

USENIX

THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Not All Is Lost: Partial Recovery & Memorability in Self-Sovereign Digital Identity

Sushanth Ambati, Dylan Jablonski, Charalampos Papachristou, Brian Pugliese, Jack Myers, and Nikolay Ivanov, *Rowan University*

https://www.usenix.org/conference/soups2026/presentation/ambati

This paper is included in the Proceedings of the Twenty-Second Symposium on Usable Privacy and Security.

August 24–26, 2026 • Hannover, Germany

ISBN 978-1-939133-59-5

Open access to the Proceedings of the Twenty-Second Symposium on Usable Privacy and Security is sponsored by USENIX.

Not All Is Lost: Partial Recovery & Memorability in Self-Sovereign Digital Identity

Sushanth Ambati
Dept. of Computer Science
Rowan University

Dylan Jablonski
Dept. of Computer Science
Rowan University

Charalampos Papachristou
Dept. of Mathematics
Rowan University

Brian Pugliese
Dept. of Psychology
Rowan University

Jack Myers
Dept. of Computer Science
Rowan University

Nikolay Ivanov
Dept. of Computer Science
Rowan University

Abstract

The shift toward Self-Sovereign Digital Identity (SSDI) promises to mitigate the epidemic of massive personal data leaks, but it fundamentally relies on users securely managing high-strength private keys. Traditional backup methods, such as hardware wallets, paper records, and cloud vaults, inevitably compromise either security, locality, or availability. While human memory offers a uniquely theft-proof and local alternative, the cognitive burden of memorizing *strong* cryptographic keys remains a critical usability barrier. Crucially, prior usable security research has largely treated key memorability as a binary outcome — either perfectly recalled or completely forgotten. We challenge this paradigm through a lab-based randomized controlled trial ($N = 92$) of 128-bit (NIST-minimum) keys. Comparing standard 12-word BIP-39 mnemonics against a graphical key scheme, we demonstrate how combinatorial search can leverage fragmented human memory to reconstruct keys without degrading theoretical security bounds. We present an empirical evaluation of how cognitive-science principles affect memorability of full 128-bit system-assigned keys, paired with a novel computational recovery technique. Our findings reveal an advantage for graphical keys, suggesting that augmenting BIP-39 mnemonics with graphical cues and combinatorial recovery of partially forgotten keys may offer a promising path forward.

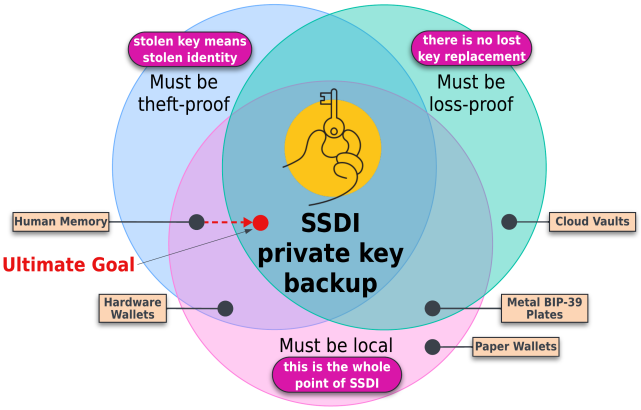


Figure 1: **SSDI private key backup desiderata.** An ideal backup is simultaneously local, theft-proof, and loss-proof, yet every existing approach sacrifices at least one. Placements are relative, not absolute: metal BIP-39 plates fall inside the loss-proof region because they resist fire, water, and decay far better than paper, though no physical medium is strictly loss-proof. Human memory already provides theft-proofness and locality; the dashed arrow marks our goal of extending it into the loss-proof region through partial recovery of imperfectly remembered keys.

1 Introduction

The past two decades of catastrophic data breaches have exposed billions of records, underscoring the inherent vulnerabilities of Centralized Digital Identities (CDIs) [2, 4, 30]. Federated Digital Identities (FDIs), offered by platforms like Google or Microsoft, suffer from similar single points of failure, exemplified by the 2018 Google+ breach where API flaws exposed millions of user profiles [42]. In response to these systemic failures, the Self-Sovereign Digital Identity (SSDI) paradigm shifts administrative control directly to users, allowing them to exclusively manage their identity data and cryptographic assertions [41, 49, 55].

While cryptocurrency wallets are the most recognized application of SSDI, the paradigm is rapidly expanding to encompass broader digital identity use cases. Unlike centralized or federated models, SSDIs are not stored in the proprietary

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23–26, 2026, Hannover, Germany.

databases of identity brokers; instead, they rely on a locally managed single private key[1]. While this architecture promises unprecedented privacy and security, it introduces severe usability hurdles [65]. Chief among these is the challenge of reliable, user-friendly *private key backup*, which forms the focal point of this work.

As illustrated in Figure 1, an ideal key backup mechanism must satisfy three conflicting properties: it must be local, theft-proof, and loss-proof. Traditional physical backups — such as paper wallets, metal seed plates, and hardware wallets — substitute digital vulnerabilities for physical ones. Theft, accidental loss, insider threats, and environmental degradation remain ever-present risks [12]. A misplaced steel plate is no less a single point of failure than a breached server; it is merely a different one.

1.1 The Limitations of Centralized Vaults

A widely adopted response to credential-management complexity is the centralized credential vault, of which password managers and custodial key-backup services are representative examples: a single master credential decrypts a stored collection of many independent secrets. This consolidation creates a high-value target. The 2022 LastPass breach exposed millions of encrypted credential stores to offline attack [26], and Chatzoglou et al. [16] showed that several popular vault implementations retain master secrets and decrypted material in plaintext within main memory, permitting direct RAM extraction by an attacker with local access. Even when the cryptographic store itself remains intact, convenience features expand the attack surface: autofill enables phishing and confused-deputy attacks on mobile platforms [43], while desktop evaluations reveal susceptibilities to clickjacking [44, 53]. More fundamentally, SSDI systems derive their security guarantees from the elimination of trusted third parties, so delegating key custody to an external cloud service or proprietary software ecosystem reintroduces the exact concentration of trust these protocols seek to avoid.

1.2 Reimagining Backup via Human Memory

The above limitations motivate investigation into a different, highly local backup medium: *human memory*. Human memory is inherently non-transferable, requires no third-party trust, and accompanies the user at all times. However, the cognitive burden of memorizing *strong*, 128-bit cryptographic keys has long been considered an insurmountable usability barrier, motivating an extensive line of work on memory palaces, mnemonic devices, and spaced rehearsal as mitigations [9, 10, 13]. Crucially, prior usable security research has almost exclusively treated key memorability as a binary outcome — a key is either perfectly recalled or completely forgotten. We investigate whether fragmented recall retains enough information to support computational recovery.

1.3 Our Research Plan and Contributions

Our core contributions are as follows:

- **Cognitive-Driven Key Space Design:** We present a principled design for a 128-bit graphical key space grounded in cognitive psychology literature (§4). Rather than relying on arbitrary parameters, we carefully balance the entropy requirements of NIST-compliant keys with human cognitive limits, formulating the foundation for our short-term empirical study.
- **Baseline Evaluation of 128-bit Keys:** We design and execute a controlled, lab-based randomized AB/BA crossover trial ($N = 92$) to measure the baseline memorability of 128-bit private key seeds (§5). By comparing standard 12-word BIP-39 mnemonics against our graphical key scheme, we empirically evaluate fundamental differences between alphanumeric and non-alphanumeric cognitive retention over a one-week period[2].
- **Single-Scene Representation:** We encode a full 128-bit key within one static scene rather than a multi-scene journey or memory-palace traversal. This choice isolates representation from spatial navigation and, critically, breaks the key into independent attribute slots that our bounded recovery procedure (§6.2) can search efficiently.
- **Partial Key Recovery:** We challenge the binary paradigm of memorability by introducing and evaluating computational error correction for human memory (§6.2). Applying combinatorial search to fragmented user recall, we successfully rescue 50% of "lost" graphical keys and show 34.1% recovery potential for “lost” BIP-39 mnemonics, demonstrating the potential to securely recover lost identities without degrading cryptographic bounds.
- **Proposed Path Forward:** Building on our behavioral findings, including the observation that only 16.2% of users failed to recall both key types, we formulate the hypothesis that graphical and BIP-39 representations provide complementary memory cues. Our findings motivate future investigation into whether graphical cues paired with BIP-39 mnemonics, alongside combinatorial recovery techniques, can improve memorability and recovery from partial recall failures.

Availability: In the spirit of open science, we openly share all code, anonymized datasets and other artifacts necessary for reproducing our study: https://github.com/acsusla

[1] A nearly infinite number of unique keys can be derived from the initial seed using a hierarchical deterministic scheme, such as BIP-32.

[2] The cognitive science-based justification of the study timeline is provided in §5.5.

b/napks-artifacts and https://github.com/acsusla b/napks

2 Background

BIP-39 Mnemonics: Bitcoin Improvement Proposal 39 (BIP-39) specifies a standard for generating mnemonic codes that serve as human-readable representations of cryptographic keys [45]. BIP-39 mnemonics are considered the de facto state-of-the-art local key backup standard for crypto wallets, used in paper backups, hardware wallets and metal vaults[3].

Key Strength: The term "key strength" refers to the effective security provided by a cryptographic key — commonly expressed in bits — and reflects the amount of work (e.g., number of operations) required in an optimal attack. It is often quantified in terms of the min-entropy of the key generation process, rather than simply its nominal length [6]. NIST prefers this notion of key strength [5] because it more accurately captures the real-world feasibility of brute-force attacks, whereas raw key length may overestimate security if the key material is derived from a generator with insufficient entropy. The concept of "bit entropy" refers to the actual unpredictability inherent in the key material; in many contexts, key strength and bit entropy align, but the key strength metric emphasizes what an attacker can practically exploit. As an illustrative example, the popular 12-word BIP-39 mnemonic generates a 256-bit seed, but due to the inclusion of checksum bits derived from SHA-256 (specifically, 128 bits of entropy plus a 4-bit checksum) the true randomness, and thus the effective entropy, is only 128 bits [7, 24].

3 Related Work

3.1 Forms of Memory

Human memory is not a single, uniform faculty; cognitive psychology distinguishes various forms of memory.

Episodic vs. Semantic: Human memory has two main systems described in declarative memory: episodic for memory of personal events and semantic for facts [59]. Passwords are stored primarily as semantic facts, which may be difficult to retrieve without contextual cues [1, 54]. Our design leverages the context-to-memory benefit by embedding cryptographic secrets into visual scenes that provide strong visual cues [19].

Recall-based vs. Recognition-based: When retrieval is recall-based, the secret must be generated from memory, as in free-entry text passwords, BIP-39 mnemonics, or unaided graphical recall. This imposes the highest cognitive load and is more susceptible to memory decay. Recognition-based retrieval involves selecting from visible alternatives, which typically reduces cognitive load compared to recall-based retrieval [38].

Cued recall: Cued recall occupies an intermediate position between pure recall and recognition. Structured cues (e.g., images, spatial layouts, or environmental context) assist retrieval by reinstating contextual information associated with encoding [58], but users must still reconstruct portions of the secret from memory. Examples include click-based graphical passwords and loci-based schemes, where visual context or spatial layouts guide recall without fully presenting the secrets [58, 60].

3.2 Loci-Based Authentication Approaches

A memory palace is a mnemonic technique, based on the Method of Loci [35], in which a person imagines a familiar physical space and places items to be remembered at specific locations within it, allowing them to recall the information by mentally walking through the scene.

LociMotion [20] adopts a loci-based approach, investigating whether users can learn a strong authentication secret through a dynamically generated video house tour, associating twelve landmarks with password characters in sequence. The scheme restricts passwords to lowercase letters, yielding approximately 56 bits of entropy. While LociMotion demonstrates that a single-session, loci-based presentation can support short-term learning of a moderate-strength secret, the evaluated security level is lower than modern security standards.

Clark et al. [18] investigated whether users could memorize strong, random, system-assigned passwords using visual techniques. In their approach, participants mentally traverse a structured route through multiple scenes, observing various objects in different states as specific spatial locations ("loci") along the way. The variations in scenes, objects, and object states collectively contribute to the password's cryptographic strength. Their design was based on SeedQuest [11], which encodes a 12-word BIP-39 mnemonic as a memory-palace *SeedQuestPath* of six scenes with objects in four possible states, but they measured recall of a *SeedQuestPath* based on *five* BIP-39 words (55 bits of entropy). Their graphical technique achieved higher recall than passphrases.

Our study was inspired by their explicit call for further memorability evaluation of strong private keys and builds on theirs in two key ways: (1) we evaluate memorability at the full 128-bit entropy level, and (2) we simplify the interaction model by replacing a multi-scene "journey" with a single static image.

Vlahavas et al. [62] conducted a study evaluating memorability of BIP-39 mnemonic phrases using the Method of Loci within a "serious game" environment called *MnemonicMaker*. Their control group, which attempted to memorize a 12-word BIP-39 phrase without the game, reported a 21% success rate (4/19 participants) after 7 days, compared to approximately

[3] Many metal vaults only use the first four letters of each word, which are unique among the 2048-word dictionary.

50% in our study. Differences in study design, participant populations, and memorization procedures may contribute to this discrepancy, although direct comparison remains difficult.

MnemonicMaker and our study examine related questions from somewhat different perspectives. While *MnemonicMaker* focuses on game-based memorization of BIP-39 phrases, our work examines memorability under a controlled study design and additionally considers recovery from partially forgotten keys (§6.2). This broader framing allows us to explore not only whether strong keys can be remembered, but also whether fragmented recall can still be useful during recovery.

Reminisce is a key-generation approach that aims to improve memorability by linking cryptocurrency keys to meaningful personal photographs rather than random word lists. It leverages spatial and visual memory cues. Users select distinctive images associated with real-life locations, and recalling those experiences helps regenerate the key [51]. Unlike our work, the method relies primarily on remembered contextual metadata rather than the images themselves. Reminisce incorporates user-selected personal data and therefore does not rely solely on fully random system-assigned secrets. In contrast, our design avoids personal information and encodes entropy through abstract graphical elements. Furthermore, while Seo et al. provide entropy estimates, the resulting key strength is not fully characterized within their evaluation.

3.3 Spaced Repetition

Spaced repetition, a memorization protocol in which review intervals are gradually expanded across multiple sessions, is an extensively studied technique for committing high-strength secrets to memory [15]. *Bonneau and Schechter* [13] showed that users can reliably store 56-bit secrets when training is distributed across multiple sessions with expanding intervals. *Blocki et al.* [9] proposed a framework in which strong system-assigned passwords are anchored to a personal memory palace and naturally rehearsed through routine logins, providing theoretical bounds on how many such secrets a user can sustain. Their subsequent empirical study [10] confirmed that mnemonic devices combined with scheduled rehearsal enable recall of multiple strong passwords. *Shay et al.* [52] showed that the representation of a system-assigned secret (e.g., random characters vs. passphrases) has a measurable impact on memorability, foreshadowing the representational shift that motivates our graphical scheme. *Biddle et al.* [8] survey the first twelve years of graphical password research, situating recall-, recognition-, and cued-recall-based schemes within a broader design space.

Our study differs from this body of work along three deliberate axes. First, we evaluate memorability at the full 128-bit NIST minimum rather than the 56-bit level common in prior spaced-repetition studies, closing the gap between what cognitive work measured and what modern cryptographic systems require.

Second, we use a single static scene rather than a multi-scene journey or memory palace. This is a methodological commitment, not a simplification. The scene serves two purposes. It isolates representation from spatial navigation, which prior schemes [11, 18, 20] confound, and it removes the sequence-and-transition errors that compound with content forgetting over long retention intervals.

A single static scene also enables our recovery method. A single scene factors the key cleanly into independent, slot-addressable attributes, and that structure is what makes bounded computational recovery (§6.2) tractable. Journey representations entangle these attributes through scene ordering and transitions, thereby expanding the search space that recovery must cover. The static scene is therefore a precondition for our contribution to recovery, not incidental to it.

Third, and most centrally, we introduce *computational recovery* (§6.2): a partially correct recall is treated as a useful signal rather than a failure, with bounded search closing the residual gap. We do not employ spaced repetition because an SSDI backup is created in a single enrollment session and recovered only on rare, high-stakes occasions (§4.1); a single-session protocol therefore mirrors the realistic operational profile of SSDI backup and yields a conservative lower bound on recall.

3.4 Partially Forgotten Keys

Recovering a partially remembered graphical key can be viewed as a constrained search process, where users reconstruct the secret by exploring remembered visual components rather than recalling it perfectly. In this context, bounded brute-force exploration becomes a practical recovery strategy guided by partial knowledge of attributes such as icon appearance, color, size, or spatial placement. Previous research in graphical authentication helps explain why such structured search is effective.

Van Oorschot et al. [61], in studies of click-based graphical passwords, show that users tend to select visually salient regions, producing clustered “hotspots” rather than uniform choices. Although originally discussed as a security concern, these patterns suggest that recovery can focus on cognitively meaningful elements instead of the entire key space.

Stobert et al. [57, 58] enable users to authenticate by clicking specific locations on images; they show that users are able to remember recognition-based graphical passwords better than recall-based graphical passwords. The heatmaps they report show consistent positional biases in graphical password entry (some parts of an image are more "clickable" than others), suggesting that structured visual encoding can guide constrained recovery when memory is incomplete. Our work builds on these insights by framing constrained brute-force exploration as a user-assisted mechanism for reconstructing partially remembered graphical keys. Their research, despite

studying lower entropy passwords, focuses heavily on the user experience and highlights the location of a click as a core authentication component, which we could explore in future work.

4 System Design and Implementation

We implement our human memory testing system as a web application using Node.js and Express.js on the backend, with EJS templates for server-side rendering. A MySQL database stores all study data, with a schema designed to capture participant responses and performance metrics. The application is deployed on an AWS server to enable parallel human trials.

We use the `bip39` module in Node.js to generate 12-word mnemonic phrases, providing 128 bits of security strength for the alphanumeric key. For the graphical key, we curate a set of exactly 810 distinct SVG icons. To ensure equal probability during graphical key generation, we employ a JavaScript-based randomizer in combination with the Fisher-Yates shuffle algorithm. During training, real-time voice prompts are provided using the browser's built-in *SpeechSynthesisUtterance* interface from the JavaScript Web Speech API rather than prerecorded human audio.

To support this comparative study, the system includes two distinct subsystems — one for BIP-39 keys and another for graphical keys. Each subsystem includes its own training and authentication flows, developed to mirror each other in structure and functionality. This alignment is vital to ensure consistency across both approaches and maintain fairness in evaluating their usability and effectiveness.

4.1 Usage Model

SSDI key backup serves a fundamentally different role from a routinely-entered password. In normal operation, the user's device retains the unlocked private key and grants access through a local PIN or biometric check, in the same manner as a hardware wallet or mobile authenticator. The memorized backup is invoked only when this device is lost, stolen, damaged, or replaced (e.g., during migration to a new phone), events that for most users arise rarely. SSDI seed recovery therefore occupies the same operational category as hardware-wallet seed recovery today: an infrequent, high-stakes operation rather than a daily friction point.

This usage profile reshapes what counts as acceptable usability. Because the seed is entered so seldom, a multi-minute authentication ceremony is tolerable, particularly when the alternative is the loss of the user's identity. Unlike a forgotten website password, a forgotten SSDI seed cannot be reset, because the SSDI paradigm deliberately removes the trusted third party that would otherwise reissue credentials. Failure is therefore not an inconvenience but a terminal event, with the affected identity becoming permanently unrecoverable. This asymmetry, where the cost of total failure vastly exceeds the cost of slow success, motivates our central design choice: rather than insist on perfect recall, we treat partial recall as a useful signal and apply computational search to close the gap.

4.2 Threat Model

The recovery procedure introduced in this work is designed so that the theoretical security bound of the underlying key is preserved while the practical attack surface remains tightly scoped. We assume the standard SSDI setting: a single private key seed is the sole basis of identity, no trusted third party can reissue it, and the user's device stores whatever auxiliary material is needed for routine wallet operation.

What is stored. The wallet stores its key material following best practices observed in modern cryptocurrency wallets [23, 32]: the seed is never written to disk in plaintext but only as an authenticated ciphertext, and a memory-hard key derivation function gates any attempt to verify a candidate seed. During recovery, the routine treats this stored ciphertext as a verification oracle, accepting a candidate only when its authentication tag verifies. No representation of the seed itself (graphical or BIP-39) is otherwise retained on the device.

Who can run recovery, and under what conditions. The recovery algorithm runs locally on the user's device and requires both the stored check value and a user-supplied attempt sufficiently close to the true key. A zero-knowledge attacker lacking these inputs must search the full 2^{128} space. An attacker who steals the device may obtain the check value, but without substantial partial knowledge of the key, the attacker cannot reach the reduced search space explored by recovery and remains effectively faced with the full 2^{128} bound.

Rate limiting and containment. Because the check value can be used to verify candidate keys, production deployments should employ standard hardware-wallet protections such as rate limiting, bounded authentication attempts, and secure storage of verification material. These measures preserve the full theoretical security bound for attackers lacking substantial prior knowledge of the key while still enabling recovery for legitimate users who remember most of their seed.

4.3 Training and Authentication Interfaces

The training interfaces incorporate serial learning, where participants first learn and recall one subset of their key components (either a BIP-39 word or a graphical icon) before the next subset is introduced. Word recall is cumulatively tested at each training step, as shown in Fig. 2. Icon recall is cumulatively tested in a similar fashion.

During training and authentication, participants are presented with choice suggestions, allowing them to use lists or grids to select all values of their key segments through a cued-recall process supported by visible alternatives. During training, words and icons are presented in both visual and

auditory modalities. At any point during training, if a participant enters an incorrect word or identifies an incorrect icon, the step is repeated until the participant successfully recalls the correct information before proceeding to the next memorization step (see §4.7 for the rationale behind these design decisions). In all training sessions for both subsystems, prerecorded instructional videos provide consistency of communication between subjects and investigators.

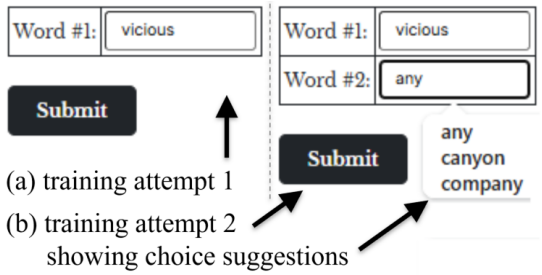


Figure 2: **BIP-39 training interface.**

4.4 Design Rationale

To match the 128-bit strength of BIP-39 mnemonics, our graphical key needs sufficient variability, but must also remain within the limits of human memory. Our graphical key most closely resembles a cued-recall paradigm while incorporating recognition-based elements.

During authentication, users select backgrounds from a list of possibilities that will serve as retrieval cues. They later place correctly colored and sized icons onto these cues. This placement constitutes a recall task, while the terrain associated with that region serves as a contextual cue to aid icon identification. Once users recall a target object, visible icon and attribute grids provide recognition-based support during selection.

The initial idea was to place images within a map-like layout, segmented into five distinct regions to introduce contextual relevance in a natural scene, maximizing memorability [31]. We embed our icons into a background context[4] rather than relying on the icons alone, as people have better memory when they focus on the *meaning of pictures* (i.e., elaborative encoding through semantic processing) rather than just on their surface features (perceptual processing) [34, 56].

Placing two images per region allows us to use 5 pairs of images (10 total). To reach the required level of bit strength, each image would need to be selected from a pool of 1,620 distinct images (see Section 4.5 for details).

However, if all 10 images were selected from the *same* shared pool, it could lead to confusion — especially if similar images appear more than once. To avoid this, we assign each image to a separate pool of 1,620 unique images. This would require 16,200 total images, a clearly unrealistic and inefficient design choice for both development and usability. To avoid the need for thousands of unique images, we shift from photographic images to modifiable SVG icons, which can be altered programmatically on demand.

We handpick a set of 810 distinct icons, divided evenly into ten pools of 81 unique icons each. Using 81 icons per pool provides sufficient visual variety while keeping the overall icon set manageable. Additional visual attributes — size and color — are then introduced. Allowing each icon to have one of ten colors and one of two sizes effectively yields 1,620 unique combinations per pool, reaching our target.

A 128-bit key requires participants to remember 10 icons along with their colors, sizes, and terrain types. Counted as separate items, that is a heavy memory load. However, people do not have to remember items individually. As noted by Miller [39], grouping individual items into larger, more meaningful chunks allows a person to hold and retain more information [22]. Therefore, what matters for our design is not how many attributes there are, but how many chunks each participant forms.

Our design helps participants form these chunks. The regions and terrains give the scene a structure, and items that sit together are easy to group into a small story or idea. Participant feedback (Appendix C) shows users did exactly this, for example, tying an icon to its terrain through a small story. By encouraging this kind of meaningful grouping, the design works with how memory consolidates over time rather than against a fixed capacity limit.

Prior work in information visualization and human–computer interaction shows that users can reliably distinguish approximately eight to twelve categorical hues when colors are perceptually separated and presented simultaneously [28, 29, 63]. We therefore use a ten-color palette that remains within this supported range while expanding the key space. The colors are selected from the Krzywinski Colorblind 15 palette[5] for accessibility [37].

We deliberately limit the number of discrete sizes to two (large and small) to force a binary decision. Visual encodings should be sufficiently distinct to support reliable discrimination during identification; closely spaced values increase the risk of confusion and misclassification [63]. Together with color, the size attribute contributes to the overall key space without introducing fine-grained visual distinctions.

Other combinations of attributes are possible, but the proposed configuration achieves a reasonable balance between usability and the need for a sufficiently large key space. This allows the system to produce one randomly generated, system-assigned scene with sufficient variation, as depicted in Fig. 3.

[4] Icons and backgrounds were designed by the authors and informally reviewed by the research team and students in a research methods course for recognizability, clarity, distinguishability, and appropriateness.

[5] While the Krzywinski Colorblind palette was selected for accessibility reasons, some color pairs (e.g., pink/orange) may remain perceptually similar under particular viewing conditions, potentially affecting memorability and retrieval performance.



Figure 3: **Sample 128-bit graphical key.**

Beyond icons, colors, and sizes, the final component of our graphical key is the terrain. Our model includes two terrain options per region. Fig. 3 shows a scene comprising sand, water, snow, sky, and trees. The alternate terrains (in order) are rocks, a road, grass, mountains, and a cave (see Fig. 4).

4.5 Constrained Randomness

Although individual icon attributes (color and size) are selected at random, scenes are generated dynamically according to a fixed regional structure and terrain constraints. Each scene is divided into five regions to promote visual coherence and memorability. The constraints are as follows: (a) **Interchangeability**, the contours of each terrain must align with the shape of its assigned region; (b) **Uniqueness**, no terrain appears more than once in a given scene; (c) **Real-world consistency**, terrains must follow everyday logic, such as roads never appearing above the sky; (d) **Region exclusivity**, each terrain is restricted to a specific region.

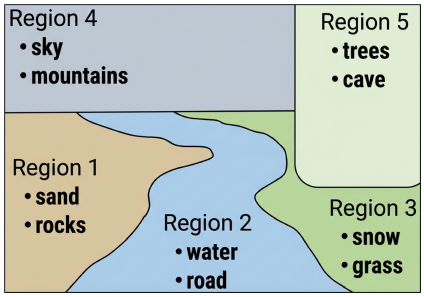


Figure 4: **Dynamic scene composition: regions and terrains.**

Given these constraints, there are 32 unique terrain configurations per background, henceforth referred to as t. Each region is populated by one of two possible terrains (see Fig. 4) and exactly two icons, which are randomly placed into the regions. Ten icons can be placed into five regions in 113,400 possible arrangements,[6] which we will refer to as r.

[6] Calculated as $\binom{10}{2} \times \binom{8}{2} \times \binom{6}{2} \times \binom{4}{2} \times \binom{2}{2}$, assigning two icons per region without repetition.

4.6 Calculating the Graphical Key

While the scene setup of icons with specific attributes across the regions defines the graphical key, it cannot be used directly for authentication. Conversion to a numerical key is required for integration with standard systems that rely on cryptographic public/private key pairs. This conversion must preserve the bits of security strength while producing a *continuous, gap-free numerical sequence in which every value occurs with equal probability*.

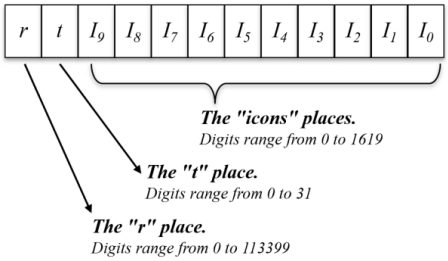


Figure 5: **Places of the graphical key seed.**

Our solution employs a positional number system that begins by converting the icon-specific information into a single base-1620 value. Just as decimal numbers use base 10 and binary numbers use base 2, our encoding uses base 1,620 because each icon can assume 1,620 distinct combinations of icon (i), color (c), and size (s), resulting in a value from 0 to 1,619 for each icon.

The following equation:

$$I = i \times 20 + c \times 2 + s, \quad \text{where } i \in [0,80],\ c \in [0,9],\ s \in [0,1] \quad (1)$$

determines the value of an individual icon from its attributes, which then becomes one of the first ten base-1620 "places" in our positional encoding, as shown in Fig. 5. However, the key is not solely based on the randomly selected icons. Two additional components reflect t (the terrains chosen for the five regions) and r (the placement of the ten icons into the regions), as shown in Fig. 5. For example, a particular icon placement might position the panda in the snow and the pie in the sky. These two components become the 11th and 12th places of the key, with bases of 32 and 113,400, respectively.

Icons, terrains, and placements together form a mixed-radix positional number system. Each position has a fixed base, so every configuration maps to one unique integer, producing a *continuous, gap-free sequence with no collisions.*

The complete key can be computed using the formula:

$$\text{key} = ((\text{LN}_2 + 1) \cdot r) + ((\text{LN}_1 + 1) \cdot t) + \sum_{n=0}^{9} I_n \cdot 1620^n \quad (2)$$

where t encodes the terrain selection (base-32), r encodes the icon-to-region placement (base-113,400), and I_n encodes each

icon including its attributes in base-1620. LN_1 is the largest possible value obtained from the ten icon terms $(I_0 \ldots I_9)$, and LN_2 is the largest possible value obtained by combining those ten icon terms with t. See Tables 5 and 6 in Appendix A for sample values.

Since the 810 icons are divided into 10 fixed "pages" (pools) of 81 icons each, an icon's page number determines which I_n term it fills in the positional formula. For example, any icon from page 5 always contributes to I_4, regardless of where it appears in the scene layout. Using the values associated with the graphical key depicted in Fig. 3, the corresponding key (to 8 significant digits) is 3.3103241×10^{38}. For this scenario, $t = 4$ and $r = 83{,}094$.

Applying the same logic to the maximum possible values of the key components yields the highest possible key: `45176562091853602030662451199999999999`, corresponding to an entropy of 128.41 bits.

4.7 Designing for Memorability

Having defined how scenes are generated and encoded into 128-bit key values, we now describe the interface and training choices used to support reliable learning and recall. To strengthen recall of the terrains in their graphical key, participants first train on the five background regions of the scene before learning their ten icons.

Multimodal Training for Enhanced Memorability: When training users on their graphical key, we present an image of the icon against its background, accompanied by descriptive text (e.g., "small red panda in the snow") and an auditory recitation of the same description. Dual Coding Theory [17] posits that the combination of visual information (images) and verbal information (phrases) results in stronger recall than either alone. Our multimodal presentation also incorporates spatial and auditory information, engaging multiple cognitive resources [3, 64].

Rehearsal and Serial Position Effect: The repetition of words and icons in the training phases is built into the design because of the value of rehearsal in memorization [46] and to address the serial position effect [40]. In free recall, early and late items are recalled better than middle items (primacy and recency). The later items in a list are typically recalled better because they are more recently added to short-term memory, but this recency effect is short-lived and typically disappears when recall is delayed [27]. The primacy effect, however, is driven by greater rehearsal and consolidation of earlier items [50]. While the delayed testing of the secrets reduces the impact of the recency effect, the primacy effect may provide some benefit through repeated rehearsal.

5 Human Study Design

To ensure a fair and meaningful comparison between the BIP-39 and graphical key approaches, our human study is carefully designed to support strong participant recall for both methods. As detailed below, we implement deliberate strategies to promote memorization and reduce bias across the two conditions. Each participant completes training with both a graphical key and a BIP-39 key. In both conditions, secrets are randomly generated and system-assigned; participants memorize the assigned key and do not select or personalize any components.

5.1 Ethics Statement

Our memorization study received IRB approval and included informed consent procedures covering publication of code and anonymized study data (See Appendix F). Temporary records linking participant identities to study IDs were securely destroyed after study completion. We designed the study to avoid exposing new vulnerabilities or negatively affecting participants. While our work aims to improve the privacy and resilience of SSDI systems, broader societal implications of identity technologies warrant continued consideration.

5.2 Sample and Recruitment

The sample comprised 92 participants from the Rowan University community, the majority aged 18 to 24 (n = 44) and the remainder distributed across older brackets (25–34, n = 12; 35–44, n = 12; 45–54, n = 6; 55–64, n = 13; 65+, n = 5). Forty-eight identified as female, 42 as male, and two as nonbinary. Educational attainment ranged from high school or GED (n = 33, many enrolled in a bachelor's program) through associate's (n = 6), bachelor's (n = 19), master's (n = 21), and doctoral (n = 13) degrees.

Two measures bear on the visual and recall demands of the task. Most participants had normal color vision (n = 86); the remaining six had deutan or protan colorblindness of varying severity, with no tritan variation. On self-assessed memory, responses centered on average (n = 47) and somewhat-above-average (n = 32), with nine somewhat below, four well above, and none well below.

Participants were recruited using a general-purpose flyer distributed by email, social media, classroom announcements, bulletin boards, and Rowan University digital signage. Participation was voluntary, with $50 compensation for completing all sessions. Eligibility was open to staff, faculty, and community members, but because recruitment ran solely through the Rowan University channels, the sample skewed toward a university-affiliated population, making it a convenience sample subject to self-selection bias.

5.3 Experimental Setup

Our crossover study [36] incorporates the principles of randomized controlled trials [14]. As shown in Fig. 6, participants are randomly assigned to one of two *experimental*

groups: one group is tested after 3 days, and another group is tested after 7 days. Exceptions are made when participant scheduling necessitated one schedule or the other. Scheduling allowances are made for follow-up sessions, with a tolerance of ± 1 day; the two groups will be referred to as "3 day" and "7 day" to reflect the midpoint of the range.

Individuals differ in psychological, perceptual, and cognitive dimensions. The within-subjects components of the design address this by using each participant as their own baseline, so that individual differences do not confound comparisons. For the between-groups comparison (3-day vs. 7-day recall), random assignment serves the same end, distributing individual variation across conditions. While the sample's skew toward a university-affiliated population constrains how far the results generalize (§5.2), it does not bias these comparisons. To address fatigue and order effects, training order is counterbalanced across subgroups: some participants train on graphical keys first, others on BIP-39 (see Table 1).

To confirm that there were no order effects due to the type of password the participants learned first, we used the Chi-Square tests of Independence to compare the memory results between the graphical and BIP-39 groups. No significant differences are found for graphical recall, $\chi^2(1, N = 92) = 2.71, p = .100$, or for BIP-39 recall, $\chi^2(1, N = 92) = 0.01, p = .941$, regardless of whether participants trained on graphical or BIP-39 keys first.

Table 1: **Number of participants in experimental groups and subgroups.**

Experimental Group	Experimental Subgroup	Number of Participants[7]
3 day recall	Graphical, then BIP-39	20
3 day recall	BIP-39, then Graphical	25
7 day recall	Graphical, then BIP-39	20
7 day recall	BIP-39, then Graphical	27

The study is administered by five investigators, all trained on our study design and IRB procedures. To ensure consistency and minimize investigator bias and variability, all investigators follow standardized, printed protocol templates, receive the same training, and participate in several rehearsals.

We store participant and trial information in a MySQL database hosted on an AWS server. The database contains *demographic information* (see §5.2 for details), *participant-specific data* (assigned icons, assigned BIP-39 words and their order, and both quantitative and qualitative UI feedback), and *training/authentication timings* (start and end times per attempt as well as per-item timings, such as per icon or per word). In addition, we store *key generation values*, including color (10 options), size (2 options), icon name (810 options), region (5 options), terrain (10 options), and terrain combinations (32 options). Each participant is assigned a unique ID by the database, and no identifiable information is stored in the database.

[7] Group sizes differ slightly because late-stage participant attrition altered the initially overenrolled sample

5.4 Environmental Setup

All sessions are conducted in four dedicated rooms within the same controlled lab environment to ensure consistency. Up to four sessions run in parallel, with each investigator managing one participant at a time.

Participants train on identical monitors, all configured to provide an equivalent visual experience for everyone. Each workstation is equipped with flashcards, pens, and a printed list of recommended memorization strategies to guide participants during training. These techniques are suggestions rather than rules, giving participants the freedom to rely on their natural strategies while still having helpful prompts handy. To improve ecological validity, we emphasized the need for long-term recall to all participants.

To further reduce variability in training delivery, we use prerecorded instructional videos for both the graphical key and BIP-39 key training sessions. This way, every participant hears the same explanations, at the same pace, with the same visual cues regardless of investigator.

5.5 Study Timeline

Our study timeline comprises three lab visits for both the 3-day and 7-day participant groups, as shown in Fig. 6. We test recall after 3 days and 7 days, representing shorter and longer delayed retention intervals within our design. Because memory retention without rehearsal declines monotonically over time, performance at a one-week delay already constitutes a conservative stress test for feasibility; schemes that fail at 7 days are unlikely to remain usable at longer delays [25]. These recall intervals align with prior work [57,61] and enable a focused trial that validates the software, characterizes participant experience, and evaluates graphical design effectiveness and user fatigue.

The procedures for the two major groups are identical across all three visits, differing only in the number of days between visits. Detailed visit procedures are provided in Appendix E.

5.6 User Surveys

Surveys are administered through Qualtrics and collect no personally identifiable information. Selected survey results are presented in this paper, while additional responses are reserved for future analysis.

Initial Survey: The initial survey collects anonymized demographic and self-assessment information. It includes questions about age, gender, education level, relative strength of memory and recall abilities. Participants are also asked to provide

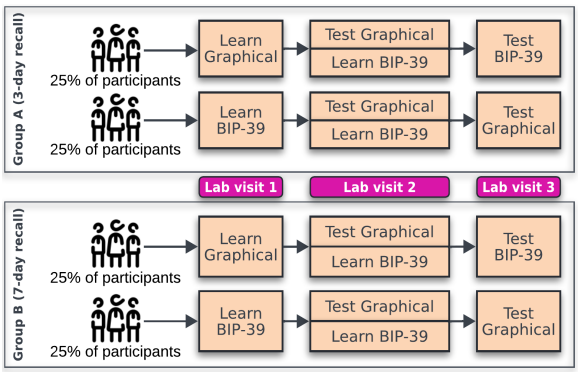


Figure 6: The plan and timeline of our crossover RCT.

results from the Pilestone [48] and Enchroma [21] online color blindness tests, which they complete during the first phase of their first visit. The summary of the initial survey can be found in §5.2.

Exit Survey: Administered during the second phase of the final visit, the exit survey includes open-ended and Likert-scale questions about participants' experiences across the three visits. Questions address memorization strategies, the most and least useful aspects of the training interface, overall experience and satisfaction, and suggestions for improvement. Analysis of selected responses is presented in §6.4.

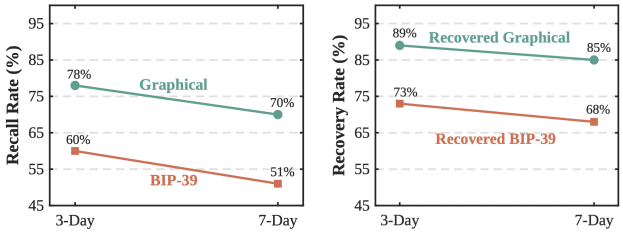
6 Results

After one week, graphical keys were forgotten substantially less often than BIP-39 mnemonics, and our combinatorial recovery approach rescued a meaningful fraction of the partially forgotten keys of both types. Notably, few participants failed to recall both key types, suggesting that graphical and BIP-39 representations may provide complementary memory cues. We organize the results below and note that each can be mapped to one or more of the three usability components defined in ISO 9241-11 [33]: effectiveness (§6.1, §6.2, and §6.4), efficiency (§6.3), and satisfaction (§6.4).

6.1 Comparative Recall Rate

We compare participant recall accuracy for BIP-39 and graphical keys across both the 3-day and 7-day groups. As shown in Fig. 7a, graphical keys consistently achieve higher recall rates than BIP-39. After three days, 78% of participants successfully authenticate using graphical keys compared to 60% for BIP-39. At seven days, graphical recall decreases slightly to 70%, while BIP-39 drops more sharply to 51%.

A McNemar's test [47], using the asymptotic chi-square approximation, indicates a significant difference in key recall between graphical and BIP-39 formats, $\chi^2(1, N = 92) = 7.31, p = .007$. Using the exact binomial method yields a slightly smaller p-value, $p = .006$. Of the 92 participants, 26 recall their graphical key but forget their BIP-39 key, compared to 9 participants who forget their graphical key but recall their BIP-39 key (Table 2). This asymmetry suggests a clear advantage for graphical keys.



(a) Pure recall (b) Recovery via search

Figure 7: **Percentage of users who successfully recalled their keys, along with the percentage of keys eventually restored through a search of partially forgotten seeds.** Recovery includes users who successfully recalled the key, as well as those whose keys were restored through the combinatorial key search.

The odds of forgetting a BIP-39 key are 2.89 times higher than forgetting a graphical key, 95% CI [1.31, 7.01]. The two failures also tend to co-occur: among participants who forgot their graphical key, 62.5% (15/24) also forgot their BIP-39 key, compared with 38.2% (26/68) of those who recalled their graphical key. When recovered keys are considered, Fig. 7b shows that the effective success rates increase to 89% for graphical keys and 73% for BIP-39 after three days, and to 85% and 68%, respectively, after seven days. The method used to recover these keys is discussed in Section 6.2.

To confirm that the $\pm$1-day scheduling tolerance did not drive this result, we re-ran the McNemar comparison using only the 41 participants tested exactly on the target day for both key types. The graphical advantage held (McNemar exact $p = .031$). Because the two keys were tested on different days in nearly all 3-day cases, this subset comprises almost entirely 7-day participants (39 of 41), so it most directly confirms robustness at the longer retention interval, where recall is lowest and return timing was held exactly constant.

Table 2: **Crossover recall success rate evaluation.**

Success Category	Number of Users
Both BIP-39 and Graphical	42
Graphical, not BIP-39	26
BIP-39, not Graphical	9
Neither BIP nor Graphical	15

6.2 Partially Forgotten Key Recovery

Our study stores data from all user authentication attempts including their responses (BIP-39 words entered or icons

identified).

To evaluate recovery from fragmented recall, we implemented a combinatorial key search algorithm to recover from the participant's failed authentication attempt. The recovery process assumes that failed authentication attempts may still contain substantial portions of the original graphical key, differing only in a limited number of attributes such as size, color, or icon selection. The algorithm systematically varies icon attributes according to a fixed search order, prioritizing lower-cost (i.e., smaller permutation space) attributes such as size and color before icon-name substitutions. Each generated key is verified using the same cryptographic authentication check as the original key. The algorithm runs until either the correct key is recovered through cryptographic verification or the search reaches a bound of 29,524,500 candidate keys,[8] corresponding to the permutations up to one color and two icon-name substitutions. Additional implementation details and pseudocode are provided in the Appendix B.

Table 3: **Successful graphical authentications via combinatorial key search. (12 users recovered)**

User	Attempt ID	Errors Made	Recovered
529	1308	3 colors	✔
	1311	1 color	✔
547	1486	1 color	✔
562	1513	1 color, 1 size	✔
571	1696	1 color	✔
574	1637	1 color, 1 size	✔
	1639	1 color	✔
586	1712	1 color	✔
	1716	1 color	✔
597	1855	2 colors, 1 size	✔
598	1980	1 color	✔
600	1952	4 colors, 1 size, 1 terrain	✘
	1953	4 colors[a]	✔
601	1890	1 color, 1 size	✔
607	1928	3 colors	✔
617	1976	1 color	✔

[a] We recovered participant 600's graphical key, *despite four recall errors*, in about 6.4 minutes. In contrast, four recall errors in a BIP-39 key could require up to 8.708×10^{15} permutations, which could take approximately 95,000 years on consumer-grade hardware.

Table 3 shows the authentication attempts of users who were unable to authenticate successfully on their own, but were recoverable. 12 users were recovered from the 24 users (50%) who failed graphical authentication. We believe that, due to the granularity of our graphical key (see Fig. 8), four mistakes in a BIP-39 key would be more difficult to recover from than four mistakes in our graphical key, as users can partially remember the correct icon.

[8] On a consumer-grade laptop (MacBook Air M4 with 16GB RAM), evaluating 29,524,500 candidates (corresponding to the number of perturbation sequences evaluated in Appendix B) required approximately 2 hours and 45 minutes.

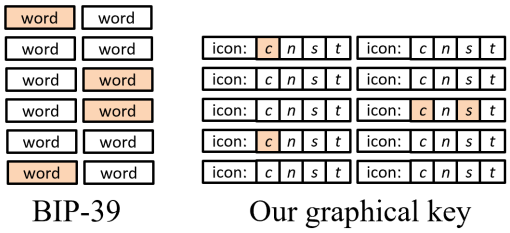


Figure 8: **Comparison of error granularity between BIP-39 and our graphical key.** Note that an error in an icon attribute is less severe than error in a BIP-39 word.

For comparison's sake, we examine the failed BIP-39 authentication attempts to see how many users could have been recovered within the same time frame as our limit for recovering graphical users, i.e., 168.5 minutes. Applying this limited recovery time to BIP-39, we theoretically recover 14 users out of the 41 users (34.1%) who failed BIP-39 authentication as shown in Table 4. These results are graphically depicted in Fig. 7b.

Table 4: **BIP-39 authentication attempts recoverable under same recovery constraints as the graphical key. (14 recoverable users)**

User	Attempt ID	Errors Made
523	1285	1 wrong word
535	1342	1 wrong word
536	1363	1 wrong word, 1 flipped order
538	1424	1 wrong word, 1 flipped order
541	1327	1 wrong word
544	1430	1 wrong word, 1 flipped order
548	1487	1 wrong word
564	1602	1 wrong word
565	1596	1 wrong word
574	1700	1 wrong word
593	1841	1 wrong word
595	1913	1 wrong word
607	1989	1 wrong word
612	1948	1 wrong word, 1 flipped order

6.3 Memorization and Recall Time

Participants required less time to memorize graphical keys (8.06 min) than BIP-39 mnemonics (12.00 min). BIP-39 keys were recalled more quickly during authentication. Fig. 9 summarizes these results; detailed statistics appear in Appendix D.

6.4 User Feedback

We evaluate the perceived satisfaction and effectiveness of our graphical and BIP-39 training interfaces, two of the usabil-

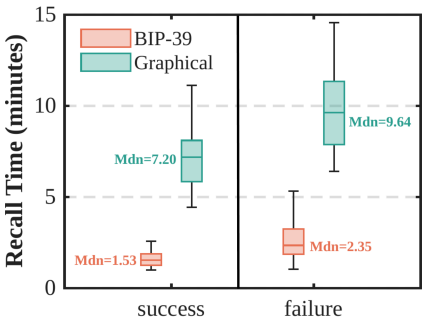


Figure 9: **Key recall times.**

ity components in ISO 9241-11 [33], with two Likert-scale items in the exit survey. We ask whether the participant is effectively able to complete their training, and whether their overall experience with the training interface is satisfactory. As shown in Fig. 10, fewer than 10% of participants report a neutral or negative user experience with either training interface. Fewer than 10% of users feel unable to complete their training using the interface. See Appendix C for some examples of qualitative feedback.

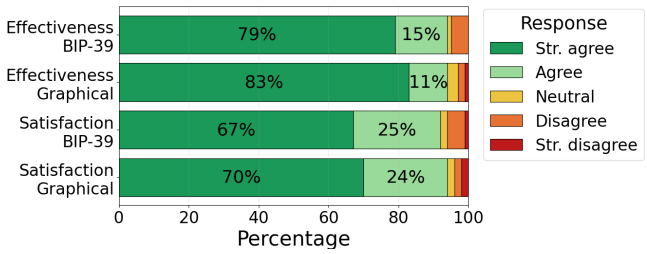


Figure 10: **Self-reported measures of effectiveness and satisfaction.**

7 Insights and Way Forward

7.1 Study Limitations and Generalizability

Our study should be interpreted as an early feasibility evaluation rather than evidence that memorized high-strength keys are deployment-ready. Participants completed training under controlled laboratory conditions with structured rehearsal, standardized instructions, and explicit expectations of future recall. Real-world private-key backup scenarios may differ substantially: users may rehearse less frequently, underestimate the consequences of forgetting, or encounter recovery needs only after long periods of nonuse.

While a week-long trial sufficed for this study (§5.5), future evaluations should examine longer retention intervals under more naturalistic conditions. Future studies should also adopt standardized usability instruments (e.g., SUS or UMUX), incorporate adaptive accessibility features (e.g., optional voice guidance), and compare cued-recall and fully recognition-based representations at equivalent entropy levels to better characterize memorability tradeoffs.

Although we collected colorblindness information and employed an accessible color palette, our participant population was limited primarily to sighted adults. Recovery mechanisms for digital identity systems should ultimately support broader populations, including users with visual impairments, color perception differences, cognitive disabilities, and age-related memory decline. Future work should investigate accessibility accommodations and alternative representations that reduce reliance on visual and memory-specific assumptions.

7.2 Future Research Directions

Building on our findings, an important direction for future work is to investigate whether graphical representations can act as retrieval cues for BIP-39 mnemonic words. By mapping each word to an icon grid, participants could associate words with unique graphical representations. This mapping reduces the candidate set from 2048 words to small, visually identifiable groups, making it easier for users to select the correct word. Our study suggests that combinatorial key search can support recovery of partially forgotten private key seeds. Integrating such recovery techniques into future investigations could further improve restoration success rates. Since key search is highly parallelizable, leveraging consumer-grade GPUs represents a promising avenue. Additionally, asking users to provide a confidence level for each recall attempt (graphical or BIP-39) would allow an AI model to prioritize certain components over others, potentially improving both speed and accuracy in reconstructing partially forgotten keys.

When comparing in-group results for users who failed to recall their key, 53.8% failed in only one category (graphical or BIP-39), suggesting that the two representations may provide complementary memory cues. This motivates a hybrid approach in which BIP-39 mnemonic memorization is enhanced with graphical cues. However, it remains uncertain how memory performance would be affected when memorizing both representations simultaneously. Further research and controlled human trials are required to test this hypothesis.

Durable memorization of strong private keys remains a future goal. However, our findings suggest that partially forgotten key material may be more recoverable than previously assumed, and that graphical cues together with combinatorial recovery may support more resilient long-term key recall.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2348344. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999. https://doi.org/10.1145/322796.322806.

[2] Ramakrishna Ayyagari. An exploratory analysis of data breaches from 2005-2011: Trends and insights. *Journal of Information Privacy and Security*, 8(2):33–56, 2012.

[3] Alan D Baddeley, Neil Thomson, and Mary Buchanan. Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, 14(6):575–589, 1975. Elsevier.

[4] Eranga Bandara, Xueping Liang, Peter Foytik, Sachin Shetty, and Kasun De Zoysa. A Blockchain and Self-Sovereign Identity Empowered Digital Identity Platform. In *2021 International Conference on Computer Communications and Networks (ICCCN)*, pages 1–7. IEEE, 2021.

[5] Elaine Barker and Allen Roginsky. Transitioning the Use of Cryptographic Algorithms and Key Lengths. Technical report, National Institute of Standards and Technology, 2019.

[6] Elaine Barker, Allen Roginsky, and Richard Davis. Recommendation for Cryptographic Key Generation. Technical Report NIST SP 800-133, National Institute of Standards and Technology (NIST), 2020.

[7] Tom Bennet. How secure is your bitcoin wallet's mnemonic seed phrase? https://bennet.org/learn/how-secure-is-your-bitcoin-wallets-mnemonic-seed-phrase/, 2025. Accessed: 2026-05-20.

[8] Robert Biddle, Sonia Chiasson, and Paul C Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM computing surveys (CSUR)*, 44(4):1–41, 2012.

[9] Jeremiah Blocki, Manuel Blum, and Anupam Datta. Naturally rehearsing passwords. In *Advances in Cryptology – ASIACRYPT 2013*, volume 8270 of *Lecture Notes in Computer Science*, pages 361–380, Berlin, Heidelberg, 2013. Springer.

[10] Jeremiah Blocki, Saranga Komanduri, Lorrie Faith Cranor, and Anupam Datta. Spaced repetition and mnemonics enable recall of multiple strong passwords. In *Proceedings of the 22nd Annual Network and Distributed System Security Symposium (NDSS Symposium 2015)*, San Diego, CA, USA, 2015.

[11] W Bodily, M Mendoza, B Hammond, and N Telfer. SeedQuest: A 3D Mnemonic Game for Key Recovery. Technical report, accessed 2025-08-02.[Online]. Available: https://github.com/reputage/seedQuest, 2023.

[12] Joseph Bonneau, Andrew Miller, Jeremy Clark, Arvind Narayanan, Joshua A Kroll, and Edward W Felten. Sok: Research perspectives and challenges for bitcoin and cryptocurrencies. In *2015 IEEE symposium on security and privacy*, pages 104–121. IEEE, 2015.

[13] Joseph Bonneau and Stuart Schechter. Towards reliable storage of 56-bit secrets in human memory. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 607–623, 2014.

[14] Laura Bothwell. *The Emergence of the Randomized Controlled Trial: Origins to 1980*. Phd thesis, Columbia University, New York, NY, 2014.

[15] Nicholas J Cepeda, Harold Pashler, Edward Vul, John T Wixted, and Doug Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, 132(3):354, 2006.

[16] Efstratios Chatzoglou, Vyron Kampourakis, Zisis Tsiatsikas, Georgios Karopoulos, and Georgios Kambourakis. Keep your memory dump shut: Unveiling data leaks in password managers. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 61–75. Springer, 2024.

[17] James M. Clark and Allan Paivio. Dual coding theory and education. *Educational Psychology Review*, 3(3):149–210, 1991.

[18] Michael Clark, Scott Ruoti, Michael Mendoza, and Kent Seamons. A Comparison of Three Approaches to Assist Users in Memorizing System-Assigned Passwords. In *Symposium on Usable Security and Privacy (USEC) 2024*. Internet Society, 2024.

[19] Rachna Dhamija and Adrian Perrig. Deja {Vu–A} user study: Using images for authentication. In *9th USENIX Security Symposium (USENIX Security 00)*, 2000.

[20] Jayesh Doolani, Matthew Wright, Rajesh Setty, and SM Taiabul Haque. Locimotion: Towards learning a strong authentication secret in a single session. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[21] EnChroma. #1 Online Color Blind Test | Test for Color Vision Deficiency. https://enchroma.com/pages/test. Accessed: 2025-05-20.

[22] K. Anders Ericsson, William G. Chase, and Steve Faloon. Acquisition of a memory skill. *Science*, 208(4448):1181–1182, 1980.

[23] Ethereum Foundation. Web3 secret storage definition. https://ethereum.org/developers/docs/data-structures-and-encoding/web3-secret-s

torage/, 2026. Ethereum developer documentation, version 3 of the keystore JSON format. Last updated April 3, 2026; accessed May 2026.

[24] Exodus. Everything you need to know about your 12-Word secret recovery phrase. https://www.exodus.com/support/en/articles/8598627-everything-you-need-to-know-about-your-12-word-secret-key, 2025. Accessed: 2026-05-20.

[25] Jerry S Fisher and Gabriel A Radvansky. Linear forgetting. *Journal of Memory and Language*, 108:104035, 2019. Elsevier.

[26] Jessica Gentles, Mason Fields, Garrett Goodman, and Suman Bhunia. Breaking the Vault: A Case Study of the 2022 LastPass Data Breach. *arXiv preprint arXiv:2502.04287*, 2025.

[27] Murray Glanzer and Anita R. Cunitz. Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, 5(4):351–360, 1966.

[28] Mark Harrower and Cynthia A Brewer. ColorBrewer. org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40(1):27–37, 2003. Taylor & Francis.

[29] Christopher G Healey. Choosing effective colours for data visualization. In *Proceedings of Seventh Annual IEEE Visualization'96*, pages 263–270. IEEE, 1996.

[30] Michael Hill, Dan Swinhoe, and John Leyden. The 15 biggest data breaches of the 21st century. https://www.csoonline.com/article/534628/the-biggest-data-breaches-of-the-21st-century.html, 2021. CSO Online, Accessed: 2026-05-20.

[31] Andrew Hollingworth. Object-Position Binding in Visual Memory for Natural Scenes and Object Arrays. *Journal of Experimental Psychology: Human Perception and Performance*, 33(1):31, 2007. American Psychological Association.

[32] Sabine Houy, Philipp Schmid, and Alexandre Bartel. Security aspects of cryptocurrency wallets—a systematic literature review. *ACM Computing Surveys*, 56(1):1–31, 2023.

[33] International Organization for Standardization. Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. ISO Standard 9241-11:2018, March 2018.

[34] Antônio Jaeger, Thuan Henrique Pedrosa Gomes Martins, João Pedro Parreira Rodrigues, Bruno Felipe Barbosa Muniz, Ana Luísa Santiago da Silveira Fonseca, and Ariel de Oliveira Gonçalves. The benefits of elaborative encoding over retrieval practice for associative learning. *Memory & Cognition*, 53(5):1592–1607, July 2025. Springer.

[35] Eric LG Legge, Christopher R Madan, Enoch T Ng, and Jeremy B Caplan. Building a memory palace in minutes: Equivalent memory performance using virtual versus conventional environments with the method of loci. *Acta psychologica*, 141(3):380–390, 2012.

[36] Chi-Yeon Lim and Junyong In. Considerations for crossover design in clinical study. *Korean journal of anesthesiology*, 74(4):293–299, 2021. Korean Society of Anesthesiologists.

[37] Lospec. Palette list: Colorblind 16. https://lospec.com/palette-list/colorblind-16, 2025. Accessed: 2025-08-07.

[38] George Mandler. Recognizing: The judgment of previous occurrence. *Psychological review*, 87(3):252, 1980.

[39] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[40] Bennet B Murdock Jr. The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5):482–488, 1962.

[41] Alexander Mühle, Andreas Grüner, Tatiana Gayvoronskaya, and Christoph Meinel. A survey on essential components of a self-sovereign identity. *Computer Science Review*, 30:80–86, 2018. Elsevier.

[42] Lily Hay Newman. Google+ API Bug Exposed Data of 52.5 Million Users. https://www.wired.com/story/google-plus-bug-52-million-users-data-exposed/, December 2018.

[43] Sean Oesch, Anuj Gautam, and Scott Ruoti. The Emperor's New Autofill Framework: A Security Analysis of Autofill on iOS and Android. In *Proceedings of the 37th Annual Computer Security Applications Conference (ACSAC)*. ACM, 2021.

[44] Sean Oesch and Scott Ruoti. That was then, this is now: A security evaluation of password generation, storage, and autofill in browser-based password managers. In *Proceedings of the 29th USENIX Conference on Security Symposium*, pages 2165–2182. USENIX Association, 2020.

[45] Marek Palatinus, Pavol Rusnak, Aaron Voisine, and Sean Bowe. Bip-0039: Mnemonic code for generating deterministic keys. https://github.com/bitcoin/bips/blob/master/bip-0039.mediawiki, 2013. Accessed: 2026-07-09.

[46] Milind Parle, Nirmal Singh, and Mani Vasudevan. REGULAR REHEARSAL HELPS IN CONSOLIDATION OF LONG TERM MEMORY. *Journal of Sports Science and Medicine*, 5(1):80–88, 2006.

[47] Matilda QR Pembury Smith and Graeme D Ruxton. Effective use of the McNemar test. *Behavioral Ecology and Sociobiology*, 74(11):133, 2020. Springer.

[48] Pilestone. Free Online Color Blind Test | Test for Color Vision Deficiency. https://pilestone.com/pages/color-blind-test. Accessed: 2025-05-20.

[49] Alex Preukschat and Drummond Reed. *Self-sovereign identity*, chapter 1, pages 3–20. Manning Publications, 2021.

[50] Dewey Rundus. Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, 89(1):63–77, 1971.

[51] Jungwon Seo, Deokyoon Ko, Suntae Kim, Vijayan Sugumaran, and Sooyong Park. Reminisce: blockchain private key generation and recovery using distinctive pictures-based personal memory. *Mathematics*, 10(12):2047, 2022.

[52] Richard Shay, Patrick Gage Kelley, Saranga Komanduri, Michelle L. Mazurek, Blase Ur, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Correct horse battery staple: exploring the usability of system-assigned passphrases. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*, pages 1–20, New York, NY, USA, 2012. ACM.

[53] David Silver, Suman Jana, Dan Boneh, Eric Chen, and Collin Jackson. Password managers: Attacks and defenses. In *23rd USENIX Security Symposium (USENIX Security)*. USENIX Association, 2014.

[54] Steven M Smith and Edward Vela. Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review*, 8(2):203–220, 2001.

[55] Reza Soltani, Uyen Trang Nguyen, and Aijun An. A Survey of Self-Sovereign Identity Ecosystem. *Security and Communication Networks*, 2021:1–26, 2021. Wiley Online Library.

[56] Mellanie V. Springer, Anthony R. McIntosh, Gordon Winocur, and Cheryl L. Grady. The relation between brain activity during memory tasks and years of education in young and older adults. *Neuropsychology*, 19(2):181–192, 2005.

[57] Elizabeth Stobert and Robert Biddle. Memory retrieval and graphical passwords. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, pages 1–14, 2013.

[58] Elizabeth Stobert, Alain Forget, Sonia Chiasson, P.C Van Oorschot, and Robert Biddle. Exploring usability effects of increasing security in click-based graphical passwords. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 79–88, 2010.

[59] Endel Tulving. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press, New York, 1972.

[60] Endel Tulving and Donald M Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):352, 1973. American Psychological Association.

[61] P. C. van Oorschot and Julie Thorpe. Exploiting predictability in click-based graphical passwords. *Journal of Computer Security*, 19:669–702, 2011. SAGE Publications Sage UK: London, England.

[62] George Vlahavas, Christos Karakasis, and Athena Vakali. MnemonicMaker: A Serious Game for Reconstructing Bitcoin Wallet Mnemonic Phrases. *International Journal of Serious Games*, 12(2):53–75, 2025.

[63] Colin Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2019.

[64] Christopher D. Wickens. The multiple resource theory and model: Some misconceptions in data interpretations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 68(1), 2024.

[65] Razieh Nokhbeh Zaeem, Manah M Khalil, Michael R Lamison, Siddhartha Pandey, and K Suzanne Barber. On the Usability of Self Sovereign Identity Solutions. *University of Texas at Austin Center for Identity, UTCID*, pages 21–02, 2021.

A Sample values for our graphical key

Table 5: **Values for icon attributes in sample icon set.**

Icon	Description	Page	*i* Value
I_0	small red panda in the snow	1	51
I_1	large green pie in the sky	2	53
I_2	small violet corkscrew in the sky	3	22
I_3	small pink picnic table in the trees	4	49
I_4	large orange roller coaster in the sand	5	53
I_5	large pink trombone in the trees	6	76
I_6	small red piano in the sand	7	78
I_7	large yellow janitor in the snow	8	24
I_8	small teal plane in the water	9	53
I_9	large purple sauna in the water	10	41

Table 6: **Sample icon set used to calculate key seed.**

Icon	Icon #	Color #	Size #	I_n value
I_0	51	1	0	1022
I_1	53	5	1	1071
I_2	22	9	0	458
I_3	49	0	0	980
I_4	53	3	1	1067
I_5	76	0	1	1521
I_6	78	1	0	1562
I_7	24	4	1	489
I_8	53	6	0	1072
I_9	41	8	1	837

B Combinatorial key search

B.1 Unsuccessful recovery attempts

Table 7 presents authentication attempts that could not be salvaged by the combinatorial recovery procedure.

Classifying mistakes relies on a user correctly identifying an icon. If a user identified an icon that is part of their key, we could verify whether they identified its color, its size, and its terrain correctly. If a user identified an icon that is not part of their key, this was recorded as a "name" error. In these cases, the error status of the other icon attributes (color, size, and terrain) is coded as indeterminate since we no longer have a uniquely identified value to look up those correct attributes. In some cases, we could *infer* where other errors might be, but not with absolute certainty. Thus when we report errors in icon components, they will be the *absolutely* determined errors. If an attempt misidentified the icon itself (name), it is possible and perhaps likely that other indeterminate errors exist.

Table 7: **Unsalvageable graphical authentications.**

User	Attempt ID	Total Errors	Errors Made
515	1396	8	3 colors, 1 name, 1 size, 3 terrains
	1397	8	4 colors, 1 name, 3 terrains
517	1294	3	1 name, 2 terrains
525	1402	7	1 color, 1 name, 5 terrains
	1404	4	1 color, 1 name, 2 terrains
527	1406	5	1 size, 4 terrains
	1407	5	1 size, 4 terrains
534	1334	7	5 colors, 1 name, 1 size
543	1481	5	4 colors, 1 size
545	1431	6	1 name, 1 size, 4 terrains
	1433	2	1 name, 1 terrain
549	1412	5	3 names, 2 terrains
550	1505	15	7 colors, 2 names, 1 size, 5 terrains
572	1698	3	1 color, 2 terrains
585	1711	5	3 colors, 2 sizes
593	1905	4	2 colors, 2 terrains
596	1912	12	5 colors, 1 name, 3 sizes, 3 terrains
610	1943	3	2 names, 1 terrain

B.2 Our recovery algorithm

As discussed in Section 6.2, our recovery approach treats a failed graphical authentication attempt as partial success rather than a complete failure. The algorithm performs a bounded combinatorial search over nearby graphical keys by varying the icon attributes according to a fixed ordering.

Table 8: Attribute perturbation identifiers used by the recovery algorithm.

Identifier	Attribute Type	Possible Choices
0	Size	2
1	Color	10
2	Icon name	81
3	Terrain	Existing terrain reassignment

Although Table 8 includes terrain reassignment as one possible attribute modification explored by the recovery algorithm, the current recovery evaluation only considered independent graphical attributes (size, color, and icon-name). Terrain assignments introduce structural dependencies between icon placements and valid regions, making their combinatorial recovery substantially more constrained. Hence, terrain-based searches were not included in the bounded recovery experiments reported in this study.

Table 8 defines the attribute identifiers used by the recovery algorithm, while Table 9 shows representative perturbation sequences explored during recovery. Sequences are evaluated using a fixed ordering based on combinatorial search

cost, with lower-cost perturbations explored before higher-cost searches as shown in the current implementation of Algorithm 1.

The number of permutations generated by each sequence is determined by the number of icon positions and the available choices for each perturbed attribute shown in Table 8. For example, as shown in Table 9, sequence `{0}` evaluates all single-size substitutions across the ten icon positions, resulting in $10 \times 2 = 20$ candidate permutations.

Algorithm 1 performs a bounded search controlled by a search budget B, which denotes the maximum number of candidate keys evaluated during recovery for a single failed authentication attempt. In our implementation, we set $B = 29{,}524{,}500$, corresponding to the largest perturbation sequence evaluated in Table 9.

Table 9: **Perturbation sequences explored by the recovery algorithm in order.**

Sequence	Description	Permutations
`{0}`	1 size	20
`{1}`	1 color	100
`{0,0}`	2 sizes	180
`{2}`	1 icon-name	810
`{0,0,0}`	3 sizes	960
`{0,1}`	1 size, 1 color	2000
`{0,0,0,0}`	4 sizes	3,360
`{1,1}`	2 colors	4,500
`{0,2}`	1 size, 1 icon-name	16,200
`{0,0,1}`	2 sizes, 1 color	18,000
`{1,2}`	1 color, 1 icon-name	81,000
`{0,1,1}`	1 size, 2 colors	90,000
`{0,0,0,1}`	3 sizes, 1 color	96,000
`{1,1,1}`	3 colors	120,000
`{0,0,2}`	2 sizes, 1 icon-name	145,800
`{2,2}`	2 icon-names	295,245
`{0,0,0,2}`	3 sizes, 1 icon-name	777,600
`{0,0,1,1}`	2 sizes, 2 colors	810,000
`{0,1,2}`	1 size, 1 color, 1 icon-name	1,620,000
`{1,1,1,1}`	4 colors	2,100,000
`{0,1,1,1}`	1 size, 3 colors	2,400,000
`{1,1,2}`	2 colors, 1 icon-name	3,645,000
`{0,2,2}`	1 size, 2 icon-names	5,904,900
`{0,0,1,2}`	2 sizes, 1 color, 1 icon-name	14,580,000
`{1,2,2}`	1 color, 2 icon-names	29,524,500

While effective for this study, the approach is not necessarily computationally optimal. Future implementations could incorporate user-confidence-aware authentication interfaces in which users indicate certainty for individual selections such as icon, color, size, or terrain. Such information could guide adaptive recovery strategies that prioritize search dynamically rather than solely on a fixed sequence ordering.

Input: Failed authentication attempt A, ordered perturbation sequence set S (Table 9), search budget B ; `// Maximum candidate evaluations`

Output: Recovered key, attempt count, perturbation sequence, or failure

$count \leftarrow 0$; `// Total evaluated candidates`
foreach *perturbation sequence* $s \in S$ **do**
  $C_s \leftarrow \text{GENERATECANDIDATES}(A, s)$;
  `// Candidate set for sequence` s
  foreach *candidate key* $k \in C_s$ **do**
    if $count \geq B$ **then**
      return failure;
    $count \leftarrow count + 1$;
    if $\text{VERIFY}(k)$ **then**
      return $(k, count, s)$;
return failure;
Procedure GenerateCandidates(A, s)**:**
$C_s \leftarrow \{A\}$; `// Initialize candidate set`
foreach *perturbation identifier* $p \in s$ **do**
  $C'_s \leftarrow \emptyset$; `// Temporary expanded candidate set`
  foreach *candidate* $c \in C_s$ **do**
    Generate all valid perturbations of type p applied to c;
    Add generated candidates to C'_s;
  $C_s \leftarrow C'_s$;
return C_s;

Algorithm 1: Bounded Graphical Key Recovery

C Examples of User Feedback

C.1 On graphical context

"...the punch bowl in the sky made me think of the big dipper."

"When remembering the images, I did not make up one big story, but I did have small stories, such as 'It makes sense for a camel to be in the sand and for the eye mask to be in the water.'"

"[I] broke the icons down into their regions to see if they had any short connections...[for the] House and Dartboard in the snow, I imagined playing darts in a house while it's snowing."

C.2 On whether rehearsal/repetition mattered

Since **all** trials extensively relied on rehearsal learning, the study has no empirical data. But user reaction is mixed:

"[The least useful part of training was] showing the items one by one. The repetition did help, but my preference is for the full image and the voiceover to go over each item."

"For memorizing graphical keys I utilized verbal repetition and mental rehearsal."

"[The least useful part of the training was] the repetition of the words by the voice..."

D Extended Timing Analysis

Participants do more than just remember keys — they *train* for them. Tables 10 and 11summarize detailed timing statistics omitted from §6.3.

Table 10: Memorization time statistics (minutes).

Metric	BIP-39	Graphical
Mean	12.00	8.06
Minimum	5.90	7.37
Maximum	26.33	19.32

Table 11: Recall time statistics (minutes).

	BIP-39		Graphical	
Metric	3-day	7-day	3-day	7-day
Mean	1.98	2.26	7.95	8.12

E Detailed Study Procedures

Initial Visit: Participants arrive at our lab for the first visit following their scheduling. The first phase of Visit 1 consists of initiation procedures, including reading and signing the informed consent form, watching educational videos, completing an initial demographic survey, and the random assignment of experimental groups and subgroups. In the second phase, subgroup 1 receives graphical key training, while subgroup 2 receives BIP-39 training. After training, participants attempt to authenticate using the key they had just learned. The session concludes with scheduling Visits 2 and 3.
Second Visit: The second visit is also comprised of two phases, but with fewer formalities. In the first phase, participants in subgroup 1 perform graphical key recall, while those in subgroup 2 perform BIP-39 recall. In the second phase, both groups watch the instructional video for the other key training modality. After that, subgroup 1 receives BIP-39 training, while subgroup 2 receives graphical key training. Thus, each participant first completes a key recall, then undergoes key training and a practice authentication with the other modality, which is later tested for recall at the final visit.
Third Visit: The third visit concludes the study and also includes two phases. In the first phase, participants in subgroup 1 perform BIP-39 recall, while those in subgroup 2 perform graphical key recall. By the end of this phase, each participant will complete one graphical training, one graphical recall, one BIP-39 training, and one BIP-39 recall, in the order predetermined by subgroup assignment and with time intervals determined by the experimental group. In the second phase, participants complete an exit survey and then receive a $50 virtual VISA card, processed through the ClinCard system.

F IRB Materials

F.1 Participant Recruitment

Population: In this study, we are focusing exclusively on sighted adult participants affiliated with Rowan University such as students, faculty members, and staff and will include all gender and ethnic backgrounds. We will gather crucial demographic information to examine potential influences on memory retention for image-based keys.

Recruitment: To recruit participants, we will utilize a multi-purpose flyer which can be emailed, posted to social media, distributed in Rowan University classrooms, posted on bulletin boards, and displayed on the university's digital signage systems. These methods of communication will provide the purpose of a study, an overview of study requirements, study eligibility criteria, the IRB approval information, the study location, and contact information for potential participants to learn more. Participation was voluntary, and those who completed all sessions received $50 in compensation. To ensure equity, recruitment will be pervasive and generic as opposed to targeted towards certain socio-economic groups. Factors such as gender, age, or education level will not be asked or identified during the recruitment phase.

Inclusion Criteria: The only inclusion criterion for participants is being of adult status; i.e., 18 years of age or older.

Exclusion Criteria: Anyone under the age of 18 will not be permitted to participate. 100 participants who will eventually be divided equally into two experimental groups will be recruited by disseminating IRB-approved flyers and posters via email or in THE university classrooms, official Rowan University online venues), placement on designated university wall areas, and through word of mouth.

F.2 Informed Consent

Informed consent will be obtained from all participants on-site immediately before the start of the experiment, ensuring they understand the study's purpose, procedures, risks, and benefits. For efficiency, we will share the informed consent sheet with the participants during the scheduling, so that they have

an opportunity to familiarize themselves with the document in advance. In a separate Demographic / Metrics Qualtrics survey, we are collecting the following information to explore the possibility of any demographic differences that may affect participants' recollection or expose weaknesses in our user interface design: age of subject; education level; a self-assessment of one's memory; results of a color blindness test such as the Pilestone test; Experience with computers.