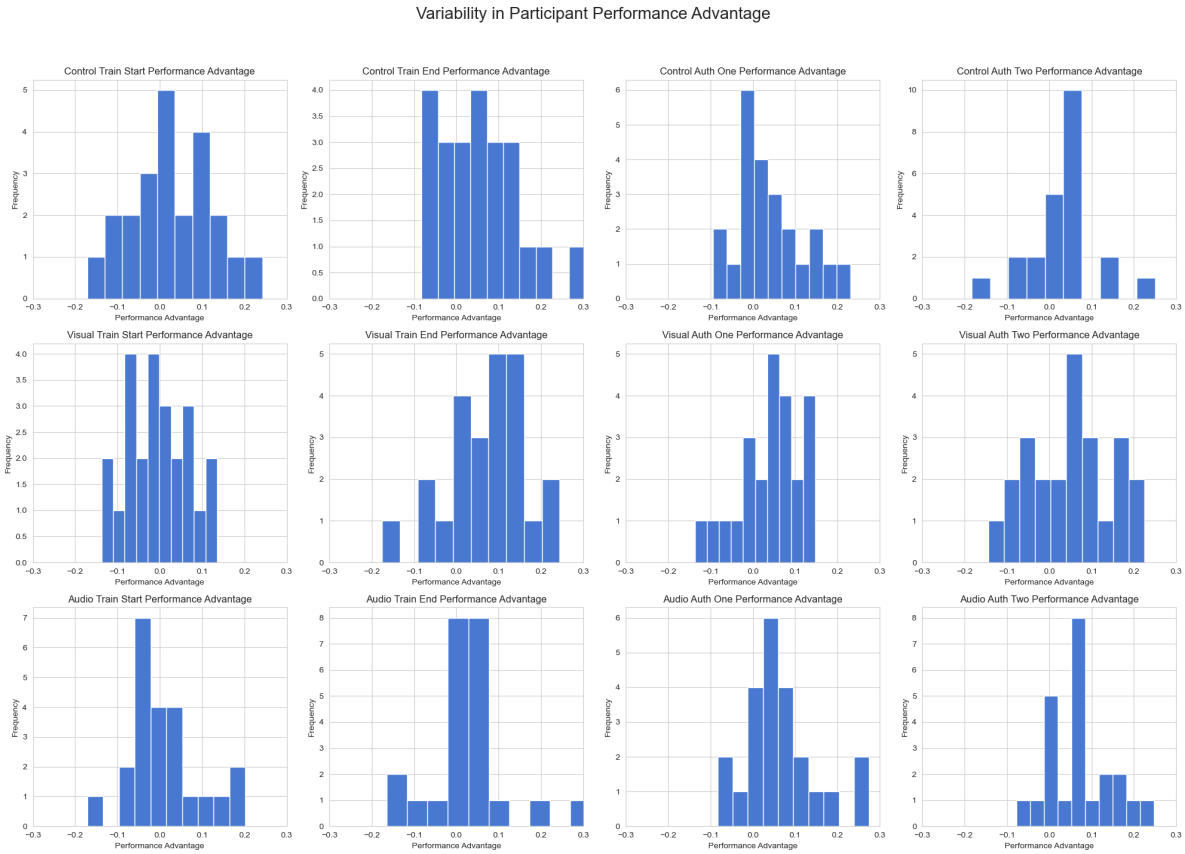


Figure 5: **Participant Performance Advantage variability across modalities and sessions**

D Visualization of the effect of Authentication Threshold Values (σ_{ATV}'s)

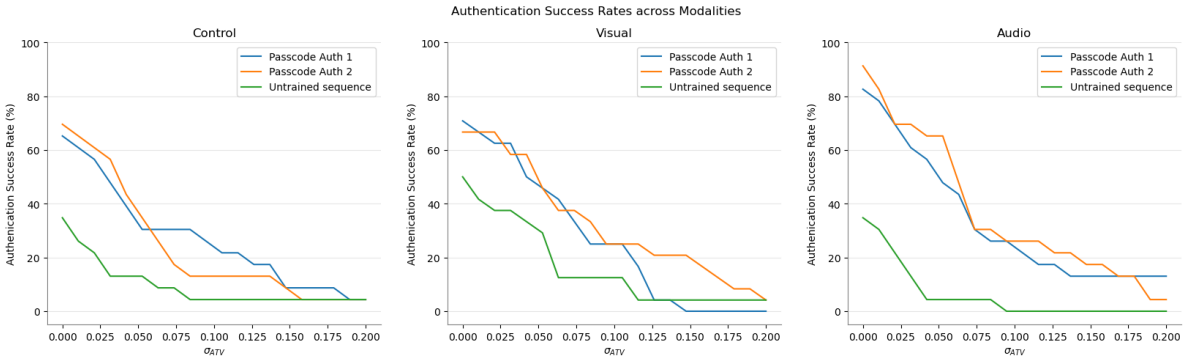


Figure 6: **Authentication success rates for Auth One, Auth Two and untrained sequences**

Sigma	Diff (Control)	Diff (Vis)	Diff (Aud)
0.000	32.61	18.75	**52.17**
0.025	**43.48**	**27.08**	47.83
0.050	19.57	16.67	**52.17**
0.075	17.39	22.92	23.91
0.100	13.04	12.50	23.91

Table 2: Difference in true authentication success rate (from Auth One and Auth Two) and the untrained, false authentication success rate. The values in bold are the maximum value for each modality – corresponding to optimal σ_{ATV}'s

E Explicit Familiarity Results

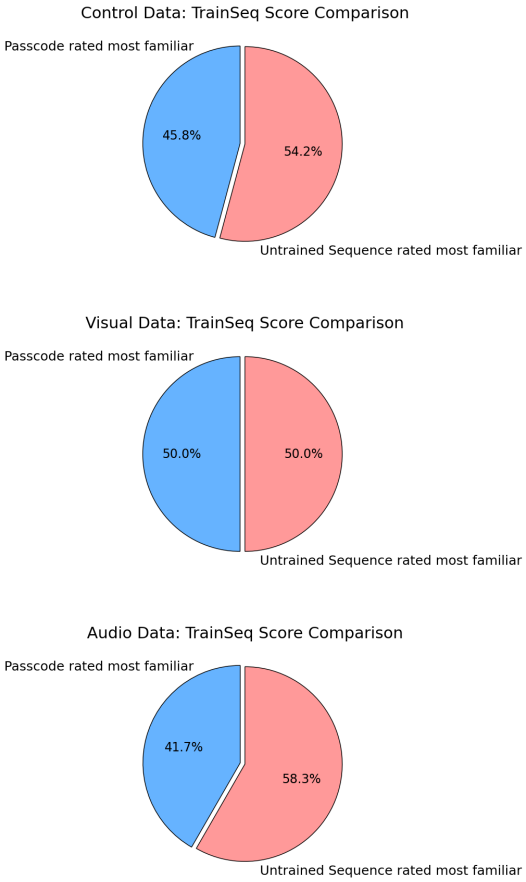


Figure 7: **Explicit familiarity ratings for different modality groups**