Applied Cognitive Psychology

WILEY

RESEARCH ARTICLE

Unlocking Memories: Selection and Effectiveness of Retrieval Cues in Security Question-Like Metamemory Tasks

Yi Shao[1] | Yuan Cheng[2]

[1]Department of Psychology, Oklahoma City University, Oklahoma City, Oklahoma, USA | [2]School of Computer Science, University of Nottingham Ningbo China, Ningbo, Zhejiang, China

Correspondence: Yi Shao (yshao@okcu.edu)

Received: 5 June 2026 | **Revised:** 22 October 2026 | **Accepted:** 27 October 2026

Keywords: metacognition | metacognitive monitoring | metamemory | recall consistency | retrieval cue selection

ABSTRACT

Security questions remain a common authentication method despite newer alternatives, but frequent real-world failures raise the question of why. This study investigated how individuals evaluate and select retrieval cues for autobiographical facts when memorability is the sole consideration. Participants selected personal fact questions to use or avoid for a future memory test, answered all questions, and rated memorability and phenomenology. Recall success, consistency, and memorability judgments were evaluated after 3 and 30 days. Participants demonstrated higher recall success for selected than for neutral or rejected questions but exhibited overconfidence regarding long-term consistency for selected questions while remaining well-calibrated for others. Ease of retrieval predicted both selection and rejection; rejection was additionally driven by importance and low identity centrality. Although many reported security question failures, about half indicated relying on memory alone rather than external aids. These findings clarify why security questions fail and advance understanding of metamemory in applied contexts.

Security questions, also known as cognitive passwords (Zviran and Haga 1990), personal knowledge questions, or challenge questions, are a type of static knowledge-based authentication (KBA). They are used in online banking and internet services to verify a user's identity during password recovery. The core idea is straightforward: ask users questions to which they have previously provided answers (Zviran and Haga 1990). Two types exist: sensitive security questions, which involve confidential information such as social security numbers and account numbers, and personal security questions, which pertain to personal history and family background. The present study focused on personal security questions. During registration, users are typically presented with a set of system-generated security questions from which they select two or three. In doing so, they must balance privacy, security, and usability (Just 2004).

Selecting effective questions matters because failures in this process carry real costs: forgotten answers lock users out of accounts. Most studies on security questions have examined them from a memory perspective, focusing on individuals' ability to successfully recall their answers. Research has found that fact-based questions yield better accuracy than opinion-based questions (Haga and Zviran 1991). A study drawing on Google account recovery attempt data indicated that 40% of English-speaking users in the United States were unable to retrieve their answers (Bonneau et al. 2015). Why does such a seemingly simple task fail so often? Answering this question requires looking beyond recall itself to the metamemory processes that shape how people evaluate and select retrieval cues in the first place. Memorability, a key usability criterion, is closely tied to metamemory judgments. Although security questions have declined in use, they remain a valuable real-world case for studying how

© 2026 John Wiley & Sons Ltd.

All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies.

people judge the future memorability of autobiographical information, a phenomenon with implications beyond authentication.

Metamemory refers to an individual's awareness and understanding of their own memory capabilities and strategies (Flavell 1979). It involves the processes by which people monitor and control their memory performance (Metcalfe 2009). Individual differences in metamemory appear to be task-dependent (Kelemen et al. 2000).

In contrast to the memory-focused research on security questions, far fewer studies have approached security questions from a metamemory perspective. In one such study (Dhekane 2020), participants were asked whether they remembered the answers they had previously provided to security questions. Almost half responded "definitely yes," while fewer than 10% expressed any doubt about their memory. Research on passwords, another type of authentication credential, offers a useful starting point for investigating metamemory in security questions. Luna (2019) found that participants' behavior suggested a metacognitive belief that more memorable passwords were less secure, as participants were less likely to use memorable passwords in critical settings. Beliefs like this suggest that when a task is explicitly framed as security-related, participants' metamemory judgments may be shaped by security-specific assumptions rather than reflecting metamemory judgments on their own terms. For this reason, the present study examined memorability in isolation from the privacy and security considerations inherent to real-world authentication. We refer to the tasks used here as security question-like metamemory tasks rather than security questions.

The task of selecting security questions based on memorability has no direct counterpart in existing metamemory research, though several related concepts provide useful insights. It can be understood as a form of prospective metamemory judgment, in which individuals assess whether they will be able to retrieve specific information in the future (Schwartz 1994). One of the most widely studied prospective metamemory judgments is feeling-of-knowing (FOK), which refers to the subjective sense that one possesses information in memory even when it cannot be immediately retrieved (Schwartz et al. 2016). In a typical FOK paradigm, participants first attempt to recall a target item such as the answer to a general knowledge question, and then, for items they fail to recall, rate their confidence that they would recognize the correct answer if it were presented to them. Security question selection resembles a modified FOK procedure in which participants make judgments about all fact items rather than only unrecalled ones, as in traditional FOK research. A key distinction remains, however: whereas FOK research typically uses recognition as the focal memory task, security question selection requires predicting future recall performance.

Judgment of learning (JOL) is another widely studied prospective metacognitive judgment, referring to an individual's prediction of future memory performance when encountering to-be-learned information (Schwartz 1994). Research on JOLs has examined both immediate and delayed judgments (Rhodes and Tauber 2011). Delayed JOL tasks share similarities with the security questions context, as both involve assessing the likelihood of successful recollection after some time has elapsed since the information was first acquired. A key question in this literature concerns how accurately these judgments reflect actual memory performance, which is typically evaluated using two indicators (Rhodes and Tauber 2011; Schwartz and Efklides 2012). Relative accuracy (resolution) refers to the ability to differentiate the relative memorability of different to-be-learned materials. Absolute accuracy (calibration) refers to the extent to which percentage-based judgments align with subsequent memory performance for to-be-learned materials. Notably, delayed JOLs have exhibited moderate to high accuracy across both indicators (Nelson and Dunlosky 1991; Schwartz and Efklides 2012).

How such judgments are formed remains debated; Koriat (1997) has proposed that cue-utilization, a heuristic approach, may be involved in addition to the approach based on memory trace strength (Schwartz et al. 1997). Further, he identified three types of cues that guide metamemory judgments: intrinsic, extrinsic, and mnemonic cues. Intrinsic cues derive from characteristics of the to-be-remembered information itself. For security questions, intrinsic cues might include personal relevance (Brown et al. 1977), emotional salience (Hourihan et al. 2017), and distinctiveness of autobiographical facts. Cue effectiveness studies have found that learner-selected cues are indeed associated with better memory than unchosen cues (Tullis and Fraundorf 2022). Given that approximately 50% of users incorporate personal information into their passwords (Taylor and Garry 2019), individuals likely favor security questions with greater personal relevance when given a choice. Extrinsic cues relate to the external conditions under which the information is learned. For security questions, extrinsic cues may include the frequency of rehearsal in the past. Mnemonic cues concern the internal phenomenal experiences that accompany an individual's memory processes. In security question contexts, mnemonic cues might include the perceived ease or fluency with which autobiographical information comes to mind.

While previous prospective metamemory research provides valuable insights, security question-like metamemory tasks differ in at least four aspects. First, they rely on information already known to the individual rather than requiring the learning of new information, as in immediate JOLs research. Second, the information is typically autobiographical rather than consisting of general knowledge or semantic facts, as in previous prospective metamemory research (Rhodes and Tauber 2011). This autobiographical nature introduces unique factors shaping metamemory judgments, such as emotional intensity, personal significance, and rehearsal frequency. Meanwhile, traditional item-level approaches generally assume that item characteristics (e.g., word frequency) affect participants in similar ways, an assumption that sits uneasily with judgments about the memorability of autobiographical facts. Idiosyncratic influences on metamemory therefore warrant particular attention in this context (Undorf et al. 2022). Third, the retention intervals studied in delayed JOL research, with the longest typically spanning weeks (Rhodes and Tauber 2011), are not comparable to those involved in security question tasks, which commonly extend across years. Fourth, security questions impose unique demands on recall outcome. Unlike previous prospective metamemory research, in which accuracy can typically be verified by researchers against an objective standard, the accuracy of security questions

responses cannot be externally verified. Only the answers provided by the users initially serve as the reference against which subsequent recall is measured. Meanwhile, security questions demand strict recall consistency, in contrast to some JOL research where partial matches of verbatim or shared gist information may be scored as correct (e.g., Koriat 1997). Together, these differences highlight the distinctive characteristics of selecting effective security questions as a metamemory task and point to an area that remains underexplored in the literature.

In addition to prospective metamemory research, the study of metamemory appraisals in remembering autobiographical events represents another relevant line of inquiry. Scoboria et al. (2015) examined different types of autobiographical judgments, of which the belief in the accuracy of recollection is most pertinent to security question-like metamemory tasks. Although security questions typically involve recollecting autobiographical facts rather than autobiographical events, both processes share a reliance on this initial accuracy belief. However, the metamemory problem posed by security questions extends beyond that initial belief: it also requires believing that accurate recollection will remain stable over an extended period. Again, this additional dimension makes metamemory judgments in the context of security questions a particularly distinctive and underexplored area of investigation.

In metamemory appraisal, a particularly relevant study examined individuals' beliefs about the retention of autobiographical events (Shi et al. 2024). Participants rated items related to the reasons why certain events are remembered. Their predictions of retention were associated with rehearsal frequency, perceived importance, and emotionality, but not with "being a part of the life story" or "important to the self." Although their study focused on autobiographical events rather than autobiographical facts, the factors it identified may nonetheless apply when evaluating the memorability of personal information in security question-like metamemory tasks.

Researchers have examined the relationship between metamemory appraisals of autobiographical events and the accuracy of subsequent recollection. Memorability ratings made at the time of recording events were associated with accurate dating only over short intervals, such as a week, but not over longer periods (Thompson 1982). In the context of security question selection, individuals draw on memory traces for autobiographical facts from the distant past, which may provide a stronger basis for memorability judgments than ratings made for newly experienced events. Whether such memorability judgments, as reflected in the selection of security questions, translate into accurate subsequent recollection remains an open question warranting further investigation.

In conclusion, research on FOK, JOL, and metamemory appraisal of autobiographical events has provided a foundational understanding of the metamemory processes involved in the selection and effectiveness of retrieval cues in security question-like tasks. However, the unique context of security questions presents a novel metamemory paradigm that has not yet been examined, and this study aims to bridge that gap.

Against this background, the first research question investigates the metamemory processes involved in selecting retrieval cues for security question-like tasks. Typically, individuals select security questions they believe will support effective future recall. To develop a more complete understanding of these processes, this study also asked participants to identify questions they believed would be ineffective and should therefore be avoided. To mimic real-world scenarios, participants selected three questions they would want to use and an equal number they would not, ensuring a balanced comparison between the two deliberate choice conditions. A total of 12 questions were included to balance the need for sufficient data against the risk of cognitive overload.

The study then compared ratings of variables such as ease of retrieval, importance, emotional valence, perceived memorability, distinctiveness, rehearsal frequency, and perceived stability across selected-yes, selected-no, and neutral questions. Among these predictors, we hypothesized that ease of retrieval would be the primary driver of both question selection and rejection decisions, consistent with accessibility-based accounts of metamemory judgments (Begg et al. 1989). We also expected to replicate Shi et al.'s (2024) findings on autobiographical events, whereby importance, emotional valence, and rehearsal frequency are associated with selection decisions. The remaining predictors were treated as exploratory, as Koriat's (1997) model suggests that any of them could plausibly contribute. A second exploratory research question examines which question types (first-experience facts, favorable experience facts, and other-related facts) were more likely to be selected.

The second research question addresses the efficacy of retrieval cues by comparing memory performance across three conditions: selected-yes, selected-no, and neutral. During two delayed retrieval sessions, participants provided answers to the personal fact questions and made metamemory judgments about their confidence in future recall consistency. We predicted that selected questions would yield superior recall success relative to neutral and rejection questions, replicating the benefits of self-selected cues (Tullis and Fraundorf 2022). Because consistency has received less attention in literature, we did not hold strong directional predictions for this metric. The study further examined whether initial memorability ratings and variables such as emotional intensity correlated with subsequent memory performance. Memorability ratings were collected on a 100-point scale, permitting an assessment of calibration. For calibration, we predicted that participants would be well-calibrated overall, based on prior work showing that delayed JOLs typically show high accuracy (Nelson and Dunlosky 1991). We further expected perceived ease of retrieval to predict recall success, consistent with accessibility-based accounts (Begg et al. 1989); all remaining analyses were exploratory.

Beyond the monitoring aspect of metamemory, the study also looked at control decisions, that is, the actions individuals take in response to their own memory judgments. Participants completed a survey about their security questions practices, including whether they record their questions and answers. The relationship between their reported memory-offloading strategy and their memory and metamemory performance was then explored.

One notable design feature is that participants indicated which questions they would and would not choose before providing any answers or making any ratings. This ordering was intended to reduce the likelihood that an explicit, fully formed retrieval outcome or rating values would directly determine selection decisions. This approach is also supported by research suggesting that a preretrieval stage underlies delayed JOLs (Son and Metcalfe 2005). It also mirrors real-world security question scenarios, in which individuals must select a question before providing an answer.

1 | Method

1.1 | Participants

A power analysis using G*Power 3 (Faul et al. 2007) indicated that detecting a medium effect size ($f = 0.25$) in a repeated-measures analysis of variance (ANOVA) with three measurement points requires a sample size of 28 participants to achieve 80% statistical power. The study targeted 30 participants at the final session to ensure adequate power.

Recruitment was designed to account for anticipated attrition across sessions. Based on a previous offline study that achieved a 37% return rate for memory tests administered 1 year apart (Tullis and Finley 2021), a conservative return rate of 30% was estimated. Initially, 131 participants were recruited from the United States via Prolific. After excluding those who failed to provide matched responses to all 12 questions across the two rounds of Session 1, the final sample comprised 113 participants. Of these, 100 who completed all selected questions were invited to participate in Session 2; 62 returned for Session 2, and 35 went on to complete Session 3. Table 1 shows participant demographics across sessions.

1.2 | Measures

1.2.1 | Personal Fact Questions

There were 12 personal fact questions (see Appendix A). One-third of the questions focused on first experiences, one-third were subjective preference questions, and the remaining one-third were facts about the participant's parents. The questions spanned four themes: people, cars, places, and occupations, with three questions per theme. Question orders were randomized for each participant across the selection, recollection, and rating phases, to minimize item-level carryover, whereby a participant's response to an item in one phase might be influenced by recollection of their own prior response to that item.

1.2.2 | Memorability Factors

There were 11 rating questions (see Appendix B) covering memorability, ease of retrieval, distinctiveness, positive emotion, negative emotion, emotion intensity, past retrieval frequency (thinking and talking), importance, identity, and perceived stability. Memorability was rated on a scale from 0 to 100, while all other variables were rated on 7-point Likert scales.

1.2.3 | Security Questions Practices

The study included six questions about participants' experience with security questions (see Appendix C): experience using them and failing to recall them, habits for recording questions and answers, satisfaction with choosing and using pre-set questions, and an open-ended question soliciting tips for choosing memorable security questions. Responses to the recording habits question were scored to reflect the degree to which participants offloaded memory (0 = relying solely on their own memory, 1 = using mnemonic strategies to aid memory, 2 = using external aids to offload memory).

1.2.4 | Demographics

The questionnaire collected demographic information, including age, biological sex at birth, education level, and ethnic background, as well as a single item assessing perceived memory ability ("I am good at remembering.") on a 7-point Likert scale.

1.3 | Procedure

After obtaining Institutional Review Board approval from the first author's institution, participants were recruited through Prolific to participate in "a personal fact memory study." They were asked to complete the study independently and without distraction. The study consisted of three sessions.

TABLE 1 | Participant demographics.

	Session 1 (Immediate)	Session 2 (3-day)	Session 3 (30-day)
n	113	62	35
Age (years) (SD)	37.40 (13.20)	37.98 (13.07)	40.34 (12.78)
Female (%)	56 (50)	26 (42)	11 (31)
White (%)	78 (69)	43 (69)	27 (77)
College degree or higher (%)	96 (85)	50 (81)	24 (69)

Note: Logistic regressions were conducted to predict next-session participation status using biological sex, ethnicity minority status, and highest education level as separate predictors. Of these, only education level was significant: Session 2 participants with higher education were less likely to participate in Session 3, $B = -0.78$, $p = 0.01$, OR = 0.46.

In Session 1, participants first reviewed a list of 12 questions and sorted them into preference categories. Specifically, they selected three questions they would like to be tested on, three questions they would not like to be tested on, and placed the remaining questions in a "no preference" category. Participants were instructed to prioritize memorability over privacy and security in order to operationalize the construct of interest while controlling potential confounds. The instructions read:

> You will be tested on a set of personal facts. Here are the questions you will be asked to recall the answers. Please review the list of questions and select three questions that you would like to be included in the future test of personal facts, and three questions that you would not like to be included. Please place all remaining questions in the 'No Preference' box. You can approach this task as if you were choosing security questions, but without worrying about privacy and security. Focus on memorability, that is, how likely you will be able to give the same answer to a given question on a later test.

Participants then answered all questions, entering "NA" for any they could not answer. The instructions stated, "Answer all questions to the best of your ability. If you don't know an answer, type 'NA'. Keep in mind that these questions will be used in future tests, similar to security questions." Next, participants rated their confidence in providing the same answers on future memory tests using a scale from 0 to 100, serving as the study's memorability measurement, and rated factors influencing memorability using 7-point Likert scales (see Appendix B). They also reported their experience with security questions, relevant demographic information, and their Prolific ID. Finally, participants answered all questions a second time, with instructions to provide the same answers as before. Session 1 lasted about 20 min. At the conclusion of Session 1, participants were not informed that future follow-up memory tests would take place. Only participants who gave consistent answers across both Session 1 timepoints and completed all selected questions were eligible for Session 2 invitation ($n = 100$).

Eligible participants were invited to return for Session 2 after a 3-day delay and Session 3 after a 30-day delay. The average actual delay was 3.18 days (range = 3–6, SD = 0.64) for Session 2 and 31.06 days (range = 26–42, SD = 3.23) for Session 3.[1] In both follow-up sessions, participants answered all questions and completed the same confidence rating as in Session 1. Each follow-up session lasted about 5 min.

Responses from Session 2 and 3 were compared to participants' original answers from Session 1. For questions that had been answered in Session 1, a response was scored as a recall success if it exactly matched the original answer. Regardless of whether the question had initially been answered, all responses were also scored for consistency, defined as being identical to the Session 1 response. For example, an "NA" in both Session 1 and Session 2 would receive a consistency score but not a recall success score. The percentages of recall successes and consistent responses, along with mean confidence ratings, were then calculated for each question type (selected, rejected, and neutral). The inconsistent responses were further coded into three categories: omission errors, in which an "NA" response was given despite an answer having been provided during the first round; gist-consistent errors, in which the information was the same but contained minor differences, such as spelling errors, spacing variations, or punctuation differences; and gist-inconsistent errors, in which the information was not the same.

1.4 | Coding

Open-ended responses about tips for choosing memorable security questions were coded. Each complete response served as the coding unit and categories detailed in Table 2. Responses could be assigned to multiple categories as appropriate. The analysis focused exclusively on memorability; responses addressing only security considerations were excluded. The first author reviewed all responses to develop the coding scheme. A research assistant coded all data. A second research assistant independently coded 20% of the responses to establish inter-rater reliability, $\kappa = 0.87$.

2 | Results

The analyses consist of three sections: selecting retrieval cues, metamemory accuracy, and security question metamemory practice.

2.1 | Selecting Retrieval Cues

For 11 of the 12 questions, the neutral choice was the most common response. The only exception was the favorite location question, where 50% of participants selected it as their security question.

To address the first question regarding metamemory processes in selecting retrieval cues for security question-like tasks, a binomial mixed model fit by maximum likelihood was conducted with proportion of questions answered as the dependent variable and question group (selected, rejected, and neutral) as the within-subject variable. Participants showed differentiation across the three groups. Participants answered more questions they had indicated they would select ($M = 0.99$, SE = 0.004) than neutral questions ($M = 0.86$, SE = 0.02), OR = 16.79, and more neutral questions than those they had indicated they would avoid ($M = 0.75$, SE = 0.04), OR = 2.09, $ps < 0.001$.

A repeated-measures ANOVA examined perceived memorability ratings across the same three question types, again with Greenhouse–Geisser correction applied. Participants showed significant differences in their memorability ratings across question types, F (1.60, 179.59) = 28.75, $p < 0.001$, $\eta_p^2 = 0.20$. Bonferroni pairwise comparisons revealed the expected pattern: participants rated selected questions most memorable ($M = 93.91$, SD = 10.93), followed by neutral questions ($M = 83.87$, SD = 22.04), and rejected questions least memorable ($M = 78.63$, SD = 29.40). Both the selected versus neutral difference ($p < 0.001$) and the neutral versus rejected difference ($p = 0.02$) were statistically significant.

TABLE 2 | Categories of tips for choosing memorable security questions.

Category	Description	Example	Frequency, *n* (%)
General metamemory judgment	A broad description of easy-to-remember questions or answers without elaborating on specifics	"I usually think of things I know I will not forget"	14 (13.1%)
Preference of specific questions	Mentioning a preference for specific questions, such as names, places, or numbers without providing further metacognitive reasoning	"Choose pet names and family member names"	11 (10.3%)
Permanent	Questions with factual answers remain consistent over time	"I try to choose ones that don't say favorite because that tends to change over time"	19 (17.8%)
Singular answer	Questions with only one possible answer	"Make sure I choose questions for which there is only 1 answer"	5 (4.7%)
Personal	Questions about private, individual experiences that are not shared with others	"I use unique memories that only I would remember, which makes the answers easier to recall but harder for others to guess"	14 (13.1%)
Personal significance	Answers reflect importance, emotional weight, personal meaning, or aspects of identity	"Things that are very important to me"	23 (21.5%)
Familiar and everyday things	Questions tied to things encountered daily	"I choose things that I encounter mostly during my day to day activities"	12 (11.2%)
Simplicity	Answers are simple and brief	"I try to use single word answers"	5 (4.7%)
Mnemonic strategies	Mentioning active methods to aid recall	"I try to repeat them out loud a few times"	3 (2.8%)
Offload memory	Using external sources to offload memory (e.g., writing down, screenshotting, or using tools like password managers)	"I used a password manager"	10 (9.3%)
Consistency across accounts	Using the same questions or answers across multiple accounts	"Try to keep them all the same across accounts"	6 (5.6%)
Adding a personal twist	Mentioning personal tips to modify the answer	"I have a 'universal' root word I modify with regards to the account." "I pick questions that can be answered with a personal twist only I'd know"	11 (10.3%)
No strategy beyond good memory	Indicating no special tips, relying instead on natural memory or having no issues	"uh, I just don't have memory problems, I do not think about this at all, its not that deep"	8 (7.5%)

Note: $N = 107$.

To examine factors influencing choice ratings, a mixed-effects model was fitted with the outcome variable coded as −1 (rejected), 0 (neutral), or 1 (selected). Ten potential influential factors were included as fixed effects, while random intercepts were specified at the participant level. VIFs were generally below 3, with two exceptions: importance (VIF = 4.20) and identity (VIF = 3.42). Parameters were estimated using restricted maximum likelihood (REML). Of the 10 predictors, ease of retrieval emerged as the only statistically significant predictor (Table 3).[2] A parallel model with memorability ratings as the outcome variable identified three significant predictors: ease of retrieval, past thinking frequency, and perceived stability (Table 3).[3] All models converged without a singular fit warning, with the exceptions noted.To disentangle potential differential influences on question selection versus rejection decisions, two binomial mixed-effects models using maximum likelihood estimation were fitted to the selection and rejection data (Table 3). The rejection model, comparing rejection against neutral responses, revealed that participants were significantly more likely to reject a question when its content was perceived as important to daily life, OR = 1.21, but significantly less likely to reject it when the content was central to personal identity, OR = 0.87 or easy to retrieve, OR = 0.90. In contrast, the selection model, comparing selection against neutral responses, showed that ease of retrieval was the only significant predictor of selection likelihood, OR = 1.31.

2.2 | Memory Performance

We examined whether selection status could predict memory performance across three metrics (Figure 1): recall success, the percentage of valid answers provided, an important measure for security question-like tasks; recall consistency, the stability of

TABLE 3 | Mixed models predicting initial selection, memorability ratings, and 30-day delayed recall success and consistency.

	Selection				Rejected vs. neutral				Selected vs. neutral			
Fixed effects	***B***	**SE**	***t***	***p***	***B***	**SE**	***z***	***p***	***B***	**SE**	***z***	***p***
Ease of retrieval	**0.01**	**0.01**	**4.40**	**< 0.001**	−0.10	**0.05**	**−2.06**	**0.04**	**0.27**	**0.07**	**3.96**	**< 0.001**
Distinctiveness	−0.01	−0.01	−0.35	0.73	0.08	0.05	1.40	0.16	0.06	0.06	1.00	0.32
Positive emotion	0.02	0.02	0.97	0.33	−0.10	0.07	−1.38	0.17	< −0.01	0.08	−0.05	0.96
Negative emotion	−0.01	0.02	−0.97	0.33	0.08	0.05	1.48	0.14	0.05	0.06	0.89	0.37
Emotion intensity	−0.01	0.01	−0.48	0.63	< −0.01	0.04	−0.04	0.97	−0.03	0.04	−0.70	0.49
Past thinking frequency	−0.01	0.02	−0.79	0.43	0.06	0.06	1.11	0.27	0.02	0.07	0.30	0.77
Past talking frequency	0.02	0.02	1.38	0.17	−0.09	0.06	−1.60	0.11	< −0.01	0.06	−0.11	0.91
Importance	−0.02	0.02	−0.97	0.34	**0.19**	**0.07**	**2.53**	**0.01**	0.12	0.07	1.65	0.10
Identity	0.03	0.02	1.85	0.07	**−0.13**	**0.06**	**−2.08**	**0.04**	−0.01	0.06	−0.21	0.83
Perceived stability	0.01	0.01	0.57	0.57	0.03	0.05	0.61	0.54	0.11	0.06	1.73	0.08

	Memorability ratings				Recall success				Consistency			
Fixed effects	***B***	**SE**	***t***	***p***	***B***	**SE**	***z***	***p***	***B***	**SE**	***t***	***p***
Ease of retrieval	**5.82**	**0.46**	**12.59**	**< 0.001**	**0.50**	**0.12**	**4.01**	**< 0.001**	−0.01	0.17	−0.08	0.93
Distinctiveness	0.64	0.48	1.34	0.18	0.21	0.13	1.62	0.10	0.09	0.18	0.48	0.63
Positive emotion	−0.25	0.60	−0.43	0.67	−0.03	0.20	−0.15	0.88	**−0.74**	**0.25**	**−2.96**	**0.003**
Negative emotion	−0.78	0.50	−1.58	0.11	−0.18	0.17	−1.08	0.28	−0.14	0.22	−0.65	0.52
Emotion intensity	0.35	0.40	0.88	0.38	−0.10	0.01	−0.94	0.35	−0.05	0.12	−0.38	0.70
Past thinking frequency	**1.13**	**0.54**	**2.09**	**0.04**	0.31	0.16	1.94	0.05	**0.58**	**0.21**	**2.71**	**0.01**
Past talking frequency	−0.23	0.53	−0.43	0.67	0.03	0.15	0.20	0.85	**−0.44**	**0.21**	**−2.11**	**0.03**
Importance	1.10	0.58	1.89	0.06	−0.02	0.15	−0.11	0.91	0.03	0.18	0.15	0.88
Identity	−0.17	0.53	−0.32	0.75	−0.09	0.15	−0.62	0.54	0.05	0.18	0.27	0.79
Perceived stability	**0.93**	**0.45**	**2.07**	**0.04**	−0.02	0.14	−0.15	0.88	0.19	0.16	1.20	0.23

Note: Significant results are shown in bold. Coefficients for linear mixed-effects models are unstandardized estimates. Coefficients from binomial mixed-effects models were estimated on the log-odds scale.

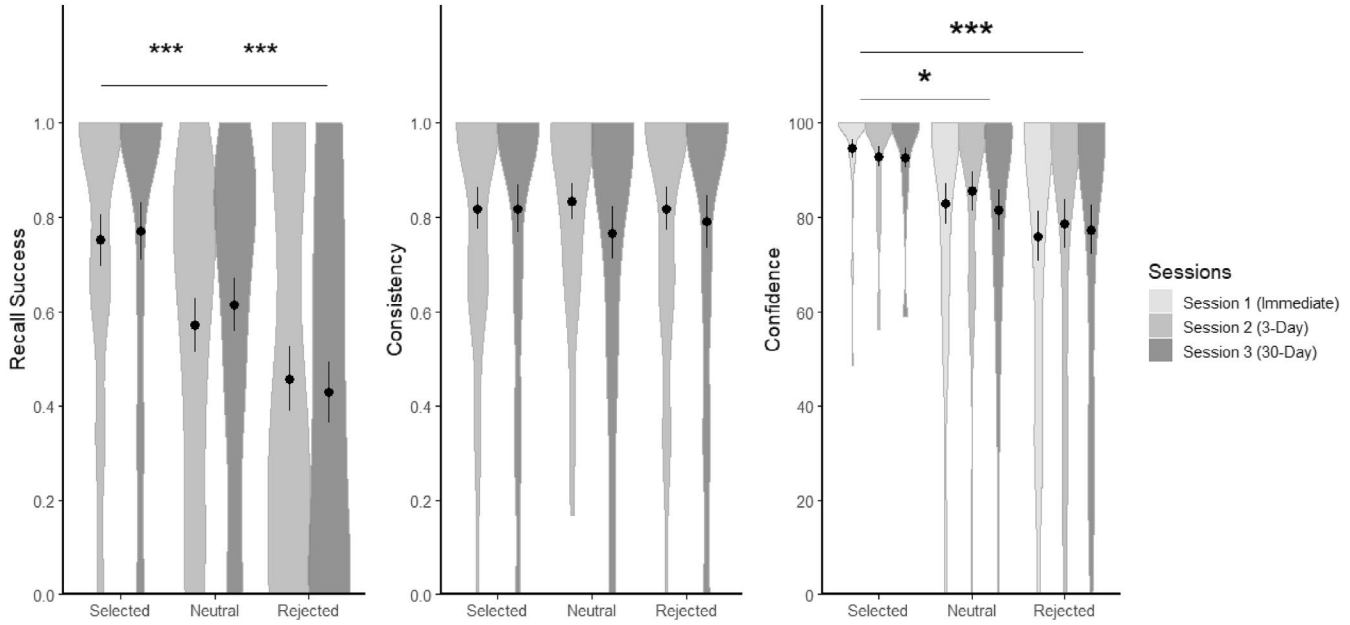


FIGURE 1 | Recall success, consistency, and confidence across selection status and sessions. *Note*. Violin plots represent the distribution of participant mean responses. Error bars represent standard errors. *$p < 0.05$, **$p < 0.01$, ***$p < 0.005$. Recall success and consistency for Session 1 were both 100%, a direct result of the inclusion criteria, and are therefore not presented.

responses over time regardless of answer validity; and memorability ratings, participants' confidence in their ability to provide consistent future responses. For each participant, we scored recall success and consistency by comparing later session responses with those from Session 1. Neither recall success nor consistency correlated with participants' self-reported memory ability, $rs < 0.14$, $ps > 0.44$.

Our first analysis addressed the relative accuracy of selection status. We first examined recall success as a function of question selection status (selected, rejected, or neutral) and delay (short vs. long) using a binomial mixed model. Question selection status was statistically significant. Participants were significantly more accurate for selected questions ($M = 0.81$, SE = 0.05) than for neutral questions ($M = 0.61$, SE = 0.06), OR = 2.81, and more accurate for neutral questions than for rejected questions ($M = 0.42$, SE = 0.07), OR = 2.19, $ps < 0.001$. However, neither the main effect of delay, $B = 0.14$, $p = 0.71$, nor the interaction between the two, $ps > 0.55$, reached statistical significance.

The second analysis used the same independent variables with response consistency (including "NA" responses) as the dependent variable (Figure 1), again using a binomial mixed model. None of the effects were statistically significant: selection status, $ps > 0.71$; delay, $B = 0.08$, $p = 0.84$; interaction, $ps > 0.26$.

For consistency, we also ran the following exploratory (nonpreregistered) analyses (Table 4). We examined whether the distribution of consistency categories (consistent, omission error, gist-consistent error, and gist-inconsistent error) differed between the 3-day and 30-day sessions using a multinomial mixed-effects model. Session was included as a fixed effect and item and participant as random intercepts. Consistent responses served as the reference category. The only significant difference was observed for gist-inconsistent errors: relative to the consistent responses, the odds of a gist-inconsistent error were significantly higher at the 30-day session than at the 3-day session, $B = 0.69$, OR = 1.99, SE = 0.25, 95% CI [0.20, 1.20].Interestingly, the binomial mixed model examining omission errors, with selection status and delay as fixed factors, revealed a statistically significant main effect of selection status. Participants had the lowest predicted probability of omission errors for selected questions, $p = 0.002$, followed by neutral questions, $p = 0.10$, and then rejected questions, $p = 0.29$. Pairwise comparisons indicated that all three conditions differed significantly from one another, $ps < 0.001$. The main effect of delay and the interaction were not statistically significant, $ps > 0.38$.

TABLE 4 | Percentages of consistency categories.

	Session 2 (3-day)	Session 3 (30-day)
Consistent	62.37	60.71
Inconsistent		
Omission errors	17.74	19.05
Gist-consistent errors	11.56	6.43
Gist-inconsistent errors	8.33	13.81

Analyses of gist-consistent errors revealed a different pattern. Only rejection status was a statistically significant predictor, $B = -1.51$, SE = 0.61, $p = 0.01$. The predicted probability of a gist-consistent error was lower for rejected questions ($p = 0.02$) than for selected questions ($p = 0.06$) or neutral questions ($p = 0.04$). Pairwise comparisons, however, indicated that these differences were not statistically significant. Analyses of gist-inconsistent error found no significant predictors, $ps > 0.14$.

TABLE 5 | Metamemory calibration biases by question selection status.

Metamemory ratings	Consistency	Selection status	*M*	SD	*t*	*p*	*d*
Session 1 (Immediate)	Session 2 (3-day)	**Selected**	**0.13**	**0.27**	**3.83**	**< 0.001**	**0.49**
		Neutral	0.05	0.33	1.21	0.23	0.15
		Rejected	0.02	0.38	0.33	0.75	0.04
Session 1 (Immediate)	Session 3 (30-day)	**Selected**	**0.15**	**0.29**	**3.02**	**0.005**	**0.51**
		Neutral	< 0.01	0.34	0.01	0.99	0
		Rejected	−0.02	0.43	−0.22	0.82	−0.04
Session 2 (3-day)	Session 3 (30-day)	**Selected**	**0.10**	**0.29**	**2.06**	**0.047**	**0.35**
		Neutral	0.08	0.35	1.35	0.19	0.23
		Rejected	−0.01	0.42	−0.19	0.85	−0.03

Note: Calibration bias = Metamemory ratings (in percentages) − Consistency. Significant results are shown in bold.

The third analysis used a repeated-measures ANOVA to examine how confidence ratings differed across question selection status and delay (immediate, short, and long; Figure 1). The main effect of selection status was statistically significant, F (2, 68) = 10.24, $p < 0.001$, $\eta_p^2 = 0.23$. Participants were more confident about providing identical answers in the future for selected questions ($M = 93.44$, SE = 1.69) than for both neutral questions ($M = 83.35$, SE = 4.01), $p = 0.02$, and rejected questions ($M = 77.33$, SE = 4.94), $p = 0.002$. The difference between the latter two was not statistically significant, $p = 0.19$. Neither the main effect of delay, F (2, 68) = 0.43, $p = 0.65$, nor the interaction, F (3.06, 103.88) = 1.27, $p = 0.29$, reached statistical significance. Goodman-Kruskal gamma correlations (G) could not be calculated for most participants because they used the same rating for all items: 33 participants (53.2%) in the short-delay session and 25 participants (71.4%) in the long-delay session. We ran a binomial mixed model with long-delay consistency as the dependent variable, first-session confidence ratings as the fixed effect, and participant as random effect. Confidence ratings did not significantly predict consistency, $B = -0.004$, SE = 0.01, $z = -0.52$, $p = 0.60$.

Because of study attrition, we also ran a parallel set of analyses on short-delay session results to determine whether the findings replicate. Recall success differed significantly across selection statuses. Participants demonstrated lower recall success for both neutral questions ($M = 0.59$, SE = 0.05, OR = 0.33) and rejected questions ($M = 0.54$, SE = 0.06, OR = 0.27), $ps < 0.001$, compared to selected questions ($M = 0.81$, SE = 0.04). Consistency in short-delay session did not differ by selection status (selected: $M = 0.88$, SE = 0.03; neutral: $M = 0.87$, SE = 0.03; rejected: $M = 0.88$, SE = 0.03), $ps > 0.98$. For confidence ratings, the main effect of selection status was again statistically significant, F (2, 122) = 12.46, $p < 0.001$, $\eta_p^2 = 0.17$. Participants expressed greater confidence in providing identical answers for selected questions ($M = 94.90$, SE = 1.16) than for neutral questions ($M = 87.86$, SE = 2.42), $p = 0.005$, and rejected questions ($M = 83.58$, SE = 3.02), $p < 0.001$. The difference between the latter two was not statistically significant, $p = 0.13$. Neither the main effect of delay (immediate vs. short), F (1, 61) = 0.01, $p = 0.93$, nor the interaction, F (2, 122) = 0.81, $p = 0.45$, reached statistical significance. To examine whether participation in a different number of rounds influenced results, we repeated all analyses using long-delay session participation status as the between-subjects variable. Neither the main effects nor the interactions involving participation status reached statistical significance ($ps > 0.08$[4]), and the conclusions reported above remained unchanged.

To examine the absolute accuracy of metamemory judgments, we calculated bias scores by subtracting consistency scores from memorability ratings (in percentages) for each question selection status (Table 5). Positive bias scores indicate overconfidence, whereas negative scores indicate underconfidence. One-sample *t*-tests comparing these bias scores against perfect calibration (zero bias) revealed that calibrations did not differ significantly from perfect calibration for either neutral or rejected questions. In contrast, participants showed significant overconfidence for selected questions. This pattern emerged consistently across all combinations of memorability rating sources and consistency sources.Additional binomial mixed-effects analyses examined associations between phenomenological ratings and 30-day memory performance outcomes (Table 5). Participant was included as random intercepts, and the 10 phenomenological ratings served as fixed effects in two separate models: one predicting recall success and one predicting consistency. Ease of retrieval was the only statistically significant predictor of recall success, OR = 1.65. For consistency, three variables emerged as statistically significant predictors: positivity, OR = 0.48, past thinking frequency, OR = 1.78, and past talking frequency, OR = 0.65. While past thinking frequency was associated with higher consistency, both positivity and past talking frequency showed the opposite pattern. Higher levels of each were associated with lower consistency.

2.3 | Security Question Metamemory Practice

Of the 113 participants, two reported no prior experience with security questions and were excluded from further analysis. Among the remaining 111 participants, 57 (51.4%) reported having failed to remember answers to security questions. We ran one-sample *t*-tests against the scale midpoint of 4 ("neither agree nor disagree"), testing whether responses differed significantly from neutrality. Participants were generally satisfied with

pre-set security questions for both selection, $M = 5.56$, SD = 1.38, $t(110) = 11.90$, $p < 0.001$, $d = 1.13$, and use, $M = 5.35$, SD = 1.43, $t(110) = 9.95$, $p < 0.001$, $d = 0.94$, with mean scores indicating "somewhat agree" levels of satisfaction.

Regarding storage strategies, participants most commonly relied solely on memory ($n = 56$, 50.5%) to store their security questions and answers, χ^2 (2, $N = 111$) = 14.65, $p = 0.001$. Twenty-seven (24.3%) reported using mnemonic strategies and 28 (25.2%) reported offloading memory to external sources.

Content analysis of open-ended question responses is presented in Table 2. Participants tended to select questions perceived as personally significant, permanent, and personal in nature. A combined 20% of participants were unable to articulate specific selection strategies, instead offering vague justifications such as "easy-to-remember" or acknowledging no clear rationale.

We performed mixed-ANOVAs to examine the effect of memory offloading strategy on memory performance outcomes (recall success and consistency) at the long-delay session, with selection status as the within-subject variable and offloading strategy as the between-subject variable. We found no significant main effects or interactions involving offload strategy, Fs < 0.77, ps > 0.46.

3 | Discussion

The present study examined an applied problem with significant real-world consequences: how people select and evaluate autobiographical retrieval cues in a task modeled on security question selection without genuine security concerns. In a novel metamemory task, participants predicted which autobiographical facts they would or would not be able to recall at a later time. Participants also rated various characteristics of the questions, allowing us to explore how these ratings related to their metamemory judgments. Beyond investigating retrieval cue selection, we assessed the validity of participants' metamemory judgments and examined whether subjective ratings of question characteristics predicted that validity. In addition to these metamemory monitoring processes, we investigated metamemory control processes by examining participants' actual security question practices.

This study yielded three primary findings. First, participants relied primarily on retrieval ease when selecting questions, but drew on a broader set of factors that extended beyond ease of retrieval alone when rejecting them. Second, self-selected questions produced superior recall success relative to neutral and rejected questions, an advantage that did not differ significantly across the 3-day and 30-day delays. However, participants showed significant overconfidence specifically for their selected questions, while their metacognitive accuracy for neutral and rejected questions was well-calibrated. Third, despite high self-reported rates of security question failures, most participants relied solely on their memory rather than external aids. The discussion addresses each of these three research findings in turn.

3.1 | Selecting Retrieval Cues

As expected, participants answered more of their selected questions during the initial assessment and rated them as highly memorable, compared to both neutral and rejected questions. This pattern was replicated when comparing neutral and rejected questions, with neutral questions outperforming rejected ones.

Analyses of how subjective question characteristics influenced selection decisions revealed an asymmetry between selection and rejection. Ease of retrieval influenced both, but rejections were additionally modulated by daily importance, which increased rejection likelihood, and identity centrality, which decreased it. These findings extend to autobiographical fact judgments the principle, previously established for JOLs, that metamemory decisions are "multiply determined" (Koriat 1997). Such inferential processes draw upon multiple cues, with cue utilization varying by situational context. Notably, even within this single task, selection and rejection were guided by different cue configurations. Methodologically, this aligns with research by Brainerd and colleagues (Brainerd et al. 2022) in recognition memory, which demonstrated that judging a test item as "new" may yield different results than judging it as "not-old." These findings suggest that affirmative and negative judgments rely on different processes rather than representing inverse operations along a single continuum.

Consistent with previous research (Shi et al. 2024), self-relevance appeared to be neither a necessary nor a sufficient condition for item rejection. Perceived importance was associated with a higher likelihood of rejection, whereas identity-relatedness was associated with a lower likelihood of rejection. Exploratory analyses of question preferences revealed a bias against other-related facts, with three out of the four other-related questions more likely to be rejected. This bias may reflect a self-reference effect in memory and metamemory, whereby self-referential information benefits from deeper encoding and enhanced retrieval (Brown et al. 1977).

3.2 | Cue Selection and Memory Outcomes

We examined two performance metrics: recall success and consistency. With respect to recall success, analyses confirmed that participant-selected cues yielded superior performance on the security question-like tasks. This finding replicates previous research on the efficacy of self-selected cues using word materials (Tullis and Fraundorf 2022). Selected questions yield more recall successes than neutral or rejected ones across both sessions. Recall successes did not differ significantly from 3 to 30 days, suggesting that the selected autobiographical facts were relatively resistant to forgetting over this timeframe.

Examining response consistency through multiple analytic lenses, overall consistency rates alongside finer-grained error types illustrates the need to look beyond a single summary metric. Response consistency, including the handling of "NA" cases, did not vary by selection status. Although selection status significantly predicted whether a participant could *recall*

an answer, it did not predict whether they would provide a *consistent* answer. Delay interval (short vs. long) similarly did not significantly affect either recall success or consistency. Error analyses, however, revealed a more nuanced role for both delay and selection status. As more time passed, participants were increasingly likely to give responses inconsistent with the gist of the original information, relative to consistent responses. Participants were most likely to give no responses at all for rejected questions, and second most likely for neutral questions, despite having answered these same questions initially. This graded pattern suggests that participants' metamemory judgments at the time of selection may exist along a continuum: initial responses to rejected and neutral questions were less likely to reflect a unique, stable memory trace, and in some cases may have been fabricated. Gist-consistent errors, by contrast, indicate that the underlying information was successfully retrieved but reproduced in an altered form. Because rejected questions were not tied to a unique, stable answer to begin with, they also produced the fewest gist-consistent errors. These findings show the value of examining error types at a finer grain than overall consistency alone. Collectively, these findings suggest that participants' initial judgments were reliable. On the one hand, selected questions yielded more valid responses than the other categories. On the other hand, their initial memory responses largely remained stable as time elapsed.

Analyses of phenomenological predictors further clarified the distinctive mechanisms underlying these two performance metrics. Ease of retrieval uniquely predicted recall success, whereas positivity, past thinking frequency, and talking frequency predicted consistency.

Beyond performance metrics, confidence ratings revealed that participants maintained differentiated metacognitive judgments throughout the study period. Confidence ratings remained higher for selected questions than for neutral and rejected questions across all three time points, without significant decay. This pattern contrasts with the underconfidence-with-practice effect observed in JOL research (Koriat et al. 2002), showing the importance of studying this novel metamemory paradigm. This pattern aligns with recent work by Shi et al. (2024), who found that participants judged more remote autobiographical events as more likely to be retained than recent ones.[5] Because most autobiographical facts used in the current study fell into the older category identified by Shi et al. (2024), participants may have relied on a similar heuristic, that long-retained memories are more likely to persist, leading them to assign these items higher confidence ratings.

Adopting an idiographic approach (Undorf et al. 2022), the present study extended previous research (Shi et al. 2024), which examined only the metamemory judgments about autobiographical information, by additionally investigating the association between those judgments and actual recall performance. Our findings demonstrate that specific item characteristics influence both metamemory judgments and memory performance. Perceived ease of recall was associated with both types of metamemory judgments (selection decisions and memorability ratings) as well as with recall success. This finding aligns with accessibility-based accounts of metamemory judgments (Begg et al. 1989). Similarly, thinking frequency predicted higher memorability ratings and greater recall consistency. Using Koriat's (1997) framework, perceived ease of recall represents a mnemonic cue, whereas thinking frequency represents an extrinsic cue. Our results suggest that Koriat's cue utilization framework extends to this previously underexamined metamemory task. The present correlational approach, which relies on naturally occurring variation in participants' subjective ratings rather than experimental manipulation (e.g., Begg et al. 1989) or established norms (e.g., Tullis and Fraundorf 2022), provides convergent evidence for previous experimental work. Despite this difference in method, the resulting associations converge with, and extend, this prior experimental work, suggesting that these item characteristics functionally guide metamemory processes and predict memory outcomes. It also illuminates how individuals spontaneously utilize different types of cues in ecologically valid memory situations, demonstrating the real-world applicability of laboratory-derived theoretical frameworks.

Notably, thinking frequency and talking frequency diverged in their associations with consistency. While thinking frequency promoted consistency, talking frequency decreased it. The finding itself is unsurprising given robust evidence that talking about past experiences reduces memory consistency (Hirst and Echterhoff 2012). However, this divergence carries an important methodological implication. In some phenomenological characteristic research, talking frequency and thinking frequency are combined into a single item (e.g., Shi et al. 2024), potentially obscuring their different effects.

The current findings also contribute to understanding the role of emotion in memory. Personal fact questions, by their nature, likely elicit less emotional variability than autobiographical events or emotionally loaded laboratory stimuli, which probably accounts for the null effects of emotion on metamemory judgments (Yin et al. 2023). Notably, however, higher positivity ratings were still negatively associated with recall consistency. This pattern aligns with research on autobiographical experiences showing that positivity negatively predicts memory consistency (Holland and Kensinger 2012; Raw et al. 2023; cf., Kensinger and Schacter 2006), and suggests that the relationship between positive valence and consistency may be a meaningful and robust phenomenon even in paradigms with relatively muted emotional content.

It is important to note that most of the factors investigated did not significantly predict selection, confidence ratings, or actual memory performance. One possible explanation for these largely null findings is restricted range, given that our materials consisted exclusively of well-learned autobiographical facts, many drawn from those used in security questions.[6] This raises the possibility that our conclusions may not generalize well to other types of autobiographical facts.

The present study solicited metacognitive judgments in the form of probabilities, with carefully worded instructions ensuring that predictions focused specifically on memory consistency. This enabled examination of calibration, the absolute correspondence between subjective and objective performance, a measurement of metacognitive accuracy that is sometimes underemphasized due to methodological limitations (Koriat 1997). Given that delayed JOLs are typically well-calibrated (Nelson

and Dunlosky 1991), we predicted high calibration across the three types of questions. Contrary to this prediction, participants were overconfident about future memory consistency for selected questions, whereas predictions for neutral and rejected questions were well-calibrated. The accurate metacognitive monitoring observed for the latter two categories aligns with the delayed JOL research. This alignment is unsurprising given that both delayed JOL research and the current study involve a substantial interval between encoding and judgment. Freedom from interference by recently learned information, combined with contextual similarity to the actual memory test, likely both contribute to enhanced accuracy (Rhodes and Tauber 2011). The overconfidence in selected questions, however, was not so easily explained: it persisted even when memorability ratings from the short-delay session were used to predict long-delay consistency. Despite having experienced both the passage of time and the test itself, participants did not adjust their ratings accordingly. This overconfidence specific to self-selected questions, the very questions users would choose in real authentication contexts, may provide a partial explanation for the high failure rates observed in real-world security question use (Bonneau et al. 2015). Given that metacognitive accuracy is largely task-dependent (Kelemen et al. 2000), these findings highlight the value of studying this novel metamemory task and may inspire further inquiry into this underexamined aspect of metacognitive monitoring.

3.3 | Security Question Metamemory Practice

Participants' self-reported experiences revealed a notable paradox with security questions. Over half had previously forgotten security question answers, yet they expressed moderate satisfaction with pre-set options. This disconnect suggests that users may overestimate the reliability of security question systems despite their own negative experiences.

Participants' self-reported strategies for choosing memorable security questions offered insights that extended beyond their observed behavior in the study. The most endorsed strategy was "personal significance," consistent with their tendency to select questions about firsts and favorites. The next most frequent categories were "permanent" and "personal," suggesting some awareness of typical security question requirements. However, approximately one in five participants could not articulate any specific selection strategy. This reveals a discrepancy between explicit, verbalizable metacognitive beliefs and implicit, experience-driven judgments.

Regarding storage strategies, most participants relied on unaided memory, with only a minority using mnemonics or external offloading. This heavy reliance on memory, despite high reported failure rates, may reflect limited awareness of alternative approaches or, alternatively, the overconfidence observed for selected questions. This finding aligns with the negative correlation between confidence and tendency to offload cognitive demands observed in a laboratory prospective memory study (Risko and Gilbert 2016). Notably, stated offloading intentions did not correlate with performance outcomes at the long delay. This suggests either that participants did not actually use such strategies during the study or that the inherent memorability of their selected questions rendered external aids unnecessary in this experimental context. Given this apparent disconnect between metacognitive awareness and actual practice, system designers cannot rely on users' metacognitive awareness to drive better storage practices.

4 | Limitations and Future Directions

The present study has several limitations worth considering when interpreting its findings. Most notably, participant attrition across sessions poses a threat to the generalizability of the longitudinal analyses. Although the observed attrition rate is consistent with the conservative 30% return rate anticipated during recruitment planning, the substantial reduction in sample size warrants careful consideration of who remained in the study. We therefore examined whether demographic characteristics predicted return for the subsequent session. Education level emerged as the only significant predictor of Session 3 participation, with more highly educated Session 2 participants being less likely to return, possibly reflecting lower motivation to repeat a task for monetary compensation. Importantly, dropout did not appear to influence memory performance: next-session participation status was not associated with recall success and consistency or confidence ratings across selection statuses, suggesting that those who dropped out were not systematically poorer performers. Future studies with larger samples and targeted retention strategies would help clarify the robustness of these findings.

Although questions were randomized in their order of presentation across the selection, answering, and rating tasks to reduce carryover effects, this procedure cannot fully rule them out. Some participants may have recalled their earlier selection choices or answers when completing a subsequent task involving that same item. Thus, the reported effect of selection status may partly reflect consistency-seeking behavior. Future research may address this limitation by including a suspicion-check item asking participants to report their beliefs about the study's purpose, which would help identify participants who may have shown such awareness.

In addition, we asked participants to select questions before answering them, intended to reduce the likelihood that participants would explicitly draw on fully formed retrieved memories as a basis for their decisions. However, the instructions' emphasis on memorability over security and privacy may have primed some participants to retrieve the memory anyway. Future research could remove this emphasis on memorability and instead simply note that security and privacy concerns need not be a factor.

The current study also faced the challenge of verifying the truthfulness of participants' autobiographical facts, as is common in much memory research on personal experience when independent corroboration is unavailable. In the context of security questions, which inspired the current project, truthfulness is not of primary importance. Nevertheless, participants may have fabricated autobiographical facts due to privacy concerns, despite our instructions emphasizing memorability and the anonymity of the platforms used in the study. This may explain why some seemingly robust autobiographical facts did

not yield the memory performance one might expect. Future research could go beyond simple "NA" responses by allowing participants to indicate whether a lack of response reflected a failure to retrieve the information or an unwillingness to provide it.

As a first inquiry into this largely unexplored area, this quasi-experimental study yielded findings that are correlational in nature. Future research would benefit from experimental designs capable of testing and extending the relationships identified here.

5 | Conclusion

Security question failure is a persistent real-world problem rooted in fundamental processes of metamemory and autobiographical recall. Using a novel paradigm examining autobiographical fact question selection and memory, this research finds that although people can effectively select questions they will later recall, they are overconfident about the consistency of their self-selected answers, a pattern that helps explain high failure rates in authentication contexts. Despite frequent real-world failures, most users continue to rely on unaided memory rather than external aids. In sum, these findings point to practical design recommendations: systems should give users feedback about their recall performance to correct overconfidence, and should encourage secure external storage of answers (e.g., through password managers) given how widespread reliance on unaided memory remains. More broadly, the paradigm introduced here offers a valuable tool for studying how people monitor and select retrieval cues in ecologically valid contexts, with applications beyond authentication to any setting where people must predict what they will remember.

Author Contributions

Yi Shao: conceptualization, investigation, funding acquisition, writing – original draft, visualization, methodology, formal analysis, project administration, data curation, resources. **Yuan Cheng:** conceptualization, writing – review and editing, methodology.

Funding

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The study was preregistered at https://osf.io/yw68j. The data of this study are openly available at https://osf.io/wbm48. The authors have no competing interests to declare. The study was supported by Psi Chi 2023–24 Faculty Research Grant to Yi Shao. We thank Zora McKeever and Jennifer Johnson for their assistance with data cleaning and coding.

Endnotes

[1] The invitation was sent in batches rather than individually, based on the Prolific work posting time. This created situations where participants who completed Session 1 several days after the initial posting but on time for subsequent sessions, resulting in a shorter delay between sessions than planned.

[2] This model and the two binomial models in the next paragraph all converged with a singular fit warning, with the participant-level variance component estimated at 0. As this indicates that random effects contributed minimally to the model, the fixed effects estimate are unlikely to be meaningfully affected.

[3] We also ran all the models using standardized scores of potential memorability factors. The conclusions were generally the same with two exceptions: in the rejection vs. neutral comparison, ease of retrieval became nonsignificant; in the memorability ratings model, past thinking frequency and perceived stability became nonsignificant.

[4] The effect most closely approaching significance was the main effect of participation status on confidence: long-delay round participants tended to report lower confidence ($M = 85.72$, SE = 2.68) than non-participants ($M = 92.89$, SE = 3.05).

[5] We thank a reviewer for suggesting this idea.

[6] We thank the reviewer for suggesting this idea.

References

Begg, I., S. Duft, P. Lalonde, R. Melnick, and J. Sanvito. 1989. "Memory Predictions Are Based on Ease of Processing." *Journal of Memory and Language* 28, no. 5: 610–632. https://doi.org/10.1016/0749-596X(89)90016-8.

Bonneau, J., E. Bursztein, I. Caron, R. Jackson, and M. Williamson. 2015. "Secrets, Lies, and Account Recovery: Lessons From the Use of Personal Knowledge Questions at Google." In *Proceedings of the 24th International Conference on World Wide Web*, 141–150. Association for Computing Machinery (ACM). https://doi.org/10.1145/2736277.2741691.

Brainerd, C. J., D. M. Bialer, M. Chang, and P. Upadhyay. 2022. "A Fundamental Asymmetry in Human Memory: Old ≠ Not-New and New ≠ Not-Old." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 48, no. 12: 1850–1867. https://doi.org/10.1037/xlm0001101.

Brown, J., V. J. Lewis, and A. F. Monk. 1977. "Memorability, Word Frequency and Negative Recognition." *Quarterly Journal of Experimental Psychology* 29, no. 3: 461–473. https://doi.org/10.1080/14640747708400622.

Dhekane, R. 2020. "Towards a Usable Fallback Authentication Mechanism." Master's thesis, California State University. https://scholarworks.calstate.edu/concern/projects/z603r4461.

Faul, F., E. Erdfelder, A. Lang, and A. Buchner. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39: 175–191. https://doi.org/10.3758/BF03193146.

Flavell, J. H. 1979. "Metacognition and Cognitive Monitoring: A New Area of Cognitive–Developmental Inquiry." *American Psychologist* 34, no. 10: 906–911. https://doi.org/10.1037/0003-066X.34.10.906.

Haga, W. J., and M. Zviran. 1991. "Question-and-Answer Passwords: An Empirical Evaluation." *Information Systems* 16, no. 3: 335–343. https://doi.org/10.1016/0306-4379(91)90005-T.

Hirst, W., and G. Echterhoff. 2012. "Remembering in Conversations: The Social Sharing and Reshaping of Memories." *Annual Review of Psychology* 63: 55–79. https://doi.org/10.1146/annurev-psych-120710-100340.

Holland, A. C., and E. A. Kensinger. 2012. "Younger, Middle-Aged, and Older Adults' Memories for the 2008 U.S. Presidential Election." *Journal of Applied Research in Memory and Cognition* 1, no. 3: 163–170. https://doi.org/10.1016/j.jarmac.2012.06.001.

Hourihan, K. L., S. H. Fraundorf, and A. S. Benjamin. 2017. "The Influences of Valence and Arousal on Judgments of Learning and on

Recall." *Memory & Cognition* 45, no. 1: 121–136. https://doi.org/10.3758/s13421-016-0646-3.

Just, M. 2004. "Designing and Evaluating Challenge-Question Systems." *IEEE Security and Privacy* 2, no. 5: 32–39. https://doi.org/10.1109/MSP.2004.80.

Kelemen, W. L., P. J. Frost, and C. A. Weaver III. 2000. "Individual Differences in Metacognition: Evidence Against a General Metacognitive Ability." *Memory & Cognition* 28, no. 1: 92–107. https://doi.org/10.3758/BF03211579.

Kensinger, E. A., and D. L. Schacter. 2006. "When the Red Sox Shocked the Yankees: Comparing Negative and Positive Memories." *Psychonomic Bulletin & Review* 13, no. 5: 757–763. https://doi.org/10.3758/BF03193993.

Koriat, A. 1997. "Monitoring One's Own Knowledge During Study: A Cue-Utilization Approach to Judgments of Learning." *Journal of Experimental Psychology: General* 126, no. 4: 349–370. https://doi.org/10.1037/0096-3445.126.4.349.

Koriat, A., L. Sheffer, and H. Ma'ayan. 2002. "Comparing Objective and Subjective Learning Curves: Judgments of Learning Exhibit Increased Underconfidence With Practice." *Journal of Experimental Psychology: General* 131, no. 2: 147–162. https://doi.org/10.1037/0096-3445.131.2.147.

Luna, K. 2019. "If It Is Easy to Remember, Then It Is Not Secure: Metacognitive Beliefs Affect Password Selection." *Applied Cognitive Psychology* 33, no. 5: 744–758. https://doi.org/10.1002/acp.3516.

Metcalfe, J. 2009. "Metacognitive Judgments and Control of Study." *Current Directions in Psychological Science* 18, no. 3: 159–163. https://doi.org/10.1111/j.1467-8721.2009.01628.x.

Nelson, T. O., and J. Dunlosky. 1991. "When People's Judgments of Learning (JOLs) Are Extremely Accurate at Predicting Subsequent Recall: The "Delayed-JOL Effect"." *Psychological Science* 2, no. 4: 267–270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x.

Raw, J., A. Rorke, J. Ellis, K. Murayama, and M. Sakaki. 2023. "Memory of the U.K.'s 2016 EU Referendum: The Effects of Valence on the Long-Term Measures of a Public Event." *Emotion* 23, no. 1: 52–74. https://doi.org/10.1037/emo0000788.

Rhodes, M. G., and S. K. Tauber. 2011. "The Influence of Delaying Judgments of Learning on Metacognitive Accuracy: A Meta-Analytic Review." *Psychological Bulletin* 137, no. 1: 131–148. https://doi.org/10.1037/a0021705.

Risko, E. F., and S. J. Gilbert. 2016. "Cognitive Offloading." *Trends in Cognitive Sciences* 20, no. 9: 676–688. https://doi.org/10.1016/j.tics.2016.07.002.

Schwartz, B. L. 1994. "Sources of Information in Metamemory: Judgments of Learning and Feelings of Knowing." *Psychonomic Bulletin & Review* 1, no. 3: 357–375. https://doi.org/10.3758/BF03213977.

Schwartz, B. L., A. S. Benjamin, and R. A. Bjork. 1997. "The Inferential and Experiential Bases of Metamemory." *Current Directions in Psychological Science* 6, no. 5: 132–137. https://doi.org/10.1111/1467-8721.ep10772899.

Schwartz, B. L., A. Boduroglu, and A. İ. Tekcan. 2016. "Methodological Concerns: The Feeling-of-Knowing Task Affects Resolution." *Metacognition and Learning* 11, no. 3: 305–316. https://doi.org/10.1007/s11409-015-9152-4.

Schwartz, B. L., and A. Efklides. 2012. "Metamemory and Memory Efficiency: Implications for Student Learning." *Journal of Applied Research in Memory and Cognition* 1, no. 3: 145–151. https://doi.org/10.1016/j.jarmac.2012.06.002.

Scoboria, A., J. M. Talarico, and L. Pascal. 2015. "Metamemory Appraisals in Autobiographical Event Recall." *Cognition* 136: 337–349. https://doi.org/10.1016/j.cognition.2014.11.028.

Shi, L., N. R. Brown, and P. J. C. Reimer. 2024. "Exploring the Metamnemonic and Phenomenal Differences Between Transitional and Mundane Events." *Memory & Cognition* 52, no. 5: 1164–1181. https://doi.org/10.3758/s13421-024-01534-7.

Son, L. K., and J. Metcalfe. 2005. "Judgments of Learning: Evidence for a Two-Stage Process." *Memory & Cognition* 33, no. 6: 1116–1129. https://doi.org/10.3758/BF03193217.

Taylor, R. J., and M. Garry. 2019. "People Infuse Their Passwords With Autobiographical Information." *Memory* 27, no. 5: 581–591. https://doi.org/10.1080/09658211.2018.1539499.

Thompson, C. P. 1982. "Memory for Unique Personal Events: The Roommate Study." *Memory & Cognition* 10, no. 4: 324–332. https://doi.org/10.3758/BF03202424.

Tullis, J. G., and J. R. Finley. 2021. "What Characteristics Make Self-Generated Memory Cues Effective Over Time?" *Memory* 29, no. 10: 1308–1319. https://doi.org/10.1080/09658211.2021.1979585.

Tullis, J. G., and S. H. Fraundorf. 2022. "Selecting Effectively Contributes to the Mnemonic Benefits of Self-Generated Cues." *Memory & Cognition* 50, no. 4: 765–781. https://doi.org/10.3758/s13421-021-01245-3.

Undorf, M., S. Navarro-Báez, and A. Bröder. 2022. "'You Don't Know What This Means to Me'—Uncovering Idiosyncratic Influences on Metamemory Judgments." *Cognition* 222: 1–9. https://doi.org/10.1016/j.cognition.2021.105011.

Yin, Y., D. R. Shanks, B. Li, et al. 2023. "The Effects of Emotion on Judgments of Learning and Memory: A Meta-Analytic Review." *Metacognition and Learning* 18, no. 2: 425–447. https://doi.org/10.1007/s11409-023-09335-0.

Zviran, M., and W. J. Haga. 1990. "Cognitive Passwords: The Key to Easy Access Control." *Computers & Security* 9, no. 8: 723–736. https://doi.org/10.1016/0167-4048(90)90115-A.

Appendix A: Personal Fact Questions

1. What is the last name of your favorite teacher in elementary school? (favorite, person)
2. What is the last name of your first elementary teacher? (first, person)
3. What is your mother's maiden name? (parent, person)
4. What was your favorite model of car? (favorite, car)
5. What was the model of your first car? (first, car)
6. What is the most recent make of your mother's car? (parent, car)
7. What is your dream job? (favorite, occupation)
8. What was your first job? (first, occupation)
9. What is the most recent job of your father? (parent, occupation)
10. Where is your dream vacation place? (favorite, location)
11. Where did you go for the first time on vacation? (first, location)
12. Where did your parents meet? (parent, location)

Appendix B: Ratings for Each Question and Answer

1. I am able to remember the answer when given the question on a later memory test. (0–100, 0 = definitely will not remember, 100 = definitely will remember)

   1 = strongly disagree, 7 = strongly agree
2. The answer to this question came to mind easily.
3. The answer to this question is unique or distinctive.
4. The emotion when I think of this question & answer is positive.
5. The emotion when I think of this question & answer is negative.

6. The emotion when I think of this question & answer is intense.
7. I have thought about this information
8. I have talked about this information
9. This information is important to my life.
10. This information is part of my identity.
11. My answer to this question will remain the same over time.

Appendix C: Security Questions Practices

1. I have encountered security questions when setting up online accounts before. (yes/no)
2. I am generally satisfied with choosing from the available pre-set security questions. (1 = strongly disagree, 7 = strongly agree)
3. Please describe any tips or strategies you use to choose memorable security questions.
4. How do you typically store or remember your security questions and answers? [I rely solely on my memory./I rely on my memory, but I also use some mnemonic strategies./I write them down on paper or in a digital document./Other (please specify)]
5. I am generally satisfied with using the selected pre-set security questions. (1 = strongly disagree, 7 = strongly agree)
6. I have failed to remember my answers to security questions. (yes/no)