Challenging Challenge Questions

Mike Just, David Aspinall
School of Informatics
University of Edinburgh
{mike.just,david.aspinall}@ed.ac.uk

***Abstract*—To authenticate human users to systems, *challenge questions* based on personal information are often used, especially for recovery when a primary authentication credential, such as a password, is forgotten. This ought to be a trustworthy mechanism, that is both reliable and accurate: personal information should be inherently memorable and not known to others. However, concerns have been raised recently about these assumptions: for example, some commonly used questions may be based on information that is available publicly. A possible improvement, then, is to allow users to choose their own questions. Here we report on an experiment which gathered user chosen questions and a subsequent security and usability analysis of them. Our experiment itself follows a novel method which is designed to engender the trust of participants, so they participate honestly. This revealed some surprising results. Although subjects sometimes seemed aware of the need for security, they often 'missed the mark' by a wide margin; similarly, there are serious concerns over the usability of freely chosen questions with free-form answers.**

I. Introduction

Authentication mechanisms are a central point of trust in any secure system. A trustworthy authentication system is one which authenticates both accurately and reliably. These concerns are aligned, but often there is a bias towards accuracy on the part of the system owner, and reliability on the part of the user. For example, users will deliberately choose weak passwords that are easy to remember, and do this more often for low security situations. In some systems this may be acceptable, but for most applications we want to find systems that are both *secure* and *usable*.

A well-known usability problem with strong passwords is that users will forget them. The most popular way to support account recovery, when this happens, is to use *challenge questions*. When a user forgets a *memorized* password, it is hoped that he or she will recall answers to challenge questions based on personal data. The theory is that since the answers to the challenge questions are information *already known* to the user (outside the context of the authentication system), the answers should be more memorable than a password. Challenge questions are even being used to complement password authentication; in addition to a password, users might be asked for the answer to one of their questions.

Despite their ubiquity, we know surprisingly little about the security and usability of challenge question authentication. Recent studies have pointed out serious security drawbacks with common *administratively chosen* questions [12] are based on information which is may be publicly available (for example, *"What is your mother's maiden name?"* [5]). Moreover, administratively chosen questions are sometimes inapplicable so that users have difficulty in choosing answers (for example, *"What was your first pet's name?"* won't apply to those that have never owned a pet). A natural usability improvement to address these problems, also commonly implemented, is to allow *user chosen* challenge questions. This paper reports on experiments with user chosen challenge questions, to investigate whether they can provide trustworthy authentication solutions.

Specifically, our contributions are:

1) The design of a novel and trustworthy experimental method supporting the collection of realistic authentication information, whereby participants submit their challenge questions but retain, and self-assess the memorability of, their answers.
2) Evidence that free-form answers to challenge questions are difficult to recall precisely, suggesting that as currently implemented, challenge question systems may not provide an adequate *fall-back* mechanism for account recovery.
3) A security analysis based upon the collection of the actual length of answers used by users for challenge question authentication, in which we determined that, with an average answer length of less than 8 characters, authentication systems that rely upon only a single question are highly vulnerable to brute force attack.
4) Evidence of the disparity between a user's confidence in the security provided by their challenge questions and answers, and the actual risk they face from attackers, by determining that while almost 90% of participants viewed little risk from an attacker, our simple analysis demonstrates that almost all questions are at risk to a brute force attack.

Besides the needs of a trustworthy authentication mechanism, performing experiments with authentication data requires a delicate balance of trust. To analyse accurately, we want to find out actual authentication information used; but to get realistic details, we want experimental subjects to trust us that it is safe for them to give their information. Balancing this is the principal idea behind our experimental method.

Our experimental method and the security and usability analyses are explained in Section III. Before then, Section II, provides some further background and motivation for our work. Section IV provides the results of our analysis of

Trust 2009: Socio-Economic Strand, 6-8 April 2009, Oxford, UK.

the experimental data and in Section V we provide some concluding remarks.

II. Background

Early work [7], [11], [14], [18] investigated the security and usability of "word associations" (such as the question-answer pair of a challenge question system) through experiments with small groups of users. These results demonstrated a reasonably high level of recall after a period of time, but not with 100% accuracy. The experiments also included a stage whereby close family or friends were asked whether they could guess the answers, revealing success rates of just under 50% in some cases. However, these experiments

- Used large numbers of word associations (e.g. 20), so that the results are difficult to compare directly to today's challenge question systems (that have at most 5 such associations between questions and answers), and
- They relied upon the success rates for *close* family and friends. For the purposes of our security analysis, we were curious to discover the possibilities of success for complete strangers that have very limited information about a user.

More recently, Just [8], [9] introduced security and usability criteria for challenge question design, though the framework wasn't rigorously applied to an actual system. Rabkin [12] discovered significant numbers of questions were either insecure or difficult when he analyzed the security and usability of *administratively chosen* challenge questions from 20 online banking sites. However, Rabkin did not examine *user chosen* challenge questions, nor did he perform experiments with actual users in order to analyzed the security and usability of challenge questions.

It is worth noting as well some upcoming work in this area by Schechter et al. [13] and Just and Aspinall [10] that complement the work in this paper.

III. Our Research Methods

Our research focused on the security and usability of user-chosen challenge questions (as opposed to administratively generated ones). We wanted to determine if users could choose questions that were secure and memorable. To do this, we needed to first collect, then analyze some user data.

As an authentication technique, challenge questions offer an interesting property that proves useful in designing trustworthy yet ethical experiments. Specifically, what most people casually refer to as a "challenge question," consists of a pair of items, namely the question and its corresponding answer. For most, if not all, scenarios the question should be assumed to be *public* information[1], while the corresponding answer is *private*. It is this *public-private* pairing that contributes to the design of our experiment, and aids our security and usability analysis.

[1]Only knowledge of a user name is generally required to retrieve questions.

A. Our Experimental Method

Experiments which attempt to collect sensitive user information can raise ethical concerns. Though, so long as users are properly informed and consent, such concerns can be mitigated. And perhaps surprisingly, research has shown users are quite willing to participate regardless. For example, Grosslags and Acquisti [6] discovered a dramatically higher willingness to accept money for personal information versus that of paying to protect the information. And beyond payment considerations, Spiekerman et al. [15] discovered a "surprising readiness to reveal private and even highly personal information," despite expressing views toward protecting such information. This was further demonstrated again by Berendt et al. [1] when experiments demonstrated that once in an online interaction, users often do not monitor and control their actions strongly, and privacy statements seem to have no impact on behaviour.

In this sense, we might be satisfied with leading an authentication experiment in which we directly ask users to contribute authentication information such as challenge questions and answers. However, we felt there to be some room for improvement for at least a few reasons:

- The above examples refer to *private* information, though it is not necessarily clear that users would view all *authentication information* as being private in the same sense. And even in this case, if users perceived passwords in one way, they might not similarly view challenge questions and answers as such (especially since this information is used in contexts external to the authentication system as well).
- We wondered whether users who might appear to be participating honestly may, in certain scenarios, "participate" but do not necessarily contribute realistic information (espeically in cases where participants are paid as an incentive to provide *some* information). This might be especially true for authentication information in which a direct exploitation could be recognizable to wary users.
- Since our focus was on challenge questions that are chosen by the participating user, users have an additional degree of freedom so that they might choose questions that avoid private or sensitive answers, especially if their answers were going to be revealed to us.
- We further challenged ourselves as to whether it was possible to analyze authentication information, but without having to see the information. Such an experiment would offer protection of the user's information recognizing that, despite our best intentions, even information collected by experimenters is susceptible to compromise.
- We want users to be accustomed to not revealing sensitive information, like the answers to challenge questions. It is normal practice for organizations, such as banks, to tell their users to never reveal such information to anyone. As researchers, we wanted to design an experiment that was consistent with this advice.

In this way, we wanted an experiment in which participating users would trust us enough to contribute realistic authentication information,

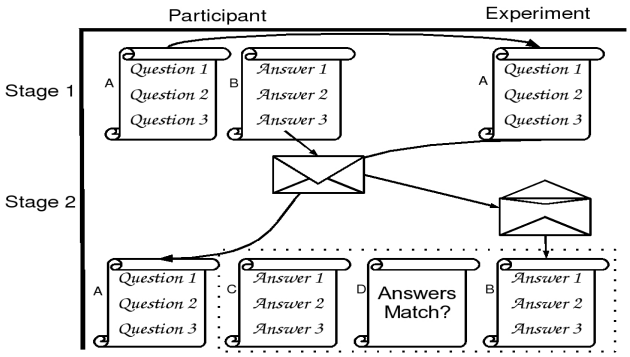

Fig. 1. Experimental Method

Our resultant experiments consisted of two stages, both of which are performed as a "pen-and-paper" exercise. The *realism* for the experiments comes not from having a perfectly simulated environment, but from allowing the participant to have greater trust and confidence in the experiment. With the pen-and-paper exercise, participants submit their challenge questions to us, but NOT their answers. The "paper answers" are retained by participants and used during a self-assessment of their memorability in the second stage. Our hope is that this allows and encourages participants to provide more realistic questions and answers. Of course, this balancing act can be quite tricky; Tsai et al.[17] note that the more emphasis placed upon privacy, the less people are willing to share. Our hope is that participants recognize the benefit of retaining their answers, but still provide us with realistic questions. Our process is depicted in Figure 1 and described as follows.

In STAGE 1 of the experiment, participants are given two sheets of paper. On Sheet A, they write their questions. On Sheet B, they write the corresponding answers. Sheet A is returned to us, while the participants retain Sheet B (in a sealed envelope). After a reasonable interval of time, participants were asked to return for the next stage, and to bring along the sealed envelope containing Sheet B. In STAGE 2, participants were returned their Sheet A, but also given two new sheets. Participants were asked to reproduce the answers to their questions on Sheet C. Then, upon opening the envelopes containing Sheet B, they performed a *self-assessment of memorability* by answering our survey on Sheet D.

B. Security Analysis

For our security analysis, we made use of the (public) questions only, but not the corresponding (private) answer (though we did ask participants for the length of their answers). Since the question behaves as a *cue* for the user to recall their answer, the question does provide some insight into the set of possible answers.[2] This gave us some confidence that our analysis of only the question provides sufficient insight into the security of the overall challenge question system.

[2]As noted, since the questions are effectively public, we did not believe that collecting the questions from users impacted our desire for an ethical experiment.

TABLE I
USABILITY CRITERIA

Property	Assessment Criteria
Memorable	EXACT same answer?
Repeatable	Answer mispelled?
Not memorable	Different answer?

Our security analysis consisted of two components. In addition to this analysis, we compared our assessment of the security to the perceptions of security offered by the questions for the participating users.

1) We determined the entropy (uncertainty) in the participant's answers based upon the answer length provided to us by the users.
2) We classified the challenge questions based upon a common set of types that define the space for the expected answer from the user, namely Proper Name, Place, Name, Number, Time/Date. The remaining challenge questions were classified as Ambiguous. From these types, we draw several conclusions regarding the expected difficulty for an attacker to guess the answers.

This approach is consistent with the *guessability* criteria of Just [9], and is similar to Rabkin's [12] analysis of administratively-chosen questions. (We have extended this even further in future work [10].)

C. Usability Analysis

While *participatory* experiments are well-known in the HCI community, they do not appear to have been leveraged with interdisciplinary security work; at least not to the degree that we propose here. For our usability analysis, we asked participants to self-assess their memorability of their answers. This included asking if the participants remembered their original answer *exactly*, or if not, the reasons for their error, which included spelling mistakes, complete "blank," or provided completely different answer (using criteria from Just [8], [9]).

In particular, we focused on the classification of answer memorability from Table I, where the second column gives a short-form of the question posed to participants for each of their challenge questions, in order to determine adherence to the property from column 1.

During STAGE 2 of the experiment, participants compared their answers to those from STAGE 1. Of course, self-assessment may be inaccurate; participants may misunderstand the instructions or be embarrassed by their poor performance. But we believe that this downside is outweighed by the ability for participants to choose realistic questions and answers, which might not happen if they were asked to share their answers with someone else (see results below).

IV. OUR RESULTS

We performed two experiments, each with a class of university students. Each student was asked to submit 3 challenge questions and provide related information as part of our survey. Experiment 1 consisted of 31 participants providing a total of 93 questions in STAGE 1; 17 of the participants returned for

TABLE II
SOME EXPERIMENT STATISTICS

	Experiment 1	Experiment 2
Participants (STAGE 1)	31	42
Participants (STAGE 2)	17	23
Questions	93	125

STAGE 2. Experiment 2 consisted of 42 participants providing a total of 125 questions in STAGE 1; 23 of the participants returned for STAGE 2. For both experiments, there was a 28-day wait from the end of the first stage till the start of the second stage. These results are summarized in Table II.

Of the total of 73 participants from STAGE 1, 23 (32%) indicated a "High" level of previous experience with challenge question authentication (used on more than 10 previous occasions), 31 (42%) indicated a "Medium" level (3-9 previous occasions), 10 (14%) indicated a "Low" level (1-3 previous occasions), and 9 (12%) indicated no previous experience. Therefore, our participants consisted of relatively well experienced, young University students. And while we are aware that our decision to survey students does limit our demographics and would indeed preclude us from showing *positive* implications, such as: if the answers to challenge questions were memorable to students, they would be memorable for everyone (which would not be correct), it does allow us to draw *negative* implications: if not memorable to young, intelligent students, then there is indeed a problem.

A. Trust in our Experimental Method

Participants were asked how likely it was that they would use the same questions to access a personal online account. Of the 59 total[3] responses 25 (42%) indicated "Very likely", 29 (49%) "Somewhat likely" and 5 (9%) "Not likely."

Of the 54 (92%) participants that responded with either "Very likely" or "Somewhat likely," 52 responded to a follow-up question regarding the impact of not submitting their answers; 8 (15%) participants indicated that not submitting their answers contributed "very much" to the decision to re-use their questions, with 25 (48%) indicating it contributed somewhat and 19 (37%) indicating it did not help.

These results are an encouraging response to our experimental method: more than $\frac{1}{2}$ of participants indicated that not submitting their answers contributed to their decision to re-use their questions. So we believe we have collected a good sample of "real" questions from our experiments, and discovered a novel way to perform trustworthy and ethical experiments. However, we also recognize that additional experiments should be performed with larger sample sizes (and more varied populations), and that our method should be compared to experiments where authentication information is exposed.

B. Security Results

We asked 42 participants in Experiment 2 to tell us the length of their answers. For the 125 questions submitted, the average length of the answers was 7.95 characters (median of 7). With answer characters drawn from the 26 letters of the English alphabet[4], according to Shannon [16] the entropy (uncertainty) per character is approximately 2.3 bits. For an 8-character answer, this gives 18.4 bits of entropy or, for an 8-character answer, $2^{18.4} \approx 350,000$ possibilities.[5] For comparison, an 8-character password uniformly chosen from the set containing uppercase and lowercase letters would have 45.6 bits of entropy. And Florêncio and Herley [3] have shown that even those web sites perceived as weak achieve greater than 20 bits of entropy for passwords. The implications here are then quite dramatic: reliance upon a single question-answer pair is significantly less secure than password-based authentication, even when just considering the basic, brute-force attack.

	Experiment 1	Experiment 2	Total
Number of Questions	51	125	176
	Difficulty for Stranger		
"Very difficult"	23	58	81 (46%)
"Somewhat difficult"	19	55	74 (42%)
"Not difficult at all"	9	11	20 (11%)
No response	0	1	1 (1%)
	Difficulty for Friend/Family		
"Very difficult"	8	11	19 (11%)
"Somewhat difficult"	17	45	62 (35%)
"Not difficult at all"	26	67	93 (53%)
No response	0	2	2 (1%)

TABLE III
PARTICIPANTS PERCEPTIONS OF SECURITY

At the same time, one can introduce knowledge of the questions. For the 218 questions submitted by participants, 99 (45%) asked for a 'Proper Name,' 47 (22%) for a 'Place,' 33 (15%) for a 'Name' including names of activities, manufacturers, 14 (6%) for a 'Number,' 8 (4%) for a 'Time/Date,' and 17 (8%) did not fit into the previous classifications. Of the 99 'Proper Name' questions, 24 (11% of the total) asked for 'Mother's Maiden Name.'[6] Knowledge of the likely space for answers further reduces the number of guesses an attacker would have to attempt beyond the already-limited entropy.

Given that 50% of questions had an answer of length less than 7 characters, it is interesting then to compare to their perception of the security of their answers as noted in Table III. The responses were to our question: "How difficult do you believe it would be for a stranger (friend or family member) to determine the answer to each of your questions?" For the 176 questions from both experiments, participants believed that for 88% of the questions that their answers would be at least "somewhat difficult" for a stranger to determine. (This reduces to 46% when considering the same difficulty for a friend or family member.) Based upon our earlier entropy calculations, participants significantly over-estimate the security offered by each of their challenge questions.

[3]We asked this during STAGE 2 of Experiment 1, with 17 participants, and STAGE 1 of Experiment 2 with 42 participants for a total of 59.

[4]Since challenge questions are often used when passwords are forgotten, the answers are typically normalized to remove case, spacing and punctuation.

[5]An attacker would not necessarily know the length of the answer, in which case they could simply make guesses starting from the minimum length.

[6]Interestingly, 3 asked for 'Grand Mother's Maiden Name,' perhaps reflecting knowledge of the insecurity of the former.

It is worth noting that our security analysis does not assume possession of information about the user that submitted the questions. Griffith and Jakobsson [5] provide the most compelling example of where such personal knowledge can further be used to easily determine the response to the "Mother's Maiden Name" question. And Haga and Zviran [7] demonstrated the ease with which friends and family can determine the answers to such questions. The indication here is that even without such personal knowledge, challenge questions provide very limited security on their own.

C. Usability Results

Our results relate to the 117 questions and answers evaluated for memorability during STAGE 2 of our experiments (51 questions from 17 participants of Experiment 1, and 66 questions from 22 participants of Experiment 1).

Given that our subjects were young university students (average age of approximately 22 years), we anticipated extremely high memorability results. To our surprise, this wasn't the case.

Of the 117 questions, 88 (75%) answers were recalled exactly while 21 (18%) of answers had different capitalization or punctuation (motivating normalization that is typically performed when registering answers). However, 8 (7%) of the answers were different: 3 of the answers were completely different, while 5 related to substitutions such as acronyms (instead of full expansion), "wrong type of street", "different units" for a numerical answer, and "adding a day of the week" to a date. If we focus on participants rather than questions, we see that 7 out of 39 participants provided a different answer (so, at least one incorrect answer from their set of three questions), equating to 18%.

For comparison purposes, Florêncio and Herley [3] note that users "forget passwords a lot" and go on to state that "4.28% of Yahoo users forgot their passwords over a three month period." Given that our experiment allowed 28 days between stages, and investigated the memorability of authentication information that often serves as a *back-up* for a lost password, it would suggest that there is a problem regarding the memorability of the answers to challenge questions.

V. CONCLUDING REMARKS

Our analysis suggests that the use of challenge questions alone is not sufficiently secure, so cannot form the sole basis of trustworthy authentication mechanisms. Further study is warranted to investigate improvements such as automated testing of responses to limit insecure questions with low entropy, enforcing a minimum length, and implementing lock out after a certain number of attempts. To protect against offline attacks, the answers to challenge questions must (at least) be hashed, and if a system is using multiple questions to increase their security, the answers should be combined to produce a single hash value (else an offline attack can be easily parallelized). Other lessons learned from similar studies on passwords should be respected as well, particularly those related to usability.

In addition, our results suggest that, despite the use of information *already known* to participants and presumed strong memories of our young subjects, there were still a surprising number of incorrect answers provided in the second stage of our experiment. The common requirement then for 100% recall of free-form answers does not appear to support a sufficient means of account recovery. However, further investigation of threshold solutions allowing a subset of questions to be answered correctly appears warranted (research initiated by Ellison et al. [2] and then Frykholm and Juels [4]). Even further, novel solutions for supporting more structured answers may be helpful (referred to as *controlled answers* by Just [8], [9]).

To follow this initial work, it is desirable to conduct experiments with a larger number of more diverse participants and further examine the ability of our method to obtain realistic authentication information. Because it is costly to scale a fully paper-based experiment, we have developed a combined online/paper-based method in which participants submit their questions online, but retain a paper copy of their answers. An electronic experiment and large sample would allow us, in principle, to add a control group to use data to test our hypothesis about enhanced trust. The control data might submit questions *and* answers and then we could compare the two sets of questions to look for meaningful differences. This seems rather difficult, but more importantly, even with a carefully secured implementation which only retained hashes of answers, it would take us back into the murky ethics of collecting authentication data, which we carefully avoided in the first place.

It is worth noting some upcoming work in this area by Schechter et al. [13] and the present authors [10] that complement the work reported in this paper.

Acknowledgements

Thanks to Andrea Szymkowiak and the anonymous reviewers for pointing out references on user behaviour discussed in Section III. We also acknowledge the support of the UK EPSRC, Grant No. EP/G020760/1, which funds the first author as a Visiting Research Fellow at the University of Edinburgh.

REFERENCES

[1] B. Berendt, O. Günther, S. Spiekermann, "Privacy in e-commerce: stated preferences vs. actual behavior," *Communications of the ACM*, **48(4)**, (Apr. 2005), 101-106.

[2] C. Ellison, C. Hall, R. Milbert, B. Schneier, "Protecting Secret Keys with Personal Entropy," *Journal of Future Generation Computer Systems*, 16(4), 2000, 3111-318.

[3] D. Florêncio, C. Herley, "A large-scale study of web password habits," in *Proceedings of the 16th international Conference on World Wide Web. WWW '07*. ACM, New York, NY, 657-666.

[4] N. Frykholm, A. Juels, "Error-Tolerant Password Recovery," in *Proceedings of the ACM Conference on Computer and Communications Security (CCS '01)*, ACM Press, 2001, 1-9.

[5] V. Griffith, M. Jakobsson, "Messin' with Texas, Deriving Mother's Maiden Names Using Public Records," *RSA CryptoBytes*, **8(1)**, 2007.

[6] J. Grossklags, A. Acquisti, "When 25 Cents is too much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information," *Sixth Workshop on the Economics of Information Security (WEIS 2007)*, Pittsburgh, PA, June 7-8, 2007.

[7] W. Haga and M. Zviran. Question-and-answer passwords: an empirical evaluation. Information Systems, 16(3):335-343, 1991.

[8] M. Just, "Designing and Evaluating Challenge Question Systems," in *IEEE Security & Privacy: Special Issue on Security and Usability*, *(L. Faith-Cranor, S. Garfinkel, editors)*, (2004), 32-39.

[9] M. Just, "Designing Authentication Systems with Challenge Questions," in *Designing Secure Systems that People Can Use*, *O'Reilly, L. Faith-Cranor, S. Garfinkel, editors*, (2005).

[10] M. Just, D. Aspinall, "Choosing Better Challenge Questions," in submission, February 2009. (Available at `http://homepages.inf.ed.ac.uk/mjust/`)

[11] R. Pond, J. Podd, J. Bunnell, R. Henderson, "Word Association Computer Passwords: The Effect of Formulation Techniques on Recall and Guessing Rates," *Computers and Security*, **19(7)**, (2000), 645-656.

[12] A. Rabkin. "Personal knowledge questions for fallback authentication: Security questions in the era of Facebook." in *Proceedings of the Symposium On Usability, Privacy and Security (SOUPS '08)*, (2008).

[13] S. Schechter, A. Bernheim Brush, S. Egelman, "It's no secret. Measuring the security and reliability of authentication via 'secret' questions," at *IEEE Symposium on Security and Privacy*, 17-20 May 2009.

[14] Y. Spector, J. Ginzberg, "Pass-Sentence - A New Approach to Computer Code," *Computers and Security*, **13(2)**, (1994), 145-160.

[15] S. Spiekermann, J. Grosslags, B. Berendt, "E-privacy in 2nd Generation E-Commerce: Privacy Preferences versus actual Behavior," *Proceedings of the 3rd ACM conference on Electronic Commerce*, p. 38-47, October 14-17, 2001, Tampa, Florida, USA.

[16] C. Shannon, "A mathematical theory of communication." *Bell System Technical Journal*, 1948, vol. 27, pp. 379–423.

[17] J. Tsai, S. Egelman, L. Cranor, A. Acquisti, "The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study," *The 6th Workshop on the Economics of Information Security (WEIS)*, June 2007.

[18] M. Zviran, W. Haga, "A Comparison of Password Techniques for Multilivel Authentication Mechanisms," *The Computer Journal*, **36(3)**, (1993), 227-237.