Personal Choice and Challenge Questions: A Security and Usability Assessment

Mike Just
School of Informatics
University of Edinburgh
Edinburgh, UK
mike.just@ed.ac.uk

David Aspinall
School of Informatics
University of Edinburgh
Edinburgh, UK
david.aspinall@ed.ac.uk

ABSTRACT

Challenge questions are an increasingly important part of mainstream authentication solutions, yet there are few published studies concerning their usability or security. This paper reports on an experimental investigation into *user-chosen* questions. We collected questions from a large cohort of students, in a way that encouraged participants to give realistic data. The questions allow us to consider possible modes of attack and to judge the relative effort needed to crack a question, according to an innovative model of the knowledge of the attacker. Using this model, we found that many participants were likely to have chosen questions with low entropy answers, yet they believed that their challenge questions would resist attacks from a stranger. Though by asking multiple questions, we are able to show a marked improvement in security for most users. In a second stage of our experiment, we applied existing metrics to measure the usability of the questions and answers. Despite having youthful memories and choosing their own questions, users made errors more frequently than desirable.

Categories and Subject Descriptors

K.4.4 [**Computers and Society**]: Electronic Commerce-Security

General Terms

Security, human factors

Keywords

Authentication, Challenge Questions, Security, Usability

1. INTRODUCTION

Challenge questions are an increasingly important part of mainstream authentication solutions, most often used as a secondary mechanism to retrieve lost primary credentials. The assumption is that querying for *already known* information may be more usable than querying *specifically memorised* information such as passwords.

The improved *memorability* is undeniable when the question is a stereotypical one such as *"What is your mother's maiden name?"*. However, further questions may suffer from usability problems of *applicability* or *repeatability* [10]. For example, questions such as *"What is your pet's name?"* fail to apply to users who have no love of animals; questions such as *"What was your first address after leaving home?"* may have formatting requirements in the answer that the respondent forgets, such as whether the answer is "115/6 Slateford Rd" or "Flat 6, 115 Slateford Road".

To address applicability, at least, some authentication solutions allow *user-chosen* challenge questions, inviting the user to choose their own questions to extend or replace administratively chosen ones. It has also been suggested that this allows users to choose more secure questions for which it is difficult to find the answers in publicly available sources. But we don't know whether users should really be trusted, without further assistance, to generate their own challenge questions. Will they really choose questions which have good usability? And what kind of security can we expect from a set of user-chosen challenge questions?

So far, there is a lack of published studies concerning challenge questions. So we have designed a series of experiments to gather data in a novel way. Our first prototype experiments were conducted by pen-and-paper on small groups of students taking degrees in Informatics-related disciplines (these are summarised below, but described in detail elsewhere [11]). The main follow-on experiment described in this paper was conducted on a larger cohort of Biology students.

The indications from our studies so far are that:

1. if allowed to freely generate their own questions, users will generate questions that are not sufficiently secure;
2. despite choosing their own questions, users still have difficulty in recalling their answers with acceptable accuracy;
3. in general, answers to challenge questions possess very limited entropy, and thus are at risk of attack if used as a primary means of authentication.

The last point should be of critical concern, since some authentication solutions allow immediate fallback authentication via challenge questions when passwords are forgotten. Our results confirm the suspicions of online commentators posted after high profile breakins over recent years

Symposium on Usable Privacy and Security (SOUPS) 2009, July 15–17, 2009, Mountain View, CA USA
.

(see e.g., [21]), that password reset mechanisms need to be carefully constructed. This is especially important since contrary to typical user expectations, an attacker may do surprisingly well even with no personal observation of the user. Though from our analysis, we have also determined that significant gains can be made by using multiple questions

In addition to the results of our experiments, the contributions of this paper include

- A security model that defines the different methods of attack given the levels of knowledge that an attacker may acquire, and considers the space of possible answers to the challenge questions.
- The design of an hybrid online-offline experimental method for analyzing challenge questions and answers that does not require participants to divulge their answers.

Evaluation by participants suggest that our experimental method contributed to the level of *realistic* authentication information received. Our security model allowed us to perform a systematic analysis of our experimental data, and also allowed us to derive some guidance towards more secure design of such systems. We believe that the model is simple, yet robust enough to allow system designers to more accurately assess the security of their authentication systems - something that has till now been lacking. Though we also recognize that our security model is but a first step, and we hope that future research work will refine and improve this model.

The rest of this paper is organised as follows. In Section 2 we provide some background on related challenge question research. Section 3 reviews our experimental method and the impact it had on participants' willingness to contribute realistic authentication information for our experiment. Sections 4 and 5 respectively review our security and usability models, and the corresponding analysis of our experimental data. In Section 6 we provide further discussion on our results, including some recommendations toward improving the security and usability of existing challenge question systems. We finish with some concluding remarks in Section 7.

2. BACKGROUND

Early work in the area of challenge questions recognized the potential advantages of using more memorable items for successful user authentication, but this work has generally been sporadic and never brought together in a complete model.

Several articles have focused on determining the usability of so-called cognitive and associative passwords [7, 16, 20, 25]. Haga and Zviran [7, 25] examined the memorability of such information with several small groups of users, and while their results report reasonably high levels of recall, their data reveals that few users were able to produce their answers with 100% recall. They also performed early tests to determine the ability of family or friends to determine a user's answers and while they wouldn't necessarily provide the answers with perfect accuracy, they recorded success rates of just under 50% in some cases. More recent results in this area, with less familiar relationships between users, have recently been published by Schechter et al. [18].

Ellison et al. [3] and Frykholm and Juels [5] both describe cryptographic techniques for tolerating errors on behalf of users that involve allowing a subset of questions to be accepted, though not for accepting answers that weren't 100% accurate.

Asgharpour et al. [1] adopt the novel approach of using browser history in order to identify users, though the techniques can be susceptible to browser-based attacks. Jakobsson et al. [8] have proposed solutions based upon user preferences (similar in spirit to some earlier solutions of O'Gorman et al. [15]) and obtained relatively positive security and usability results. However, in support of usability, it would be desireable to see longer-term studies of their solutions. While a small number of weeks is generally sufficient to evaluate the capabilities of long-term memory, approaches based upon preferences must also evalute the impact of those preferences inevitably changing over time.

Just [9, 10] provided an early framework for challenge question design, including criteria for secure and usable solutions. Though the framework wasn't rigorously applied to an actual system. Rabkin [17] analyzed the security and usability of challenge questions from 20 online banking sites and found significant numbers of questions were either insecure or difficult to use. He also noted the trade-offs with security as there appeared to be a strong inverse relationship between those questions that were secure and those that were memorable. However, while Rabkin performed a user survey, he did perform a user study using the questions he collected, and did not provide recommendations for improved design. We make further reference to this research throughout our paper.

3. OUR EXPERIMENT

In our pilot studies [11], we designed a purely manual pen-and-paper experiment conducted with two classes of Informatics students that leveraged the *public-private* relationship between challenge questions and their answers. This allowed us to focus on the public portion (the questions) for our security analysis and engage participants in a self-assessment of the private portion (their answers). The analysis examined only basic guessability of the answers and did not incorporate the systematic application of the Security Model presented in this paper. Given a scenario of providing challenge questions for password recovery at an online bank, we asked participants to write down three questions and their answers. The questions were written on one sheet, and the answers on another. We collected questions but participants kept their answers in a sealed envelope. We then returned after a period of several weeks and asked participants to answer their questions again, in a classical usability study to address memorability and repeatability of responses. The accuracy and recall were assessed by participants themselves, so we did not need to see the complete authentication information.

3.1 Method

Our larger follow on experiment was conducted with first year Biology students. To scale up, we revised the previous mechanism to use a hybrid online and paper method. Participants submitted their challenge questions and responded to additional survey information online. But they still retained their answers on a sheet of paper in a sealed envelope provided in advance.

The overall process is depicted in Figure 1. To encourage participation, we indicated that prizes would be awarded to

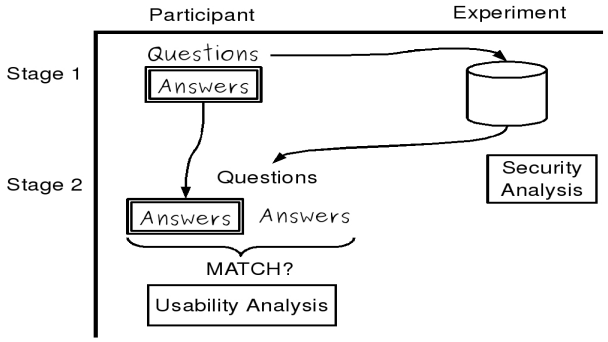

Figure 1: Experimental Method

a number of randomly selected participants at the end of the experiment.[1] Students received an *experiment package* including handout and sealable envelope, both marked with an anonymous identification number (ID).

For Stage 1, participants visited our web site and were then redirected to our survey. At our survey, they were asked for their ID, gender, year of birth, as well as their previous level of experience with challenge question authentication. They were then presented with a *banking scenario* and asked to imagine registering for an account at their bank. As part of the scenario, they were asked to enter their challenge questions online and enter the corresponding answers on the handout they received in class. Particpants were asked to seal the handout in the provided envelope until instructed to open during Stage 2. Participants retained their envelopes.

Stage 2 began 23 days after the end of Stage 1. On returning to the experiment web site, participants were asked for their ID (to facilitate analysis with Stage 1 data) and then shown their original challenge questions. They were asked to respond to their questions and write their answers on the outside of their envelope and were then redirected to our survey where they were asked to open their envelopes and respond to the following:

- Participants were asked for the length of each answer (that was submitted during Stage 1). Answer length was used in the security analysis (Section 4) (we did not want to ask users about it at the first stage lest they would be influenced.)
- For each question, their memorability of the original answer (Section 5).
- Their perceptions of the experiment, and in particular the influence (if any) of not submitting their answers, on providing realistic information for the experiment. (Section 3.2).
- Their perceptions of the security of their information by asking the perceived difficulty of strangers or friends and family to determine their answers (Section 4.3).

We handed out 289 experiment packages prior to Stage 1. In total, 97 students took part though only 94 submitted challenge questions. Their average year of birth was 1989, with respective earliest and latest years of 1970 and 1991. Of the 94 participants submitting questions, 4 (4%) indicated no previous experience with challenge question authentication, 20 (21%) indicated "low" experience (used on 1–3 previous occasions), 48 (51%) indicated "medium" experience (used on 4-9 previous occasions) while 22 (24%) indicated "high" experience (used on 10 or more previous occasions). 60 participants returned for Stage 2.

[1]Full details of our experiment, including presentation slides, handouts and webpages are available at our project web page [12].

3.2 Accuracy of the Experiment

Experiments with authentication mechanisms are fraught with difficulty, particularly with ethical concerns over gathering private information that may actually be used by individuals to identify themselves. Anecdotal data can be obtained by getting permission of a few friends or colleagues to break into their accounts [21], but a scientific study needs to do better.

Our experiment was designed to encourage participants to submit realistic data by not asking them to submit their answers. We tried to assess this by asking participants if it influenced how they responded. In other words, did the fact that they did not have to tell us their answers encourage them to participate more honestly in our experiment and use *realistic* questions and answers? While we can't answer conclusively, we saw some encouraging results. We asked participants "For this experiment, do you honestly feel that you used questions that you would actually use in an online application such as banking?" Of 60 respondents, 37 (62%) responded with "Yes, very much" while an additional 19 (32%) responded with "Yes, maybe" for a total of 94%. Only 1 (2%) participant responded with "No," while 3 (4%) participants were "Not sure." Written comments suggested that some participants had some concern regarding whether their questions and answers were sufficiently secure to resist attack (perhaps a result of their participation in this experiment).

Of the 56 respondents that answered "Yes", we asked: "Was this decision influenced by the fact that you did not have to tell us the answers to your questions for this experiment?", and 52 participants responded. While not overwhelming, we received a positive response with 5 (10%) of the respondents indicating "Yes, very much" and 18 (35%) indicating "Yes, somewhat." However, 29 (55%) of the respondents indicated "No, not at all." It thus appears reasonable to suppose that there may be some advantage to our experimental process, especially if it could influence more than 40% from participating with unrealistic authentication information.

4. SECURITY MODEL AND ANALYSIS

Our security model extends ideas of both Just [10] and Rabkin [17]. Just proposed *Guessability* and *Observability* as criteria for evaluating the security of challenge questions. Guessing difficulty is related to an attacker's likelihood of guessing an answer based upon the answer entropy, or perhaps some other analysis of the question. Observation difficulty is related to the effort needed by an attacker to discover or observe an answer, or information that might contribute to determining it. Rabkin [17] uses similar concepts, namely *Guessable*, which aligns with both our Blind Guess and Focused Guess, and *Attackable* and *Automatically Attackable*, which align to our Observations.

In this section, we revise these previous investigations, introducing a new security model which considers three different modes of attack based on the knowledge of the attacker, and three corresponding analyses. A key innovation is to allow for a closer comparison with password security, by taking the first mode of attack to be one where the attacker does not even consider the question being asked.

4.1 Security Model

Our model is depicted graphically in Figure 2. It shows three methods (modes) of attack toward the goal of guessing the answer of a challenge question, and considers the information available to an attacker for each method:

Blind Guess the attacker does not consider the question; the attack is a brute force attack, but in a space where there is considerable opportunity for using dictionaries.

Focused Guess the attacker considers the question; perhaps automatically, he may be able to cut the search space considerably by identifying the likely *type* of the data in responses.

Observation the attacker considers the user as well as the question; at this point, he may apply information gleaned from public records, social networks, and direct observation.

The methods can be applied independently, but may well be used in combination by an attacker. At first glance, it might seem excessive to consider the 'Blind Guess'; afterall, wouldn't an attacker that guesses the answers to the questions have access to the questions? However, consider that an attacker that is performing an automated attack may not be able to easily read and/or comprehend what is being asked by a question (Rabkin [17] notes that CAPTCHAs might be useful in this regard to combat such automated attacks). In this case, the attacker could still perform a blind guess, perhaps by guessing English words starting from a length of one character. As well, we anticipate that our model might also be useful to system designers for the dynamic assessment of questions and answers entered by users. In this case, an examination of the length of the answer would be an important consideration.

Below, we define each part of the model and its associated security metrics and then (in Section 4.2) apply this model to evaluate the challenge questions we received during Stage 1 of our Experiment. In Section 6, we will discuss how the model may also aid in the design of secure challenge questions.

A *Blind Guess* uses knowledge of the alphabet of characters for the set of answers, and their probability distribution. For our purposes, we assume this alphabet to be the set of 26 lowercase letters from the English alphabet (one may choose to additionally include the set of 10 numbers, where appropriate). The probability distributions for English words is well-known and most notably studied by Shannon [19]. Determining the level of security for a particular challenge question system (or set of questions) requires knowledge of the length of answers provided by users.

A *Focused Guess* involves knowing (and understanding) the challenge question, from which an attacker is (typically) able to focus on a set of candidate answers. Determining the level of security for a particular question can be made

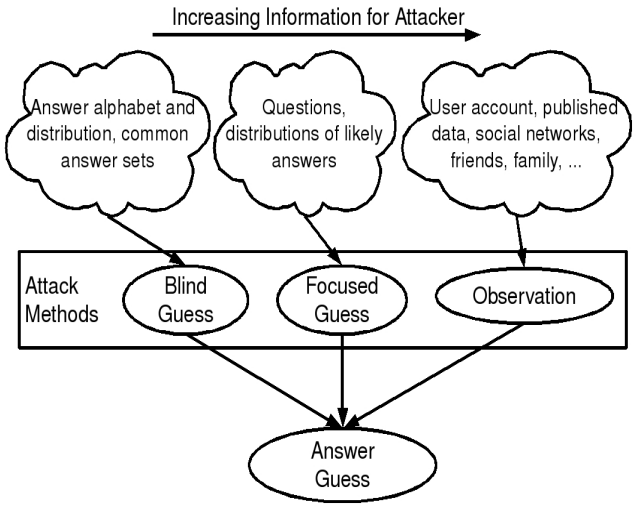

Figure 2: Challenge Question Security Model

based upon the size and distribution for the set of possible answers.

The resistance to a Blind Guess or a Focused Guess is based upon the following levels of security, relative to the calculated entropy of the answer to a challenge question:

- *Low* — less than 2^{34} possible answers
- *Medium* — between 2^{34} and 2^{48} possible answers
- *High* — greater than 2^{48} answers

where the security levels were derived from comparable levels for passwords. For *Low*, we assigned the entropy level for a randomly chosen 6-character password constructed only from alphabetic characters: $52^6 \approx 10^{10} \approx 2^{34}$. For *Medium*, we assigned the entropy level for a randomly chosen 8-character password constructed from alphanumeric characters: $62^8 \approx 10^{14} \approx 2^{48}$. This approach is generally consistent with NIST authentication guidance [14] and related assessments of secure credential length [4].[2]

A guess following an *Observation* involves *observing* information about the user that allows the attacker to either stereotype the user based upon apparent peer groups, or determine the answer itself. The resistance to such an attack is measured according to the following levels of security.

- *Low* — the answer to the question is publicly available and readily known to family and/or friends.
- *Medium* — the answer to the question is known to family and/or friends but is not believed to be publicly available.

[2]NIST includes a "Very High" level, which we might add: this is the level required to strongly resist offline attack, based upon NIST's guidance for minimum key length (until the year 2010) of 80 bits [13]. And we could similarly consider a "Very Low" level for answers with extremely low entropy, but for our purposes felt that a wider measure, in which more questions would be identified as Low, was more approriate.

Table 1: Security Levels Achieved by Question

Attack Type	Low	Medium	High	Total
Blind Guess	174	4	2	180
Focused Guess	167	0	13	180
Observation	124	54	2	180

Table 2: Security Levels Achieved by User

Attack Type	Low	Medium	High	Total
Blind Guess	5	19	36	60
Focused Guess	5	28	27	60
Observation	24	34	2	60

- *High* — the answer to the question is neither known to family and/or friends, nor is it believed to by publicly available.

It is not easy to quantify the availability of observed information, so we have opted for a subjective, qualitative measure. However, our measure does capture the two most important distinctions: namely, whether information is publicly available to anyone and whether it is known to people connected to the user. Determining the level of security for a particular question can be made based upon the apparent availability of the corresponding answer.

Observations attempt to learn more information about the account owner, from which information can be derived regarding their answers. The information will likely be helpful with some questions more than others. For example,the user identifier can sometimes reveal information about the user, particularly their first and/or last name. Such information can be very helpful in making a follow-up observation, for example, with publicly available records. An attacker can often guess a candidate user identifier as part of an account recovery process and verify its correctness based upon whether a set of challenge questions are returned.[3] From the challenge questions themselves, an attacker might be able to learn information about the gender, age range, interests, or place of evidence for the user. Similar to the information revealed by the questions, an account at a particular site might allow an attacker to learn information such as gender, age, interests. All such information might help to reduce the likelihood for other answers.

Beyond the account of the user and the site at which the account is held, the prime sources of information about a user are public information sources, and the user themselves. Rabkin [17] has already noted the utility of social networking sites and Griffith and Jakobsson [6] note the ease with which Mother's Maiden Names can be determined through observation.

When considering all the attack methods, notice that they are not necessarily increasing in strength. Thus, an answer that is susceptible to a Blind Guess may remain elusive against an Observation. Consider, for example, a last name with only two characters (e.g., "Ng") that might easily fall to a Blind Guess but (dependent upon what question solicited this answer) may not be easy to determine through Observation.

4.2 Security Analysis

We now apply this security model to our experimental data. Before explaining the detail, we present an overall summary of the results. This is shown by question in Table 1 and by user in Table 2.

The tables summarise the results for 180 questions, from the 60 users that took part in both stages of our experiment. The security levels of questions are independent assessments of the strength of each question against each type of attack. The security levels of users, meanwhile, is the consideration of the security of *all* of their 3 questions together under the given type of attack. Thus, although no questions individually achieved an entropy of between 2^{34} and 2^{48} (the delimiters of the Medium Level of security) considering the answer spaces for a Focused Guess (0 questions with Medium rating in Table 1), there are 28 users whose 3 questions taken together implied a large enough search space.[4]

4.2.1 Blind Guess Analysis

To analyse the difficulty of Blind Guess attacks for our data, we use the lengths of answers which were provided by our experiment participants. Of course, a real attacker does not know the length of the answer; our aim in this analysis is to examine the actual difficulty for the data we collected. And short answers would be particularly susceptible to simple attacks that attempt English words of ever-increasing length.

We calculated the entropy as follows (assuming only 26 lowercase letters are used). According to Shannon [19], "when statistical effects extending over not more than 8 letters are considered the entropy is roughly 2.3 bits per character." Shannon went on to show that for longer text, an entropy of 1.5 bits per character can be used. Using this guidance, and consistent with NIST [14], we assigned 4 bits to the first character, 2.3 bits to the next 7 characters, and 1.5 bits for each character thereafter.

The entropy calculation implies that we require at least 18 characters in order to reach the *Medium* security level, and 27 characters to reach the *High* security level. For the 180 challenge questions assessed as part of Stage 2, the average answer length was 7.33 characters (median of 6). Hence, the results of Table 1 contain a significant number of questions (174) with *Low* security.

For passwords, while it is functionally possible to ask users to use longer passwords, there would typically be a negative impact on usability. With challenge questions, it can also be functionallly difficult to require a longer answer: for the name of a pet, or a mother's surname, the answer length is fixed.

However, the security picture improves if we look at the use of more than one question, which we calculated by summing the respective entropy values for the questions from each user:

First question only: Low: 59, Medium: 1, High 0

First two questions: Low: 38, Medium: 13, High: 9

All three questions: Low: 5, Medium: 19, High: 36

where the result for use of all three questions is captured in the summary of Table 2.

[3]We did not ask users to choose a user identifier as part of our experiment.

[4]This is with the best case assumption that the answers are independent.

Table 3: Question Types and Answer Spaces

Question Type	Percentage	Answer space
Proper Name	50%	10^4–10^5
Place	20%	10^2–10^5
Name	18%	10^3–10^7
Number	3%	10^1–10^4
Time/Date	3%	10^2–10^5
Ambiguous	6%	10^8–10^{15}

Assuming independence between the answers, it appears that systems that are reasonably secure against Blind Guess Attacks can be achieved by using a small number of questions. We discuss this further in Section 6.

4.2.2 Focused Guess Analysis

In a Focused Guess attack, the attacker has knowledge of the challenge questions, and sets of possible answers. To analyse our data, we therefore examined the questions submitted by participants in our experiment. We consider the same security levels as for a Blind Guess, but now the space of possible answers may be partitioned. We do not consider the answer lengths in this analysis.

The first step is to consider the type of responses that may be given in (truthful) answer to the questions. Table 3 shows a classification of question type for the total 282 questions submitted, along with the range of the estimated answer space sizes we assigned.[5] This classification of types was designed to cover the user-generated questions in our experiment, but it turns out to be similar to the classification results for administratively-generated questions from Rabkin [17].

The answer space estimate uses a finer sub-classification for each question. We estimated the possible number of answers for each of these categories from public sources of information, as far as possible, and then by making reasonable assumptions.

For example, considering the largest category of Proper Name, the further breakdown considers Last Name, First Name, First and Last Name, Pet Name, and Other Name. According to the Year 2000 US Census, there were $151,671$ last names (surnames) [24], so we used 10^6 as answer space size for Last Name. This means that without even taking into account the (highly skewed) distribution of these last names, an answer with a "Last Name" response would thus be *Low* security according to our security levels. Similarly, the 1990 US Census indicates $1,219$ male and $4,275$ female first (given) names [23]. Clearly either a first or last name would offer *Low* security. Unfortunately, even the combination of both a first *and* last name for males (for example), which gives just about 184 million possibilities, is still only 27 bits of security.[6] This is still not enough to achieve *Medium* security.

Pet naming is notoriously stereotypical, and many web sites provide lists of common pet names. For this category, we chose an answer space size of 10^4. Of course, the space of actual possible names is much larger, so in cases like this our analysis is supposing that answers are taken from the most commonly chosen set. If your Pet's name is obscure and not on any common list, it will be more secure against a Focused Guess — however, if it is short, it may still be insecure against a Blind Guess.

Notice that the assumptions based on the size of public records implicitly makes some level of observation, by supposing the person to have a name that appears in the 1990 US Census (or a pet name typically chosen by US dog owners). In general, we expect the broad categorisations that we use may have similar sized answer spaces in other countries and cultures, although the exact lists will of course be different. The right list could be chosen by some easy *site* or *language* observation made by the attacker, or the list expanded by taking the union of available lists, probably with cost of at most an order of magnitude.

Some questions immediately have very low entropy. Only one of our participants chose a question that was (apparently) *factual* (*"What man made structures can be seen from the moon?"*), which could be immediately answered by anyone. But others chose questions that clearly had a very small range of possible values. This included all of the 8 questions in the number category, including *"How many steps are there in the staircase in your house?"* (we guessed less than 100), *"What year was your father born in?"* (we supposed a range of up to 100 years) and *"4 digit code for the range"* (an easy enumeration of 10^4 values).

After the estimate of answer space per-question, we made the same calculations as for the Blind Guess attack to consider the security achieved for each user as questions are accumulated:

First question only: Low: 58, Medium: 0, High 2

First two questions: Low: 46, Medium: 11, High: 3

All three questions: Low: 5, Medium: 28, High: 27

where the result for use of all three questions is captured in the summary of Table 2.

4.2.3 Observation Analysis

With observations, an attacker is concerned with either recovering the answers themselves, or further information (often personal) that reduces the set of possible answers. In this sense, for the question *"What is my mother's maiden name?"* while a Focused Analysis would be concerned with determining sets of likely last names, an observation is *targeting* the particular user in order to infer their mother's maiden name.

Our assessment was performed as follows. Based upon our security levels, we performed a subjective assessment of each question and assigned each to have either *High*, *Medium* or *Low* security. This was done independent of the results of either the Blind or Focused analysis (so that questions that might be rated *Low* in the Blind Analysis could still merit a *Medium* rating versus our observation, or vice-versa). For the most part, questions asking for Proper Names were assigned *Low*, due to their likely public availability. Though some names, that we surmized might be known to friends or family, but not publicly available, were assigned *Medium*, consistent with our definition. For example, while *"What*

[5]'Proper Names' include family members, pets, friends, lovers, community members (vicars, teachers, ...), and celebrities (movies, sports, ...). 'Places' include names of cities and schools. 'Names' include manufacturers, Internet sites, email addresses, animal types/breeds, and leisure activities (music, food, film, books, games, sports, ...).

[6]Again, this does not take into account the distributions or any potential correlations between first and last names.

is my [current] pet's name?" was assessed as *Low*, *"First pet's name?"* was assessed as medium.[7] Subsequent to this assessment, of the 282 questions, we had 15 (6%) questions with a *High* security level, 83 (29%) *Medium* and 184 (65%) *Low*. However, we felt it was not sufficient to use only this subjective analysis to determine the security levels.

As will be discussed in Section 4.3, we asked participants for their perception of the difficulty for either a Stranger or Friends and Family to discover the answers to their questions, with responses of either *"Very Difficult"*, *"Somewhat Difficult"*, or *"Not Difficult at all"*. Therefore we used the participants own perception of the security to aid our assessment, but only to provide us with an upper bound. In other words, we trusted participants to tell us that an answer was less secure than we had assessed, but we did not trust them to tell us when they thought it was more secure. So, we developed an adjusted rating using this input and based upon the following rules.

Because of the tie between public availablility of the answers (from our definition) and the difficulty for a stranger to determine the answer (from the question to participants), if a participant indicated that an answer was *"Not Difficult at all"* for a stranger to discover, we reduced our ranking to *Low*. If the user responded with *"Somewhat Difficult"* for a stranger, we reduced our ranking to *Medium*. And due to the tie between the knowledge of the information for friends and family (in both our definition and the question to participants), if a participant indicated that an answer was either *"Somewhat Difficult"* or *"Not Difficult at all"* for friends or family, we reduced our ranking to *Medium*.

Since we questioned participants about their perceptions during Stage 2, our resulting adjusted levels apply only to the 180 questions from this stage. The results were 2 (1%) questions with a *High* security level, 54 (30%) *Medium* and 124 (69%) *Low*, and are summarized in Table 1. Of the 180 questions, there were only 4 adjustments made, giving us some confidence in our original assessment.

As with the Blind and Focused Guesses we calculated the overall entropy for each user, considering the combination of all three of their questions. Since we believe that an attacker would be able to more easily attack questions in parallel using an observation (since candidate answers can be validated external to the authentication system), we concluded that it would not be realistic to compute the overall security level for the set of three questions per user to be the sum of the level for each question. Thus we assigned a user rating of *Low* if all questions had individual *Low* ratings, a user rating of *Medium* if there was at least one *Medium* question (but no *High*) and a user rating of *High* if at least one question was rated *High*. The results for use of all three questions is captured in the summary of Table 2 where, noticably, there are only 2 users with a *High* rating.

4.2.4 The Final Guess

Given the security analysis above for Blind Guess, Focused Guess and Observation, we may now consider the security of the challenge questions against simultaneous application of all of these attacks.

We have not attempted to combine the measurements into a single overall assessment. This is because, without additionally considering a *cost model* for the attacker to obtain the different amounts of information, the modes of attack remain rather separate.

Nonetheless, it is interesting to look at the submitted question sets and examine how many sets achieved particular levels within our analysis. Here are some data points, out of the 60 users:

All low: only one user gave 3 questions that together failed to achieve even a medium rating against any attack. The questions all concerned the same close family member.

No lows: 31 users managed to avoid a low security rating altogether. This suggests that, taken together, their questions could resist moderate attack efforts.

High, medium, medium: 15 (25%) users achieved this, by managing High against either a Blind Guess or a Focus Guess, and Medium for the others.

High, high, medium: 12 (20%) users achieved this, 11 with Medium against an Observation, 1 with Medium against a Blind Guess.

All high: no users achieved a set of three questions which were rated high against all modes of attack.

The final result is somewhat encouraging, then. None of our participants chose a "perfect" set of questions[8]. But it is encouraging that over half managed to achieve a moderate level of security on the three fronts. We did not bias the users' choices by giving them any encouragement to think hard about their questions, and gave them no security advice whatsoever beyond the stated scenario. Indeed, some participants chose questions that they expected had low security, at least when we asked afterwards.

Yet one must take care with extending these results too far. As noted earlier, our security model will likely be refined and improved over time. For example, we have not yet considered the impact of question dependencies where, for example, knowledge of the answer to one question may reduce uncertainty in the answer to another. As well, while the data points above consider overall ratings against each individual attack, they may not yet capture all possible attacks. For example, with three questions, two of which are Low against an Observational attack, an attacker may be able to effectively attack all three questions in parallel should the other question have Low security against either a Blind or Focused Guess. We anticipate though that our security model will now allow various attack scenarios to be systematically analyzed.

4.3 Participant Security Perceptions

As part of our experiment, we asked participants for their perceptions as to the security of their answer versus either Strangers, or Friends and Family. From 59 respondents, the following responses were given regarding user perception of how difficult it would be for a stranger to determine the answers to each of their 177 questions: 84 chose *"Very Difficult"*, 79 chose *"Somewhat Difficult"*, and 14 chose *"Not Difficult at all"*. Based upon our results above, there appears to be a large discrepancy in user understanding of the abilities

[7] We assumed that some current information might be available on social networking sites, but that historical information might only known to family and friends.

[8] Partly because, in particular, achieving high security for observation attacks was not easy

of an attacker. (We did not ask users for their opinion of the strength of all of their questions together.)

From 60 respondents, the following responses were given regarding user perception of how difficult it would be a friend or family member to determine the answers to each of their 180 questions: 15 chose *"Very Difficult"*, 67 chose *"Somewhat Difficult"*, and 98 chose *"Not Difficult at all"*. These results were not terribly surprising given that personal information is often shared amongst family and friends, and coincide quite well with our analysis. It is also interesting to question whether users are terribly concerned with the fact that they are using authentication credentials that might be known to family and friends. The fact that the information is known to family and friends confirm those early results of Haga and Zviran [7, 25] and coincide with recent results of Schechter et al. [18]. The more interesting question might be as to why users don't understand, or aren't concerned, with this result.

It was also interesting to note some novel, though unsuccessful attempts to improve security, and recognize some of the more popular shortcomings. For example, one participant asked the question *"What is your father's name spelled backwards?"*, perhaps not recognizing the ineffectiveness against an attacker that also views the question using a Focused Guess. A second participant asked the question *"What is your grandmother's maiden name?"* which only adds a small degree of difficulty to this question [6].

5. USABILITY MODEL AND ANALYSIS

We performed two types of analysis in order to learn more about the usability of the challenge questions obtained from our experiment. Firstly, we asked participants to perform a self-assessment of their answer recollection during the second stage of the experiment. Secondly, we analysed the questions alone and made some assumptions about their answer spaces.

In both cases, we used the usability criteria given by Just [10], namely:

- *Applicability*: How widely applicable is the given question?
- *Memorability*: How easy is it for the user to recall the answer?
- *Repeatability*: How accurately can the answer be replayed, without syntactic or semantic ambiguity?

Applicability is a given with user-chosen challenge questions, so we consider memorability and repeatibility only. To measure these, we looked at participants' own accuracy results and the reasons they gave for making mistakes. This information was gathered as the main response in the online survey in Stage 2.

During Stage 2, we asked participants whether or not they responded with an answer that matched their original answer *exactly*. If not, we asked them to specify the reason for the discrepancy, which included "complete blank" in which they provided no answer, completely different answer, different spelling, or some other reason. We also tracked errors due to different capitalization, spacing or punctuation, even though we believe that most systems should normalize answers to filter out such differences.

To our surprise, despite the youthful minds of our participants, 11 out of 60 Stage 2 participants (18%) were unable to reproduce at least one answer exactly after a period of approximately 23 days between the end of Stage 1 and start of Stage 2. This equated to 15[9] out of 180 questions (8%). If we compensate for errors based on capitalization, spacing or punctuation, we discover that 7 of 60 participants (12%) were unable to reproduce at least one answer exactly. These results compare quite dramatically with those of Florêncio and Herley [4] in which 4.28% of Yahoo! users forgot their passwords; especially when challenge questions are often used as a fallback to a forgotten password.

We believe this highlights the difficulty of expecting a 100% accurate response rate from participants. We suspect that two issues might be contributing to this result

- There is a degree of memorization that takes place for at least some questions simply by the user having to associate known information with a question (even though the question has been chosen by the user).
- There is a difference between *knowing* a piece of information and repeating it exactly in written form (whether typed or hand-written).

It may also be that some users did not provide honest responses to their challenge questions, and later could not recall what answer they provided. Our experiment did not ask participants whether they answered their questions honestly, nor is it clear that a participant that did not honestly answer would tell us so.

Of the 7 participant errors, 4 provided responses and comments that indicated an issue of repeatability; either spelling mistake or mistakes related to the use of multiple words in their answer. Of the remaining 3 participants, 2 indicated that they simply provided a completely different answer (in responding to the name of a school and pet name). Without a larger sample size we cannot provide a clear link to the types of questions that might more strongly correlate to such mistakes. Though despite the small sample size and limited demographics of the participants, there appears to be a question of users' ability to successfully recall and repeat their answers.

Further related to repeatability, while the duration between stages was sufficient for evaluating the efficacy of long-term memory, it does not address the changing of information over the long-term. Thus, we were also curious to classify the challenge questions based upon their notion of *time* and *preference*.

Approximately 40% of questions made reference to a first-time or past event. While first-time events are relatively stable, Just [9] indicated that they were more difficult to recall, especially for older users. General past events potentially introduce different challenges. While many questions (e.g., *"Which city were you born in?"*) refer to one-time events, others (e.g., *"Personal Best for Sport?"*) may change from the time the question is registered, till the time the user authenticates.

Approximately 25% of questions made reference to the current time, either explicitly (e.g., *"What's your favourite place?"*) or implicitly (e.g., *"Favourite holiday destination?"*) as part of the question. Similar issues can arise since although the question seems to be asking for a user's current

[9]Two participants made errors with all three questions, while nine participants made an error with only one question each.

"favourite place," the user must respond with the *original* answer that was provided at registration.

Identical issues apply to questions that ask for a preference ("favourite," "best," "most memorable," "worst"). Since one might expect such responses to change over time, one might expect that users would have difficulty in providing correct responses over the long-term. Our results were not able to justify such possibilities, but we suspect it would be worthy of long-term evaluation.

6. DISCUSSION

From our analysis of the experiment data against our models, there appear to be a number of key themes evident. While challenge question authentication can appear to be quite insecure in certain instances, we feel that with some simple improvements, secure and viable authentication solutions remain possible.

Our security analysis suggests the necessity (though not sufficiency) of multiple questions. Though systems will have their own risk profiles, we suspect that all would be wise in using more than one question and answer pair. And while *'t of n'* schemes whereby the user is posed n questions and required to successfully answer only $t <= n$ seems prudent as a way to reduce usability issues related to repeatability, this would mean that $t > 1$. Based upon the assumptions made as part of our analysis, three questions seems prudent.

And while our current analysis did not focus on this aspect, the *independence* of the questions and answers would be of critical importance in this case. For example, asking three questions about your pet would likely not provide an increased level of security than asking only one of the questions. Such analysis, likely leveraging conditional probabilities, has not yet been considered as part of applying our security model.

An alternative to posing multiple questions is to mandate minimal answer lengths. However, this would likely have strongly negative impacts on usability. Consider that, unlike passwords, the answer to a question is not typically adjustable, e.g., there are many last names of length 4 characters. This would still allow for the entropy across all answers to be measured at registration, perhaps suggesting that a user to select additional questions if needed, similar to *dynamic assessments* that are used to inform users as to the strength of their passwords.

Our analysis revealed that many users were given a low security level due to questions that could be clearly identified as not sufficiently secure. Thus, we suggest that such questions should be avoided altogether (either as administrative choices, or filtered out as user choices), for example: *"What colour is your favourite fruit?"*, and *"Favourite musical instrument?"*. Even further, there appears to be a lack of variety in the types of questions asked. Most questions aligned to administratively generated questions used by financial institutions [17]. In other words, there was a surprising lack of creativity in the user chosen questions we collected. Thus, it remains an open problem as to whether there are more creative questions and answers that could be used by users (in general), whilst providing an improved level of security and usability.

In dealing with potential *time-based* issues and users' ability to recall or repeat past events, it may be prudent to suggest that users use questions that refer to *one-time* events in order to remove ambiguity. Similarly, in dealing with preferences, while *favourites* might change, it may be possible for users to recall *most memorable* times, places, events or even people. We would recommend though that such ideas be validated first through user study as they have not been vetted by our experiment.

As an aid to repeatability, fixed format answers would likely improve repeatability, especially for questions asking for a *date* answer. One participant went so far as to include a format reminder in their question, presumably in a way that would help them remember the structure of their response, e.g. *"What date is your Mother's birthday on? (eg. 01/01/00)"*.

For system designers, current challenge question implementations readily allow attackers to validate their username guesses (unlike password authentication whereby a generic 'login failed' is returned for either a false username or password). It might be helpful then to provide *fake questions* in response to a non-existent username.

An interesting observation that can be read into our model and results is that, depending upon the risks to a particular system, it is not necessarily a bad decision to choose challenge questions that offer *Low* security against personal observation attacks, so long as the security of the question (and likewise of additional questions) is sufficiently secure against Blind Guess Analysis and Focused Guess Analysis. For example, while *"Mother's Maiden Name?"* has been shown to be quite easy for an attacker to determine [6], the question can offer some limited protection versus the other attack methods.

7. CONCLUDING REMARKS

With our models, analysis and experimental results, we hope to provide a clearer picture as to the security and usability offered with challenge question authentication. While still early days, our results do seem to suggest that it may be possible to design a reasonably secure solution. In particular, the Security Model we've introduced indicates that while individual questions typically provide very limited security, and there exist numerous individual questions that are themselves insecure, there are choices around which a secure system can be built. We have provided some suggestions for this to happen, and hope that our work can be extended to further this goal.

Though while there might be some optimism for more secure questions, usability remains a challenge. And while further experiments should be performed on larger, more diverse populations, and over longer periods of time, our initial results indeed indicate usability issues regarding the memorability and repeatability of answers. Novel solutions that ensure security and usability are needed in this area.

Our experimental method presents an interesting option for obtaining more realistic authentication information in an ethical way. Though while its use of pen-and-paper aids us in this effort, the same practice introduces some factors that are difficult to control. For example, the self-assessment of memorability places a significant amount of trust in the participant. While we might not expect a significant portion to wilfully contribute false results, some participants may not have fully understood their tasks. Additionally, in our pilot experiments [11], some participants commented that they would not have made errors with capitalization had they typed their answers. We do not know if writing, as opposed to typing, the answers would have helped or hindered partic-

ipants to recall their answers. However, we feel that despite these limitations, our experimental model does have some promise for obtaining realistic authentication information from users.

As for experimenting with challenge questions in general, it is possible that security savvy users may resort to giving password-style responses (i.e., a dishonest response to their question). While some people do this, we did not query our participants whether they did. Inventing a question and then giving a deliberately unrelated random string response as an answer is hardly likely to be recommended practice for a usable authentication system, especially one often used as a form of password recovery.

And finally, it is worth noting some other recent applications of challenge question authentication. While the present work only considered the use of questions and answers by a single individual, the work of Toomim et al. [22] and Bonneau [2] considers the use of information shared by pairs or groups of users in order to support group authentication. Though while not addressed explicitly, our model and assessments would likely be useful to measuring security and usability in these environments as well.

Acknowledgements

We would like to thank the anonymous referees for their thoughtful and helpful comments. The paper is much improved as a result of their input (especially that of suggesting a change of the original title). We would also like to thank Chris Mitchell for pointing out the work of Bonneau et al. [2], and thank Joseph Bonneau for several interesting conversations. We are grateful to all the students who took part in our experiments, including those that gave permission for their questions to be quoted here. We also acknowledge the support of the UK EPSRC, Grant No. EP/G020760/1, which is funding the first author as a Visiting Research Fellow at Edinburgh.

8. REFERENCES

[1] F. Asgharpour, M. Jakobsson, "Adaptive Challenge Questions Algorithm in Password Reset/Recovery," in *First International Workshop on Security for Spontaneous Interaction (IWIISI '07)*, Innsbruck, Austria, (2007).

[2] J. Bonneau, "Alice and Bob in Love: Cryptographic Communication Using Natural Entropy," to appear in *Proceedings of the 17th International Workshop on Security Protocols 2009*, Cambridge, UK, April 2009.

[3] C. Ellison, C. Hall, R. Milbert, B. Schneier, "Protecting Secret Keys with Personal Entropy," *Journal of Future Generation Computer Systems*, **16(4)**, (2000), 311-318.

[4] D. Florêncio, C. Herley, "A large-scale study of web password habits," in *Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07*. ACM, New York, NY, 657-666.

[5] 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.

[6] V. Griffith, M. Jakobsson, "Messin' with Texas, Deriving Mother's Maiden Names Using Public Records," *RSA CryptoBytes*, **8(1)**, (2007), 18-28.

[7] W. Haga, M. Zviran, "Question-and-Answer Passwords: An Empirical Evaluation," *Information Systems*, **16(3)**, (1991), 335-343.

[8] M. Jakobsson, E. Stolterman, S. Wetzel, L. Yang. "Love and Authentication," in *Proceedings of ACM Human/Computer Interaction Conference (CHI)*, (2008).

[9] 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.

[10] 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).

[11] M. Just, D. Aspinall, "Challenging Challenge Questions," presented at *Trust 2009: International Conference on the Technical and Socio-Economic Aspects of Trusted Computing*, 2009. (Available at [12])

[12] Knowledge-Based Authentication Project Site. `http://homepages.inf.ed.ac.uk/mjust/KBA.html`

[13] National Institute of Standards and Technology (NIST), "Recommendation for Key Management - Part 1: General (Revised)," *NIST Special Publication 800-57*, March 2007. `http://csrc.nist.gov/groups/ST/toolkit/documents/SP800-57Part1_3-8-07.pdf`

[14] National Institute of Standards and Technology (NIST), "Electronic Authentication Guideline," *NIST Special Publication 800-63*, Version 1.0.2, April 2006. `http://csrc.nist.gov/publications/nistpubs/800-63/SP800-63V1_0_2.pdf`

[15] L. O'Gorman, S. Begga, J. Bentley, "Call Center Customer Verification by Query-Directed Passwords," in *Proceedings of Financial Cryptography '04*, *International Financial Cryptography Association*, (2004).

[16] 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.

[17] 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).

[18] S. Schechter, A. Bernheim Bruch, S. Egelman, "It's no secret. Measuring the security and reliability of authentication via 'secret' questions," to appear in *Proceedings of the IEEE Symposium on Security and Privacy*, 17-20 May 2009.

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

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

[21] H. Thompson, "How I Stole Someone's Identity", Scientific American, online feature posted August 18, 2008. Retrieved from `http://www.sciam.com/article.cfm?id=anatomy-of-a-social-hack`, 23rd

February 2009.

[22] M. Toomim, X. Zhang, J. Fogarty, J. Landay, "Access Control by Testing for Shared Knowledge," in *Proceedings of CHI 2008*, Florence, Italy, April 2008, ACM.

[23] U.S. Census Bureau, *1990 Census - Names*, available at `http://www.census.gov/genealogy/names/names_files.html`.

[24] U.S. Census Bureau, *Frequently Occurring Surnames from Census 2000*, available at `http://www.census.gov/genealogy/www/freqnames2k.html`.

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