1 + 1 = You: Measuring the comprehensibility of metaphors for configuring backup authentication

Conference Paper · January 2009

DOI: 10.1145/1572532.1572544 · Source: DBLP

1 + 1 = You

Measuring the comprehensibility of metaphors for configuring backup authentication

Stuart Schechter
Microsoft Research
1 Microsoft Way
Redmond, WA 98052
stus@microsoft.com

Robert W. Reeder
Microsoft (TUX)
1 Microsoft Way
Redmond, WA 98052
roreeder@microsoft.com

ABSTRACT

Backup authentication systems verify the identity of users who are unable to perform primary authentication—usually as a result of forgetting passwords. The two most common authentication mechanisms used for backup authentication by webmail services, personal authentication questions and email-based authentication, are insufficient. Many webmail users cannot benefit from email-based authentication because their webmail account is their primary email account. Personal authentication questions are frequently forgotten and prone to security failures, as illustrated by the increased scrutiny they received following their implication in the compromise of Republican vice presidential candidate Sarah Palin's Yahoo! account.

One way to address the limitations of existing backup authentication mechanisms is to add new ones. Since no mechanism is completely secure, system designers must support configurations that require multiple authentication tasks be completed to authenticate. Can users comprehend such a rich set of new options? We designed two metaphors to help users comprehend which combinations of authentication tasks would be sufficient to authenticate. We performed a usability study to measure users' comprehension of these metaphors. We find that the vast majority of users comprehend screenshots that represent authentication as an exam, in which points are awarded for the completion of individual authentication tasks and authentication succeeds when an authenticatee has accumulated enough points to achieve a passing score.

General Terms

Authentication, Backup authentication, Password reset

Keywords

Authentication

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

1. INTRODUCTION

Backup authentication systems are used to verify the identity of users who are unable to perform primary authentication. For systems that use passwords as their primary authentication mechanism, backup authentication mechanisms guard the functionality used to reset passwords with new ones. Like all user authentication systems, backup authentication systems suffer from two well-known modes of failure: rejecting an authenticatee who is the legitimate account holder (a *false reject*, or *reliability* failure) or accepting an authenticatee who is impersonating the account holder (a *false accept*, or *security* failure).

The backup authentication systems for the four largest webmail services (AOL Mail, Gmail, Windows Live Hotmail, and Yahoo! Mail) implement two common authentication mechanisms: email-based authentication and personal authentication questions. In email-based authentication, the authentication system sends an email containing a code to an address configured by the user in advance. The user must provide that code, which is sometimes contained in a web link, as evidence of her identity. This approach is both popular and effective, but of limited use for users' primary email accounts; they may not have alternate addresses that they can access as reliably.

When a user is asked to answer a personal authentication question, she must provide the correct answer – one configured in advance – as evidence of her identity. We have previously studied the questions used by these webmail services and found they fall short in both reliability and security [9]. Roughly 20% of users forget their answers with six months, 17% of answers could be guessed by acquaintances whom account holders would not trust with their passwords, and 13% of answers could be cracked by guessing the five most popular responses for each question.

One way to reduce the likelihood of failure is to enable users to configure multiple authentication tasks chosen from a variety of different authentication mechanisms.[1] Each authentication task provides evidence that helps the overall authentication system to differentiate an account holder from an impersonator. Adding authentication tasks increases the potential pool of evidence available to the authentication

[1]Authentication that requires multiple tasks is often referred to as n-factor authentication, where n is the number of credentials (factors). We avoid this terminology as it presupposes a requirement of a fixed number of credentials regardless of the strength of each credential. We discuss tasks, rather than factors (credentials), as there needn't always be a one to one mapping between tasks and credentials.

system and thus reduces the risk of reliability failure: an account holder (and, to a lesser extent, an impersonator) will be more likely to identify tasks she can complete. One may make trade-offs between reliability and security by adjusting the *evidentiary requirements* for a given set of authentication tasks: the combinations of tasks that will be deemed sufficient to authenticate. Increasing the evidentiary requirements for authentication by requiring more demanding combinations of authentication tasks will reduce the likelihood of a security failure; an impersonator (and, to a lesser extent, the account holder) will be less likely to successfully authenticate if more tasks, or more difficult tasks, are required. So long as the legitimate account holder is more likely to succeed at each task than an impersonator, an iterative process of adding authentication tasks and increasing the evidentiary requirements has the potential to simultaneously reduce the risks of both security and reliability failures.

For example, some websites with high-value accounts, such as some banks, harden their authentication systems by requiring users to answer multiple personal authentication questions. In this case, there are multiple tasks (questions to answer) that employ the same authentication mechanism.

Different users will be best served by different choices of authentication tasks and evidentiary requirements. Some users will have accounts with little to protect (e.g. throwaway email accounts) and will want to spend the least time possible to configure backup authentication options. Users who do not have strong security requirements for their accounts, such as those who use their accounts only for backing up their music collections, may be opt for tasks and evidentiary requirements that maximize reliability. Users storing security- or privacy-critical information may prefer higher evidentiary requirements. Users with high reliability and security requirements will want to configure a large number of authentication tasks and impose strong evidentiary requirements.

While no one configuration can best meet all users' needs, most of today's website authentication systems take a one-size-fits-all approach. For example, the authentication systems for the four largest webmail services all offer only two authentication tasks: answering a single personal authentication question or requesting an email-based authentication process (see [3]) in which the service sends the user an authentication code by email. These services do not allow users to increase the evidentiary requirements of backup authentication by requiring that both tasks be completed to authenticate. Because so many websites rely on email addresses as a backup authenticator, the security and reliability of the mechanisms used by these webmail providers are especially critical.

One hurdle to enabling users to add authentication tasks and increase evidentiary requirements is that they must be able to comprehend, and possibly specify, these evidentiary requirements. If authentication requirements are presented in a manner that account holders cannot comprehend, they will be unable to make informed risk decisions about how to use their accounts. When authentication requirements are stricter than the user believes them to be she will feel betrayed when, after performing tasks she believed provide sufficient evidence to prove her identity, she is still unable to access her account; she will have believed the authentication process was more reliable than it actually was. When authentication requirements are weaker than the user believes them to be, she will feel betrayed when the system provides access to an imposter who provided less evidence than she believed would be required to change her password; she will have believed the authentication process was more secure than it actually was.

To examine whether it is possible to scale the number of authentication options without negatively impacting comprehension of evidentiary requirements, we created two metaphors with which to represent these requirements.

The *exam* metaphor associates each authentication task with a number of points awarded for completing it. Authentication requires a passing score: ten points in our implementation. We selected the exam metaphor because we believed it would be familiar to many users—quizzes, tests, and other examinations using such points are used across educational levels and cultures.

The *evidence scale* metaphor groups authentication mechanisms into three buckets: those deemed hardest for impersonators to complete provide *strong evidence* of an account holder's identity; the next hardest provide *medium strength evidence*; and the remaining (easiest) tasks provide *weak evidence*. Authentication in our implementation of the evidence scale model requires completion of either two tasks when one provides strong evidence or both provide medium strength evidence, and three tasks otherwise. We selected this metaphor because it required no math and would limit users to seemingly tractable combinations of authentication tasks—any three credentials would be sufficient to authenticate. The evidence scale metaphor cannot express all the combinations that can be represented by the exam metaphor, which in turn cannot express all the combinations that could be represented by boolean algebra.

To test comprehension of these metaphors we performed a paper-based in-laboratory survey. Participants were shown screenshots of interfaces based on these new metaphors, as well as a screenshot of the current Windows Live ID password-reset configuration page, and asked questions to test their comprehension. We found that our participants were at least as able, if not better able, to comprehend complex configurations presented with the exam metaphor as they were able to comprehend the two authentication tasks currently supported by Live ID.

2. BACKGROUND AND RELATED WORK

The personal authentication questions and email-based authentication mechanisms used for backup authentication by the top four webmail services are becoming increasingly inadequate. In part, this is a consequence of the success of those services. The usability of web-based mail has approached, and by some accounts surpassed, that of client-based email. Webmail services are enhancing their offerings to work even when users are offline [6]. Those users who now rely on webmail for their primary email accounts may not have alternate email addresses to use for backup authentication. Those who previously configured alternate email addresses may no longer be associated with the ISP, employer, school, or other organization that had provided the listed address.

Both the security and reliability of personal authentication questions have received increasing scrutiny, especially following the compromise of Republican vice presidential candidate Sarah Palin's personal Yahoo! account via her

question, which asked where she met her spouse [2, 4]. The press has not only covered the weakness of personal authentication questions, but also their failure in helping legitimate users to recover their accounts [12].

Quantitative studies on the security and reliability of personal authentication questions were first performed by Zviran and Haga in 1990 [14] and later by Podd *et al.* [7]. Both studies found that roughly 20% of answers are forgotten within three months and that close friends or significant others can guess over 30% of answers. More recently, Ariel Rabkin attempted to categorize questions by potential weaknesses [8].

Our recent work has shown that these security and reliability problems remain in the personal authentication questions in use today by the top four webmail providers, and that these questions remain guessable even when an attacker isn't the user's significant other or close friend [9]. We conducted a laboratory study of 65 pairs of participants (130 total). We asked participants to answer all of the personal authentication questions used by AOL, Google, Microsoft, and Yahoo!. Of those participants who arrived with a partner they wouldn't trust with the Live ID (Hotmail) password, 17% of answers could be guessed by their partner. There was a strong correlation between the memorability of questions and the likelihood that they could be guessed. Furthermore, it was possible to guess 13% of all answers by iterating through the five most popular answers for each question.

As different individuals have different capabilities, a broader choice of authentication mechanisms should allow authentication systems to better serve their users. Jakobsson *et al.* have created and tested mechanisms that use a series of multiple choice questions about users' preferences to authenticate them [5]. In recent work we have studied how user-selected trustees could assist in backup authentication [10], an approach previously envisioned for primary authentication by Brainard *et al.* [1]. Other possible backup authentication mechanisms include SMS messages sent to mobile phones [13], single use password sheets, and recall of previously used passwords (which may be especially useful when a user forgets a new password shortly after changing it).

To our knowledge, no existing work has addressed the question of how to convey an array of authentication options to users such that they can comprehend which combinations are sufficient to authenticate.

3. METHODOLOGY

We conducted an in-laboratory paper-based survey in which participants were asked demographic questions and then shown five screenshots of web forms used to configure backup authentication (password reset). After each screenshot we asked questions designed to gauge their comprehension of the configuration form depicted therein. We used a within-participants design, so all participants answered the same questions about all five of these screenshots. The screenshots were configured for the account of a fictional user named Jane Doe.

3.1 Screenshots presented in survey

The current Windows *Live ID* password reset settings form (Figure 1) served as our baseline form. The rest of the web forms appeared to be used to configure similar backup authentication settings for "SplendMail", a fictional webmail

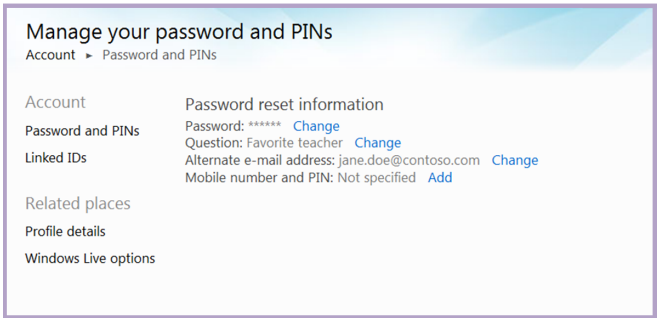


Figure 1: The Windows Live ID password reset settings form.

service which we presented as if it were a real product. The simplest SplendMail screenshots presented a short form that used the point-based exam metaphor.

The Live ID form and short exam form had the same authentication mechanisms configured: both had a personal question (*favorite teacher*) and email address (*jane.doe@contoso.com*). We did not configure mobile phone numbers because, despite appearances, mobile phone numbers cannot actually be used to reset Live ID passwords.

We presented two screenshots of the short exam form: in *short exam P5* each authentication task was worth five of the ten points needed, so both tasks would be required to authenticate (see Figure 2); in *short exam P10* each task was worth ten points, so either task would be sufficient to authenticate (see Figure 3).

A *longer exam* form featured five authentication tasks configured from six possible authentication mechanisms (Figure 4). Tasks were worth between three and seven points, such that some combinations of two tasks would be sufficient to authenticate but others would not be.

Finally, an *evidence scale* form contained five authentication tasks – two strong, one medium, and two weak – also chosen from six possible mechanisms (Figure 5). As with the longer exam form, some combinations of two authentication tasks in the evidence scale form were sufficient to authenticate whereas others were not. Two authentication tasks in the evidence scale form were classified as providing strong evidence, one as providing medium strength evidence, and two as weak evidence. We did not include a short evidence scale form as there would have been more strength levels than authentication tasks. Because the evidence scale was also presented as part of SplendMail, we described it as an interface from an earlier version of the product.

3.2 Questions accompanying screenshots

All forms were followed by questions designed to gauge participants' *evidentiary requirements comprehension*: their ability to understand which combinations of authentication tasks would be sufficient to authenticate and which would not be. The simplest way to measure comprehension of evidentiary requirements when only two authentication tasks have been configured is to ask whether one is enough or if both are necessary. We asked a *one-or-both* question for both the Live ID and short exam P5 screenshots. In retrospect we should have also asked this question for short exam P10, but we did not.

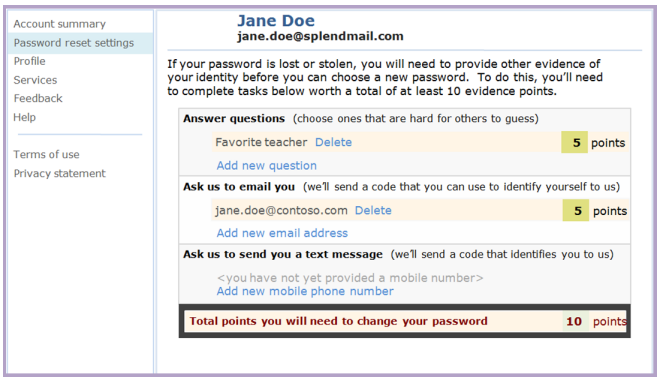


Figure 2: The short exam P5 form for SplendMail.

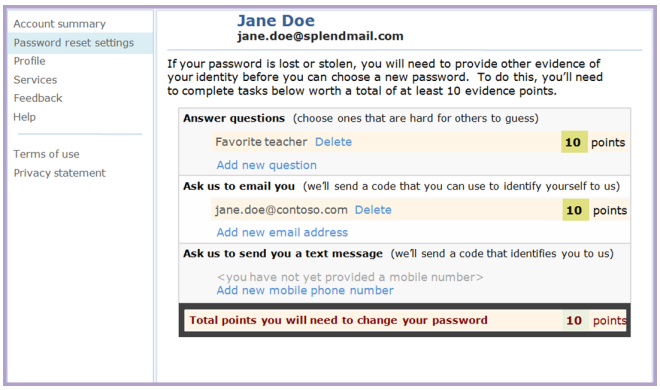


Figure 3: The short exam form P10 for SplendMail.

> To change her password, will `[Windows Live|SplendMail]` require Jane to establish her identity using both the e-mail address and the question, or is one of the two enough?

The answer options were *one*, *probably one*, *not sure*, *probably both*, and *both*.

When more than two authentication tasks have been configured, one may gauge comprehension of evidentiary requirements by presenting sample combinations of authentication tasks and asking whether these combinations would be sufficient to authenticate. We asked these *sample combination* questions for all five forms, including Live ID and short exam P5 (essentially asking the same comprehension question twice, since the one-or-both question asked for the same information). The questions differed only in the name of the service (Windows Live or SplendMail) and the bullet points that followed to identify the set of authentication tasks.

> Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to `[Windows Live|SplendMail]`?

Each question was followed by five options: *yes*, *probably*, *not sure*, *probably not*, and *no*.

We were also curious as to whether Live ID users understood how the authentication mechanisms worked; the configuration form shows only what information is configured–not how it is used. We had designed all of our exam and scale forms to explain how the authentication mechanisms worked and wanted to see if doing so was worthwhile. We

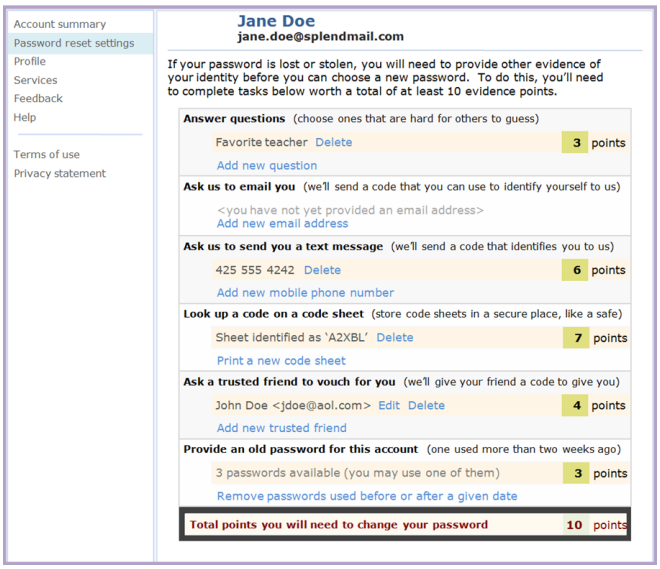


Figure 4: The longer exam form for SplendMail.

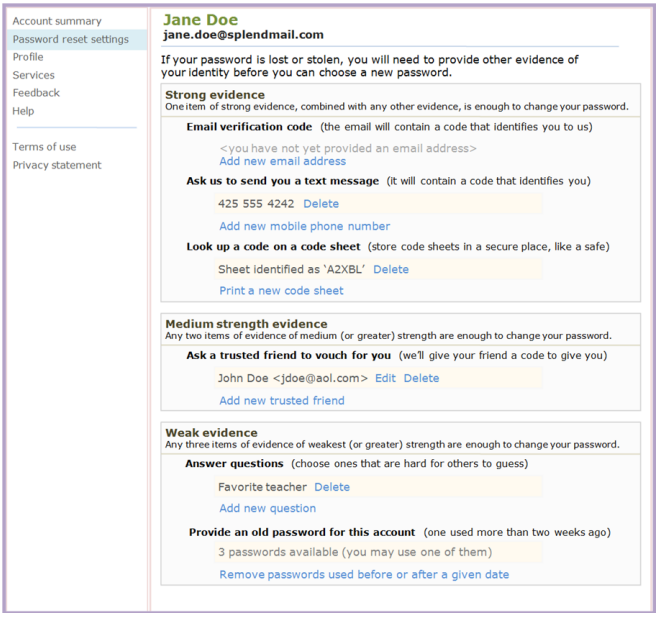


Figure 5: The evidence scale form for SplendMail.

thus asked *mechanism comprehension* questions to gauge how well users understood these mechanisms. For example, we asked whether the task associated with the alternate email address was to type in the alternate email address or to receive (and provide) a code sent to the address.

> How does `[Windows Live|SplendMail]` use Jane's question ("favorite teacher") and answer ("Mrs. Smith")?
>
> A Windows Live asks the question "favorite teacher" and verifies that the response is "Mrs. Smith".
>
> B Windows Live presents the name "Mrs. Smith" and asks Jane to identify the question that describes her relationship to Mrs. Smith (that she is Jane's "favorite teacher").

How does [Windows Live|SplendMail] use Jane's alternate e-mail address (jane.doe@contoso.com)

A Windows Live asks "what is your alternate e-mail address" and verifies that the response is "jane.doe@contoso.com".

B Windows Live sends an email to jane.doe@contoso.com containing instructions and a code (in the form of a web link) used to identify Jane.

The options for answering these two mechanism comprehension questions were *definitely A*, *probably A*, *not sure*, *probably B*, and *definitely B*.

Unless otherwise specified, all questions in the survey were graded on a zero-centered five point (i.e., integers from -2 to 2) scale. For example, if the correct answer to a question was *yes*, a participant would receive 2 points *yes*, 1 for *probably yes*, 0 for *not sure*, -1 for *probably not*, and -2 for *no*.

3.3 Survey section ordering

In designing our survey, we took steps to guard against two kinds of ordering effects. First, we anticipated that some participants who saw the Live ID form first might use what they learned to help them answer some of the short exam questions, and vice versa. Second, we anticipated that for the mechanism comprehension questions, participants might opt for the first response presented to them. To guard against these ordering effects, we had two forms, Survey Form I and Survey Form II, in which the order of the sections for the first two screenshots – Live ID and short exam P5 – was swapped and the order of the response options to the mechanism comprehension questions was also swapped. Otherwise, we used fixed ordering of the sections and of the questions within each section.

Nine users (seven of whom were Live ID users) received Survey Form I and the other nine (five of whom were Live ID users) received Survey Form II.

There was the possibility of additional ordering effects that other design constraints prevented us from guarding against. When these effects might have impacted results, we chose the most conservative ordering: the one biased (if at all) against our hypotheses. Thus, for example, all questions related to the evidence scale form were last, so that participants would have learned as much as they could about how to think about authentication options and how to complete our survey before getting to the evidence scale form. Learning time (during which participants presumably are more likely to make errors) would have been spent on the exam forms, giving an advantage to the evidence scale form. So, if the performance on the exam form was as good or better than performance on the evidence scale form, we can conclude that the exam form is at least as good as the evidence scale form. This particular ordering decision is relevant to our hypothesis H4, discussed below.

3.4 Participants

We recruited 18 total participants at two educational levels: eight had a college degree and ten had two years or less of post-secondary education. One of the latter participants had not completed high school. Because our goal is to enable the broadest possible range of users to comprehend and configure authentication options, we wanted to test whether users at both of these educational levels could perform the mental arithmetic required by our exam metaphor. To ensure that education would not be a proxy for age, we recruited participants between the ages of 30 and 49; actual ages ranged from 30 to 48. Seven of our participants were female and 11 were male. Twelve of the 18 were Live ID users; the other six had no Live ID account.

3.5 Procedure

When participants arrived at our laboratory, we randomly assigned them to a survey form[2] and had them sit down at a desk to work. We presented the survey to them one section at a time so that they could not return to work on previously completed sections. We allowed participants to ask us questions to clarify the wording in the survey, but we did not address questions about the user interfaces themselves. Upon completing the survey, participants were compensated with their choice of a software gratuity or a $50 gift card.

3.6 Hypotheses

We approached this work with five hypotheses about how well users would understand and like the various password reset settings forms we presented. We introduce the hypotheses here and report on our tests of the hypotheses in Section 4.

H1: When presented short exam P5, which describes how each authentication mechanism will be used, Live ID users are better able to comprehend the use of these mechanisms than when presented with Live ID's password-reset settings form.

	Q#	*Mech. comprehension question*
Live ID	22	How is secret question used?
	23	How is email address used?
Short Exam P5	30	How is secret question used?
	31	How is email address used?

Hypothesis: The average scores of Live ID users on the mechanism comprehension questions for the short exam P5 screenshot are greater than the average scores for the Live ID screenshot.

The screenshots for the exam and evidence scale metaphors contained text that attempted to explain the task required to satisfy each authentication mechanism. For example, personal authentication questions were under the heading "Answer questions" and the email-based authentication heading indicated that an email would "contain a code that identifies you to us".

We asked users the two mechanism comprehension questions following both the Live ID screenshot (questions 22 & 23) and the short exam P5 screenshot (questions 30 & 31). These questions examined how well users understand how personal authentication questions and email-based authentication work. We examined the results using only participants who were Live ID users, as Live ID may rely on users to learn how these mechanisms work through interfaces other than the screenshot presented. Participants were encouraged to draw upon any existing experience they had with Windows Live when answering questions about Live ID.

[2]When possible, we paired demographically similar participants and randomly assigned each to a *different* survey form.

H2: Live ID users comprehend the evidentiary requirements of authentication in the short exam form as well as they do for Live ID's current password reset settings form.

	Q#	*Question*
Live ID	26	one task or both?
Short Exam P5	32	one task or both?

Hypothesis 2a: The average scores of Live ID users on the set of one-or-both question about the short exam P5 screenshot are greater than the average scores on the same questions about the Live ID screenshot.

	Q#	*Task 1*	*Task 2*
Live ID	27	question	
	28	email	
	29	question	email
Short Exam *P5*	33	question	
	34	email	
	36	question	email

Hypothesis 2b: The average scores of Live ID users on the set of sample combination questions about the short exam P5 screenshot are greater than the average scores on the same questions about the Live ID screenshot.

	Q#	*Task 1*	*Task 2*
Live ID	27	question	
	28	email	
	29	question	email
Short Exam *P10*	38	question	
	39	email	
	40	question	email

Hypothesis 2c: The average scores of Live ID users on the set of sample combination questions about the short exam P10 screenshot are greater than the average scores on the same questions about the Live ID screenshot.

		Q#	*Task 1*	*Task 2*
Live ID		27	question	
		28	email	
		29	question	email
Short Exam	P5	33	question	
		34	email	
		36	question	email
	P10	38	question	
		39	email	
		40	question	email

Hypothesis 2d: The average scores of Live ID users on the set of sample combination questions about the short exam P5 and P10 screenshots are greater than the average scores on the same questions about the Live ID screenshot.

We asked participants to answer one-or-both questions to examine their comprehension of the evidentiary requirements to authenticate when two authentication tasks were configured: a personal authentication question and email-based authentication (hypothesis 2a). We examined the 12 responses of participants who were Live ID users because they were already relying on Live ID's behavior to match their expectations: if they believed both authentication tasks were required to authenticate then Live ID would not be providing the protection they expected.

We also generated a mean score for each user's responses to the three sample combination questions, which were also used to test comprehension of evidentiary requirements. The first instance of this question was followed by a single bullet item for the personal authentication question, the second also a single bullet for alternate email address, and the third containing both bullet points. We calculated the average of these three sample combination scores for the Live ID, short exam P5 (hypothesis 2b), and short exam P10 (hypothesis 2c) screenshots. We also took an average over both short exam screenshots (hypothesis 2d).

H3: Comprehension of the exam metaphor decreases as more authentication mechanisms are configured.

		Q#	*Task 1*	*Task 2*
Short Exam	P5	33	question	
		34	email	
		36	question	email
	P10	38	question	
		39	email	
		40	question	email
Longer Exam		43	question	text msg
		44	code sheet	question
		45	question	old pswd

Hypothesis: The average scores on questions about the three most difficult sample combination questions on short exam screenshots P5 and P10 are greater than those for the three most difficult sample combination questions about the longer exam.

Each additional authentication task a user configures increases the number of potential combinations that may or may not be sufficient to authenticate. We wondered whether comprehension of evidentiary requirements would decrease as the number of authentication tasks increased. As the longer exam was presented after the short exam screenshots, we considered that participants' increased experience with the exam metaphor might counteract the effects of complexity.

We compared the average score on the three sample combination questions asked in both short exam P5 and P10 (six total question instances) with the average score on the three most challenging questions in the longer exam. The longer exam contains sample combination questions with one, two, and three authentication tasks. To predict which questions on the longer exam would be the most challenging, we examined them based on the number of authentication tasks in the sample combinations.

Sample combinations that contained only a single authentication task were all insufficient to authenticate, and so we expected these questions to be easy. Indeed, we would find that only one participant failed to answer *no* to both of these questions.

Two questions presented combinations of three authentication tasks, which were always sufficient. The only participant who failed to answer both correctly was the one who was also unable to answer the questions about single task combinations correctly. One trick question examined whether participants were reading specific instructions about an authentication mechanism. Again, we would find that only one participant failed to answer correctly.

The remaining three questions all featured two authentication tasks and did not share a common correct answer–two were insufficient and one was sufficient. We predicted (correctly) that these would pose the most difficulty and thus used the mean responses to these questions to calculate the sample combination score for the longer exam. These were the questions used to evaluate participants' performance on the longer exam when comparing to the short exam to test this hypothesis.

H4: The evidence scale form, which does not require mental math, is more comprehensible than the exam form, which does.

	Q#	*Task 1*	*Task 2*	*Task 3*
Longer exam	41	question		
	42	code sheet		
	43	question	text msg	
	44	code sheet	question	
	45	question	old pswd	
	46	question	trustee	
	47	question	text msg	old pswd
	48	question	old pswd	old pswd
Evidence Scale	49	question		
	50	code sheet		
	51	question	text msg	
	52	code sheet	question	
	53	question	old pswd	
	54	question	trustee	
	55	question	text msg	old pswd
	56	question	old pswd	old pswd

Hypothesis: For the sample combination questions common to both the longer exam form and evidence scale form, the average participant scores are higher when these questions are asked about the evidence scale form than when they are asked about the longer exam form.

Given that usability testing often results in discoveries that users often cannot perform tasks that designers assume they can, we were concerned that it may be too optimistic to rely on users to perform mental addition. We chose the evidence scale metaphor because it could accommodate a large number of authentication tasks without requiring mental math. Because the evidence scale metaphor form cannot be scaled down to a short form, participants did not build experience with it as they did with the exam metaphor. We thought that if the evidence scale metaphor was sufficiently superior in its comprehensibility that it might still perform significantly better than the exam metaphor. What's more, because the evidence scale form came last, participants would have the most experience understanding the nature of the survey and the sample combination questions we used to gauge their comprehension of evidentiary requirements.

H5: Users prefer the exam form to the evidence scale form, or vice versa.

Our last hypothesis was that users might have a preference for using the exam form (which they knew as the "new" SplendMail interface) or the evidence scale form (the "old" SplendMail interface). We asked users which they preferred.

4. RESULTS

Our results for questions used to test hypotheses are shown in Table 1. Each row represents a single survey question, and the rows are grouped by the five screenshots presented in the survey. Each numbered column contains a participant's scores for these questions. All questions were followed by five options which were converted to numerical values ranging from -2, for the least correct answer, to 2, for the most correct answer. For example, if the correct answer to a yes or no question was *yes*, that answer received 2 points, *probably* 1, *not sure* 0, *probably not* -1, and *no* -2.

The correct answer to each question is given in the column labeled *Ans*. The *Avg* column contains the mean score for each question. A score of zero is expected if answers are chosen at random.

The mean scores for each question show a striking pattern of generally correct responses to evidentiary requirements comprehension questions on the exam metaphor. The mean scores for all questions about all of the exam screenshots (short exam P5, short exam P10, and long exam) are positive and all but three are above 1.5. Even for the long exam, for which we asked the most complex questions, average scores are all above 1.5. These results are in contrast to average scores for the Live ID screenshot, for which, of three scores, one is negative and one is effectively zero.

Average scores for the evidence scale screenshot, like those for the exam screenshots, indicate high levels of correct responses. All average scores for the evidence scale are positive, and all but one are above 1.0.

We used statistical inference testing to test each of our five hypotheses. We used nonparametric tests, the Wilcoxon signed ranks test and the binomial test, which are safer than their parametric counterparts because they make no assumption about the underlying distribution of the data. Since we ran nine statistical significance tests, we corrected for multiple testing by adjusting our significance level α from 0.05 to 0.028 according to the Benjamini-Hochberg method. We now address results of testing for each hypothesis.

H1: When presented short exam P5, which describes how each authentication mechanism will be used, Live ID users are better able to comprehend the use of these mechanisms than when presented with Live ID's password-reset settings form.

For the 12 participants who were Windows Live ID users, the mean mechanism comprehension score for the Live ID screenshot was 1.17 (s.d. 1.00) vs. 1.50 (s.d. 0.74) for exam P5. A Wilcoxon signed ranks test did not find a significant difference for this sample: $Z = -.946, p = .344$.

H2: Live ID users comprehend the evidentiary requirements of authentication in the short exam form as well as they do for Live ID's current password reset settings form.

The mean scores of the 12 participants who were Live ID users on the one-or-both question for the Live ID screenshot was -1.08 (s.d. 1.25) vs. 1.25 (s.d. 1.35) for exam P5. The Wilcoxon signed ranks test indicates the difference is strongly significant: $Z = -2.716, p = .007$.

The mean scores of the 12 participants who were Live ID users on the the sample combination questions – also used to measure comprehension of evidentiary requirements – were .28 (s.d. 1.05) for the Live ID screenshot and 1.36 (s.d. 0.74) for exam P5. The Wilcoxon signed ranks test was again significant: $Z = -2.283, p = 0.022$. However, these tests are insufficient to prove our hypothesis as users may have been predisposed to believe that both authentication tasks would be required to authenticate, thus favoring exam P5 (for which the correct answer was *both*) over Live ID (for which it was not).

Alas, we had not asked the one-or-both question for exam P10. The mean score of the 12 participants who were Live ID users on the sample combination questions for exam P10 was 0.86 (s.d. 1.42), and while this was higher than the mean for Live ID the difference was not significant: $Z = -1.431, p = 0.153$. However, there is reason to believe

Table 1: All questions used to test our hypotheses (rows) and participants' scores on each of them. Columns represent participants, who are identified by number. Boldface numbers indicate participants were Live ID users. The column labeled *Ans* contains the correct answer to each question. The column labeled *Avg* contains the mean score of all participants for each question. Columns labeled *Task* represent a sample combination of authentication tasks. All questions were scored on a zero-centered five-point (-2 to 2) scale.

Screenshot	Q#	Task 1	Task 2	Task 3	Ans	Avg	Participants assigned Survey Form A									Participants assigned Survey Form B									
							1	**2**	**3**	4	**5**	6	**7**	**8**	**9**	**10**	11	12	**13**	**14**	15	**16**	**17**	18
One-or-both questions used to gauge comprehension of evidentiary requirements for authentication																									
Live ID	26				one	-1.17	-2	1	0	-2	-2	-2	-2	-1	1	-2	-2	-1	-2	0	-1	-2	-2	0
short exam A	32				both	1.44	2	2	2	2	2	2	2	1	1	2	2	2	2	-1	2	-2	2	1
Sample combination questions used to gauge comprehension of evidentiary requirements for authentication																									
Live ID	27	question			yes	0.06	2	1	0	2	-2	2	-2	2	0	-2	-2	-1	-2	2	1	2	-2	0
	28	email			yes	-0.67	-2	1	0	2	-1	-1	-2	1	-1	-2	-2	-1	-2	0	-2	2	-2	0
	29	question	email		yes	1.61	2	2	0	2	2	2	2	2	-1	2	2	2	2	2	2	2	2	0
short exam P5	33	question			no	1.78	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	0	1	2
	34	email			no	0.94	-1	2	2	2	2	2	-2	2	1	2	2	2	2	-1	2	-2	-2	2
	36	question	email		yes	2.00	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
short exam P10	38	question			yes	0.72	1	2	2	2	2	-2	2	2	-1	-2	2	2	2	2	2	-1	-2	-2
	39	email			yes	0.78	1	2	2	2	2	-2	2	2	-2	-2	2	2	2	2	2	1	-2	-2
	40	question	email		yes	1.28	1	2	-1	2	2	2	2	2	1	2	2	2	2	2	1	-1	-2	2
longer exam	41	question			no	2.00	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	42	code sheet			no	1.83	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	-1	2	2
	43	question	text msg		no	1.56	2	2	2	-2	2	2	2	2	2	2	2	2	2	2	2	-2	2	2
	44	code sheet	question		yes	1.83	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	-1	2	2
	45	question	old pswd		no	1.72	2	2	2	2	2	2	2	2	2	2	2	2	-2	2	2	1	2	2
	46	question	trustee	old pswd	yes	1.83	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	-1	2	2
	47	question	text msg	old pswd	yes	1.94	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2
	48	question	2 old pswds		no	1.94	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2
evidence scale	49	question			no	1.72	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	-2	2	2
	50	code sheet			no	0.72	-2	-2	2	-2	2	2	2	-2	2	2	2	2	2	2	2	-1	-2	2
	51	question	text msg		yes	1.78	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	-2	2
	52	code sheet	question		yes	1.83	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	-1	2	2
	53	question	old pswd		no	1.44	1	2	2	2	2	2	-2	2	-2	2	2	2	2	2	2	1	2	2
	54	question	trustee	old pswd	yes	1.17	-1	1	2	2	-2	2	2	2	2	2	2	2	2	2	2	-1	2	-2
	55	question	text msg	old pswd	yes	1.89	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2
	56	question	2 old pswds		no	1.17	1	2	2	2	2	2	-2	-2	2	2	2	2	1	2	-2	1	2	2
Mechanism comprehension questions																									
Live ID	22	How to authenticate using question			A	1.78	1	2	2	2	2	2	2	2	2	2	2	2	2	2	1	0	2	2
	23	How to authenticate using email			B	0.5	-2	1	2	-1	2	1	2	2	2	2	2	2	-2	-2	-1	2	-2	-1
short exam A	30	How to authenticate using question			A	1.78	2	1	2	1	2	2	2	2	1	2	2	2	2	2	1	2	2	2
	31	How to authenticate using email			B	1.06	2	2	2	-1	2	2	2	1	-1	2	2	2	1	1	-1	2	-2	1

these understate the efficacy of the exam metaphor. In post-survey interviews, we found that more than one participant failed to notice that the exam P10 screenshot was different from exam P5, and thus missed two of three questions by providing the correct answers for exam P5.

The mean of the three sample combination questions over both of the two short exam screenshots (P5 and P10) was 1.11 (s.d. 0.94), significantly higher than the mean for the three questions asked for Live ID (provided above): $Z = -2.197, p = 0.028$.

We had not originally planned to compare the aggregate score on the full exam screenshot to Live ID—we thought it unfair to compare metaphors using screenshots in which one implemented a richer and more complex set of authentication combinations than the other. Yet, despite the difference in complexity and users' lack of real-world experience with the exam metaphor, the 12 Live ID users received a mean comprehension score of 1.67 (s.d. 0.83) on the the three most difficult questions about the longer exam screenshot. The mean score for these participants over all the sample combination questions about the longer exam was 1.79 (s.d. 0.58). The difference in the mean scores for the hardest questions about the exam and the mean scores and the three simple questions about the Live ID screenshot is statistically significant—in favor of the exam: $Z = -2.294, p = .022$.

H3: Comprehension of the exam metaphor decreases as more authentication mechanisms are configured.

Fifteen of our 18 participants – including the one who had not completed high school – answered all eight questions on the long exam screenshot perfectly. Of the remaining three participants, two missed only one question.

Recall that we had guessed that the three questions that contained sample combinations of two tasks would be the most difficult for participants on the longer exam screenshot. Indeed, no other questions in this section proved more challenging. Of those three questions, two received 16 correct answers from 18 participants and one received 17 correct answers. In other words, only 5 of 54 total responses were imperfect (either incorrect or not provided with full confidence).

Three of those five imperfect responses came from a single participant (#16), who appeared to have difficulty throughout the survey. One of the two remaining errors was made by a participant who made a practice of copying point values from the exam screenshot to the bullet points on the question. That participant had copied a point score from the wrong authentication task from the screenshot to the paper. The final error appeared on the one question in which the two authentication tasks were worth a total of nine points, one short of the sufficiency threshold of ten.

The mean score on these three "difficult" questions was 1.704 (s.d. .731). This compares favorably to the mean score of 1.25 (s.d. .864) on the simpler questions for the two short exam forms. Not only was there no evidence of a decrease in comprehension, but the learning effect might well have increased comprehension. This requires further investigation as it fell short of the significance threshold: $Z = -1.855, p = .064$.

H4: The evidence scale form, which does not require mental math, is more comprehensible than the exam form, which does.

All of the questions posed for the exam screenshot were also asked for the evidence scale screenshot. Thus, we could compare performance across all eight of the common questions. We expected participants to be more confident in their answers for the exam metaphor, which they had experience with, than for the evidence scale metaphor, which they did not. To reduce this effect, we treated answers containing 'probably' as if they had been given with full confidence by using a zero-centered three-point scale (-1 to 1).

The mean score on these eight questions for the exam form was 0.917 (s.d. .242), significantly higher than the mean of .75 (s.d. .271) for the same questions when asked for the evidence scale: $Z = -.2521, p = .012$. We reject this hypothesis in favor of its opposite: the exam form is more comprehensible than the evidence form as presented by our survey.

H5: Users prefer the exam form to the evidence scale form, or vice versa.

Of 18 participants, two expressed no preference between the forms. Thirteen of the 16 who expressed their preference favored the exam form, two of whom conditioned their preference as slight. One of the three who preferred the evidence scale form conditioned the preference as slight. We ignored the slight conditions to group those who had expressed preferences into two categories and performed a two-tailed binomial test. The preference for the exam form was significant, with $p = .021$.

During a post-survey interview with the one participant who had struggled with questions on all forms (#16), a researcher asked his preference between the exam form and the evidence scale form. The interview question was posed, in part, to gauge whether the participant might have simply been answering questions at random. (Though he was a high school graduate, we were concerned the survey might have been above his reading level.) The participant, who had expressed a preference for the exam form on the survey, switched to a preference for the evidence scale metaphor during the interview. We are not confident that the opinion expressed in the interview is even a valid data point; it followed interview questions in which the participant was asked to explain the exam form as best as he could. If the participant did indeed prefer the evidence scale metaphor, our binomial test would not have been significant, with $p = .077$, whereas disregarding this participant's response entirely yields $p = 0.035$.

Our H5 result should be interpreted with some caution. Because we presented the exam form as the "new" SplendMail interface, it's possible that participants assumed that a newer interface must be better. It's also possible they realized SplendMail was a fictional product, assumed that the new interface was the more recent development of the researchers, and stated their preference because they believed the researchers would like their latest development to perform the best. We had considered introducing a new fictional product for this interface, but that presented additional confounding factors. In retrospect, we should have randomly assigned which interface was presented as new to each participant.

5. DISCUSSION

Our results indicate that the exam form was comprehensible and remained so when scaled to many more authentication tasks than are configurable in today's backup authentication systems. From our results, the exam form seems to be a certain win; not only does it enable configuring authentication combinations that are not possible in the Windows Live ID form, but users actually understand it. Moreover, they appear to like it, based on their preference for it over the ostensibly simpler evidence scale form.

Nevertheless, we should consider some of the limitations of our methodology. These results do not guarantee success for authentication systems that are configured based on the exam metaphor. There are many factors that might cause participants to perform differently when role-playing on a paper survey than they would under real-world conditions [11].

One limitation of the exam metaphor is that it cannot express all possible authentication sufficiency requirements. For example, assume our fictional Jane Doe has configured four authentication tasks. Two of these, tasks A and B, could likely be completed by her on-and-off romantic partner. Both of Jane's two other tasks, C and D, could likely be completed by her occasionally disgruntled brother. To protect herself if either of these individuals is acting alone, Jane might wish to require that one of the first two tasks (A or B) and one of the latter two tasks (C or D) be completed in order to authenticate $((A \vee B) \wedge (C \vee D))$. This requirement cannot be expressed in the exam metaphor.

One feature of the exam metaphor is that the system or users themselves may choose different sufficiency thresholds (required score totals) for different authentication situations. For example, one could require ten points to reset the password on an active account, fifteen points to add or modify the authentication configuration, but only five points to reset the password on an account that had been inactive for more than two weeks. The only part of the interface that need change to support this is the choice of totals at the bottom.

The exam metaphor may also be valuable in helping to recover compromised accounts. If an account holder and an impersonator are competing for ownership of an account, the system provider could use the last known-good copy of the exam and return the account the the individual who performs best on the exam. The exam is, after all, a user-generated test of her own identity.

6. FUTURE WORK

We did not study how points would be assigned to authentication tasks in the exam metaphor. We have only focused on whether users would comprehend the decisions that have been made, regardless of who has made them. If users were to assign points themselves, they might do so in ways that some might deem recklessly insecure (too easy to authenticate to) or so paranoid as to make authentication excessively unreliable (too difficult to authenticate to). To help users make better decisions, the authentication system may suggest scores based on its estimates of the security and reliability of individual authentication tasks. The authentication system could tune scores for different threats by asking the user to provide additional information. For example, the system could ask the user to estimate the number of individuals who might know the answer to a personal authentication question or who might borrow the users' mobile phone. Finally, authentication systems could provide feedback that helps users assess the security of an authentication configuration and estimate the likelihood that she will be able to successfully authenticate if she needs to.

Our recent discoveries on the weaknesses of today's backup authentication mechanisms suggest that many webmail users should add authentication tasks and increase the evidentiary requirements as new authentication mechanisms are made available [9]. Compelling users to take such action will be challenging. Most backup authentication mechanisms are configured when users create their accounts—the moment at which users have the least invested in these accounts. There may be no point at which users notice that their gradually increasing reliance on their accounts is no long proportionate to their investments in the security and reliability of these accounts. New research is needed how to best *prod* users in security contexts: maximizing action among users who would feel compelled to act if better informed while minimizing the collective annoyance experienced by those who would deem action unnecessary.

7. CONCLUSION

Given the plethora of results in the security usability literature that show what users cannot do, we approached the problem of increasing comprehension of the evidentiary requirements of authentication systems with trepidation. If users cannot understand whether one or both of two tasks is required to authenticate, how could they be expected to understand which of five tasks would be sufficient? User authentication is, after all, a complex process.

On the other hand, user authentication is just a technical term for an examination designed to test a user's identity and examinations are a familiar concept. We found the examination metaphor extremely effective for improving comprehension of the evidentiary requirements of authentication: 15 of our 18 participants answered all eight questions about the exam metaphor perfectly. Only one of the 18 participants missed more than one question. This compares most favorably to the existing interfaces the exam is designed to replace. Moving to an interface based on the exam metaphor may thus make it possible to simultaneously broaden users' authentication options while increasing their comprehension of how these options work together.

NOTE ON THE APPENDIX

While we have written our paper to be self contained, we have attached form I of our survey instrument – warts and all – to allow full scrutiny should the reader have questions that we have neglected to answer in the main text. We have annotated most questions with counts of the number of participants who answered each option. The counts accompanying correct answers are placed in boldface.

8. REFERENCES

[1] J. Brainard, A. Juels, R. L. Rivest, M. Szydlo, and M. Yung. Fourth-factor authentication: somebody you know. In *CCS '06: Proceedings of the 13th ACM Conference on Computer and Communications Security*, pages 168–178, New York, NY, USA, 2006. ACM.

[2] T. Bridis. Hacker impersonated Palin, stole e-mail password, Sept. 18, 2008. Associated Press.

[3] S. L. Garfinkel. Email-based identification and authentication: An alternative to PKI? *IEEE Security and Privacy*, Oct. 4, 2003.

[4] N. Hines. Sarah Palin's private e-mail account accessed by hacking group anonymous, 18, 2008. http://www.timesonline.co.uk/tol/news/world/us_and_americas/us_elections/article4780133.ece.

[5] M. Jakobsson, E. Stolterman, S. Wetzel, and L. Yang. Love and authentication. In *CHI '08: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, pages 197–200, New York, NY, USA, 2008. ACM.

[6] A. Palay. New in labs: Offline gmail. http://gmailblog.blogspot.com/2009/01/new-in-labs-offline-gmail.html.

[7] J. Podd, J. Bunnell, and R. Henderson. Cost-effective computer security: Cognitive and associative passwords. In *OZCHI '96: Proceedings of the 6th Australian Conference on Computer-Human Interaction (OZCHI '96)*, page 304, Washington, DC, USA, 1996. IEEE Computer Society.

[8] A. Rabkin. Personal knowledge questions for fallback authentication: security questions in the era of facebook. In *SOUPS '08: Proceedings of the 4th Symposium on Usable Privacy and Security*, pages 13–23, New York, NY, USA, 2008. ACM.

[9] S. Schechter, A. J. Bernheim Brush, and S. Egelman. It's no secret: Measuring the security and reliability of authentication via 'secret' questions. In *Proceedings of the 2009 IEEE Symposium on Security and Privacy*, Washington, DC, USA, May 17–20 2009. IEEE Computer Society.

[10] S. Schechter, S. Egelman, and R. W. Reeder. It's not what you know, but who you know: A social approach to last-resort authentication. In *CHI '09: Proceeding of the twenty-seventh annual SIGCHI conference on Human factors in computing systems*, New York, NY, USA, Apr. 4–9 2009. ACM.

[11] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators: An evaluation of website authentication and the effect of role playing on usability studies. In *Proceedings of the 2007 IEEE Symposium on Security and Privacy*, pages 51–65, Washington, DC, USA, May 20–23 2007. IEEE Computer Society.

[12] R. Stross. What would you do if you logged onto your e-mail and received an unfamiliar message: 'user name and password do not match'? *The New York Times*, oct 5, 2008. http://www.nytimes.com/2008/10/05/business/05digi.html.

[13] M. Wu, S. Garfinkel, and R. Miller. Secure web authentication with mobile phones. In *DIMACS Workshop on Usable Privacy and Security Software*, 2004.

[14] M. Zviran and W. J. Haga. User authentication by cognitive passwords: an empirical assessment. In *JCIT: Proceedings of the Fifth Jerusalem Conference on Information technology*, pages 137–144, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

Demographics

1. What is your participant ID?

   _____

2. Do you have an email address?
   Check one:

   18 Yes
   0 No

3. What is your gender?
   Check one:

   11 Male
   7 Female
   0 Prefer not to answer

4. What is your age?
   Please write your age:

   30-48

5. What is your total household income (include all earners)?
   Check one:

   3 Under $20,000
   3 $20,001-$40,000
   6 $40,001-$60,000
   2 $60,001-$80,000
   4 $80,001-$100,000
   0 Greater than $100,000
   0 Prefer not to answer

6. What is the highest level of education that you have completed?
   Check one:

   0 No high school
   1 Some high school
   0 High school diploma or GED
   6 Some community college
   3 Associates degree or community college degree
   0 Some college
   8 Bachelor's degree
   0 Graduate degree
   0 None of the above
   0 Prefer not to answer

7. In what industry do you currently, or did you most recently, work?
Check one or more:

___ Accounting/Auditing
___ Administrative and Support Services
___ Advertising/Marketing/Public Relations
___ Aerospace/Aviation/Defense
___ Agriculture, Forestry, & Fishing
___ Arts, Entertainment, and Media
___ Automotive/Motor Vehicle/Parts
___ Banking
___ Biotechnology and Pharmaceutical
___ Building and Grounds Maintenance
___ Business Opportunity/Investment Required
___ Career Fairs
___ Computer, Hardware
___ Computers, Software
___ Construction, Mining, and Trades
___ Consulting Services
___ Consumer Products
___ Customer Service and Call Center
___ Education, Training, and Library
___ Electronics
___ Employment Placement Agencies
___ Engineering
___ Executive Management
___ Finance/Economics
___ Financial Services
___ Government and Policy
___ Healthcare – Business Office & Finance
___ Healthcare – Patient Services
___ Healthcare – General
___ Hospitality/Tourism
___ Human Resources/Recruiting
___ Information Technology
___ Installation, Maintenance, and Repair
___ Insurance
___ Internet/E-Commerce
___ Law Enforcement and Security
___ Legal
___ Manufacturing and Production
___ Military
___ Nonprofit
___ Oil/Gas/Utilities
___ Personal Care and Services
___ Publishing/Printing
___ Purchasing
___ Real Estate/Mortgage
___ Restaurant and Food Service
___ Retail/Wholesale
___ Sales
___ Sports and Recreation
___ Supply Chain/Logistics
___ Telecommunications
___ Transportation and Warehousing
___ Prefer not to answer
___ Other (please explain) _____

8. What is your current employment status?
Check one:

5 Full-time
1 Part-time
2 Full-time College/University Student
6 Self-employed
4 Unemployed
0 Retired
0 Other
0 Prefer not to answer

9. What is your current, or most recent, professional level?
Check one:

4 Administrative
6 Staff
0 Consultant
2 Managerial
2 Owner/Founder
0 Director
2 Executive
2 None of the above (please explain) 1 Sales, 1 dishwasher
___ Prefer not to answer

10. Do you use a computer daily for work?
Check one:

12 Yes
1 No
5 Sometimes

11. Which of the options below best describes how often you use the Internet?

Circle one: many times a day	almost every day	every few days	weekly or less	I don't use it
15	3	0	0	0

12. Which of the options below best describes how often you use email?

Circle one: many times a day	almost every day	every few days	weekly or less	I don't use it
13	4	1	0	0

13. Which of the options below best describes how often you use a personal, web-based, email account?

Circle one: many times a day	almost every day	every few days	weekly or less	I don't use it
13	4	1	0	0

14. How recently, if at all, have you accessed (checked mail at) a webmail account at any of the following webmail services?

AOL Mail (from America Online)

Circle one:
never | within last day | within last week | within last month | within last year | over a year ago

Hotmail (MSN Hotmail or Windows Live Hotmail, from Microsoft)

Circle one:
never | within last day | within last week | within last month | within last year | over a year ago

Gmail (from Google)

Circle one:
never | within last day | within last week | within last month | within last year | over a year ago

SplendMail (from Splendorifica)

Circle one:
never | within last day | within last week | within last month | within last year | over a year ago

Yahoo! Mail (from Yahoo! Inc.)

Circle one:
never | within last day | within last week | within last month | within last year | over a year ago

Other free webmail services open to the public

Circle one:
never | within last day | within last week | within last month | within last year | over a year ago

Other webmail services provided by a school, employer, other organizational relationship

Circle one:
never | within last day | within last week | within last month | within last year | over a year ago

15. Have you ever lost your webmail password and had to choose a new password?
Check all webmail services for which you've had to reset your password in order to get into your account:

2 AOL Mail
6 Hotmail
5 Gmail
0 SplendMail
7 Yahoo! Mail
5 Other webmail service
2 Other non-email web service

16. What search engine do you use to search the web?
Check all that apply:

18 Google
1 Live Search (Microsoft)
7 Yahoo! Search
1 Whatever search website is built into my web browser
0 I don't search the web

17. Do you use any of the following Microsoft services that require a Windows Live ID (formerly known as *Passport*) password?
Check all that apply:

10 Hotmail (also known as Windows Live Hotmail or MSN Hotmail)
10 Windows Live Messenger (also known as MSN Messenger)
2 Windows Live Spaces
1 XBOX Live
0 HealthVault
0 Zune music store
1 Other service requiring a Windows Live ID—which one: _____

18. If you have an account that uses a Windows Live ID (Passport) password, how long have you had it for?
Check one:

7 I do not have one
1 Less than three months
1 At least three months but less than one year
1 At least one year but less than two years
3 At least two years but less than four years
5 More than four years

19. If you have any accounts that use a Windows Live ID (Passport) password, have you answered a secret question to use in the event you need to reset your password?
Circle one of the options below only if you have an account that uses a Windows Live ID password:

yes	probably	not sure	probably not	no
3	2	1	1	5

20. If you have any accounts that use a Windows Live ID (Passport) password, have you provided an alternate email address to use in the event you need to reset your password?
Circle one of the options below only if you have an account that uses a Windows Live ID password:

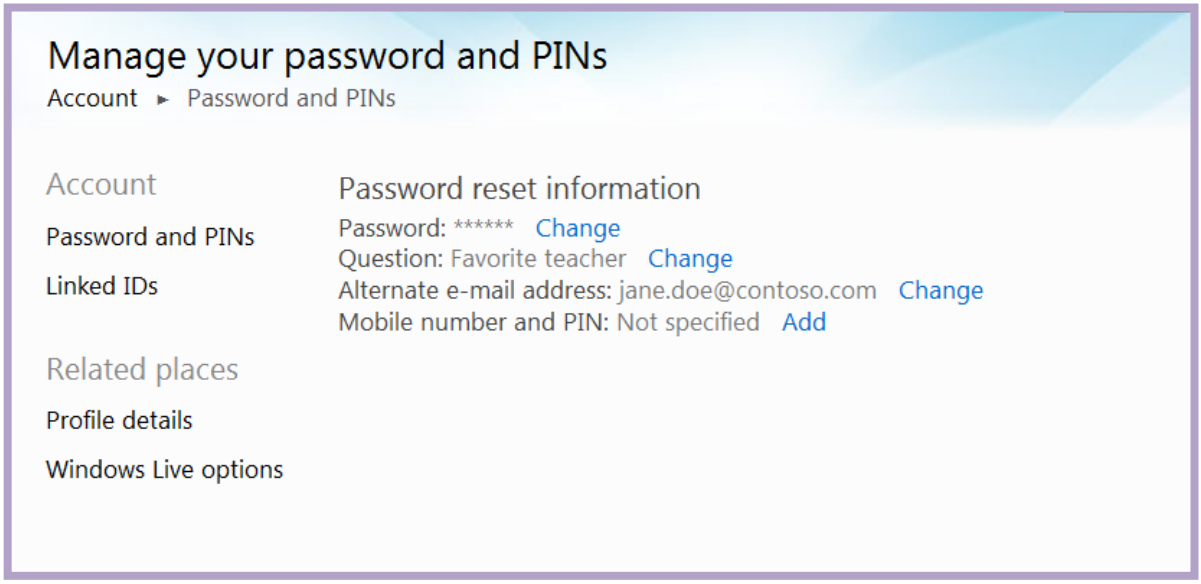
yes	probably	not sure	probably not	no
4	2	1	4	1

21. If you have a SplendMail account, how long have you had it for?
Check one:

18 I do not have one
0 Less than three months
0 At least three months but less than one year
0 At least one year but less than two years
0 At least two years but less than four years
0 More than four years

Windows Live Password Reset Settings 1

Consider the following settings page for Jane Doe's Windows Live account.



Because users lose and forget their passwords, Windows Live (Hotmail, MSN, XBOX Live, etc.) maintains other information that can be used to identify users who forget their passwords. In the above example, Jane Doe has chosen a question ("Favorite teacher"), provided the answer ("Mrs. Smith") and provided an alternate e-mail address (jane.doe@contoso.com).

When someone attempts to login to Jane's Windows Live (Hotmail/MSN/XBOX Live) account, is unable to provide Jane's correct password, and requests to change (reset) Jane's password, Windows Live uses the information Jane has provided to verify that the person asking to change the password really is Jane.

You may refer to the information on this page, as well as any existing knowledge you may already have about Windows Live, to answer the following questions about how Windows Live will verify Jane's identity should she need to reset her password.

22. How does Windows Live use Jane's question ("favorite teacher") and answer ("Mrs. Smith")?
    A. Windows Live asks the question "favorite teacher" and verifies that the response is "Mrs. Smith".
    B. Windows Live presents the name "Mrs. Smith" and asks Jane to identify the question that describes her relationship to Mrs. Smith (that she is Jane's "favorite teacher").

Circle one: definitely A	probably A	not sure	probably B	definitely B
15	2	1	0	0

23. How does Windows Live use Jane's alternate e-mail address (jane.doe@contoso.com)
    A. Windows Live asks "what is your alternate e-mail address" and verifies that the response is "jane.doe@contoso.com".
    B. Windows Live sends an email to jane.doe@contoso.com containing instructions and a code (in the form of a web link) used to identify Jane.

Circle one: definitely A	probably A	not sure	probably B	definitely B
4	3	0	2	**9**

24. If Windows Live were to ask Jane to enter her alternate e-mail address (option A in question 23), would the answer still be correct if it were capitalized differently?

Circle one: yes	probably	not sure	probably not	no
1	1	4	6	3

25. If Windows Live were to send e-mail to jane.doe@contoso.com (option B in question 23), would it reveal that jane.doe@contoso.com was the email address to which the mail was sent or keep this information private?

Circle one: keep private	probably keep private	not sure	probably reveal	reveal
4	5	3	2	4

Jane has configured both an alternate e-mail address (jane.doe@contoso.com) and a question ("favorite teacher").

26. To change her password, will Windows Live require Jane to establish her identity using both the e-mail address and the question, or is one of the two enough?

Circle one: one	probably one	not sure	probably both	both
0	2	3	3	10

27. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to Windows Live?
    - Using the question "favorite teacher"

Circle one: yes	probably	not sure	probably not	no
6	2	3	1	6

28. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to Windows Live?
    - Using her e-mail address (jane.doe@contoso.com)

Circle one:

yes	probably	not sure	probably not	no
2	2	3	4	7

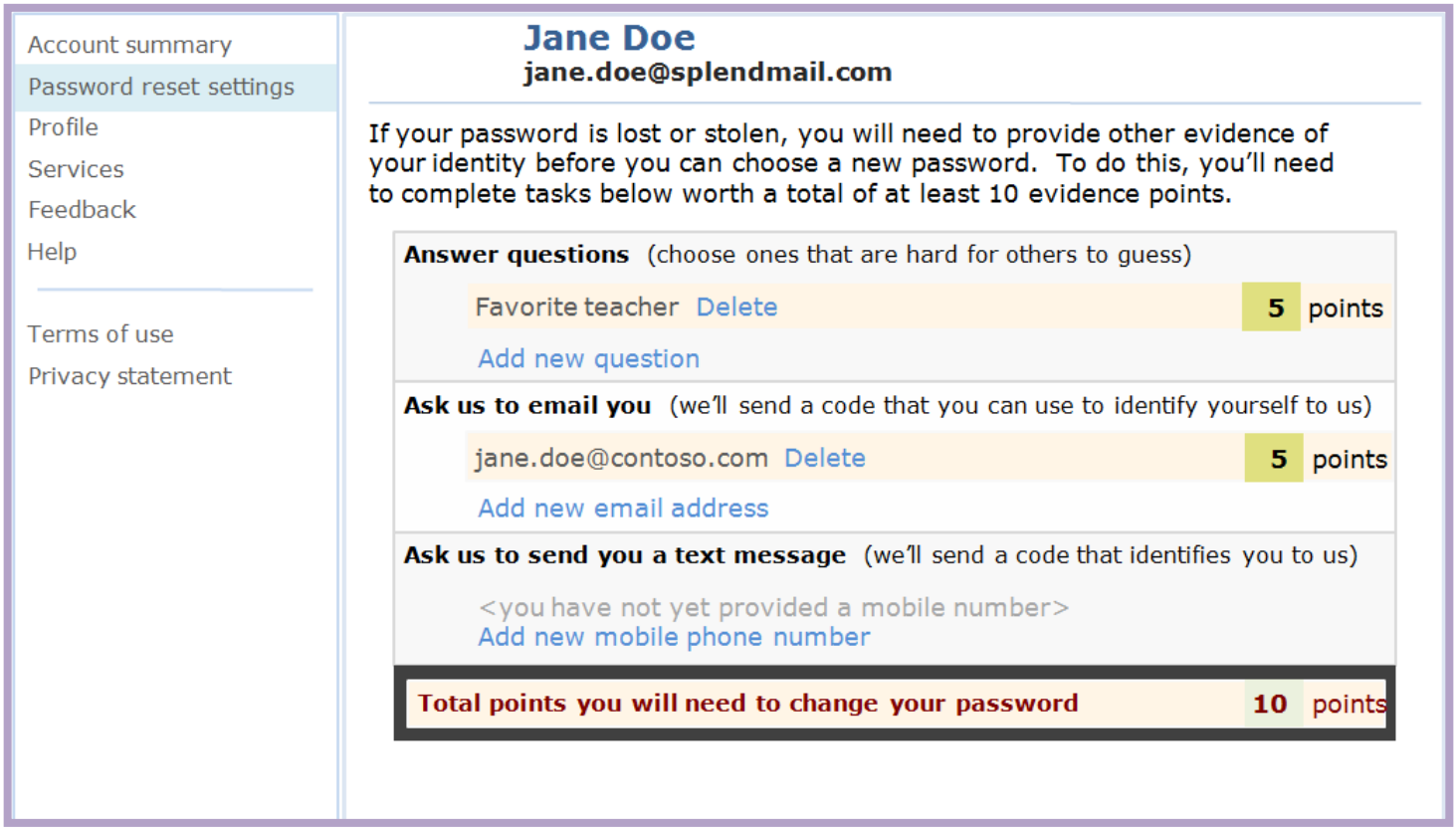
29. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to Windows Live?
    - Using the question “favorite teacher”, **and**
    - Using her e-mail address (jane.doe@contoso.com)

Circle one:

yes	probably	not sure	probably not	no
15	0	2	1	0

SplendMail Password Reset Settings 1

Consider the following settings page for Jane Doe's SplendMail account.



Because users lose and forget their passwords, SplendMail maintains other information that can be used to identify users who forget their passwords. In the above example, Jane Doe has chosen a question ("Favorite teacher"), provided the answer ("Mrs. Smith") and provided an alternate e-mail address (jane.doe@contoso.com).

When someone attempts to login to Jane's SplendMail account, is unable to provide Jane's correct password, and requests to change (reset) Jane's password, SplendMail uses the information Jane has provided to verify that the person asking to change the password really is Jane.

You may refer to the information on this page, as well as any existing knowledge you may already have about SplendMail, to answer the following questions about how SplendMail will verify Jane's identity should she need to reset her password.

30. How does SplendMail use Jane's question ("favorite teacher") and answer ("Mrs. Smith")?
    A. SplendMail asks the question "favorite teacher" and verifies that the response is "Mrs. Smith".
    B. SplendMail presents the name "Mrs. Smith" and asks Jane to identify the question that describes her relationship to Mrs. Smith (that she is Jane's "favorite teacher").

Circle one:

definitely A	probably A	not sure	probably B	definitely B
14	4	1	0	0

31. How does SplendMail use Jane's alternate e-mail address (jane.doe@contoso.com)
    A. SplendMail asks "what is your alternate e-mail address" and verifies that the response is "jane.doe@contoso.com".
    B. SplendMail sends an email to jane.doe@contoso.com containing instructions and a code (in the form of a web link) used to identify Jane.

Circle one:

definitely A	probably A	not sure	probably B	definitely B
1	3	1	4	**10**

Jane has configured both an alternate e-mail address (jane.doe@contoso.com) and a question ("favorite teacher").

32. To change her password, will SplendMail require Jane to establish her identity using both the e-mail address and the question, or is one of the two enough?

Circle one:

one	probably one	not sure	probably both	both
1	1	0	3	**13**

33. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question "favorite teacher" correctly

Circle one:

yes	probably	not sure	probably not	no
0	0	1	2	**15**

34. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Asking SplendMail to email her a code to jane.doe@contoso.com and using that code to identify herself to SplendMail

Circle one:

yes	probably	not sure	probably not	no
3	2	0	1	**12**

35. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Asking SplendMail to send her a code via text message to 425 555 4242 and using that code

Circle one:

yes	probably	not sure	probably not	no
1	1	0	3	**13**

36. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question "favorite teacher" correctly, **and**
    - Asking SplendMail to email her a code to jane.doe@contoso.com and using that code to identify herself to SplendMail

Circle one:

yes	probably	not sure	probably not	no
18	0	0	0	0

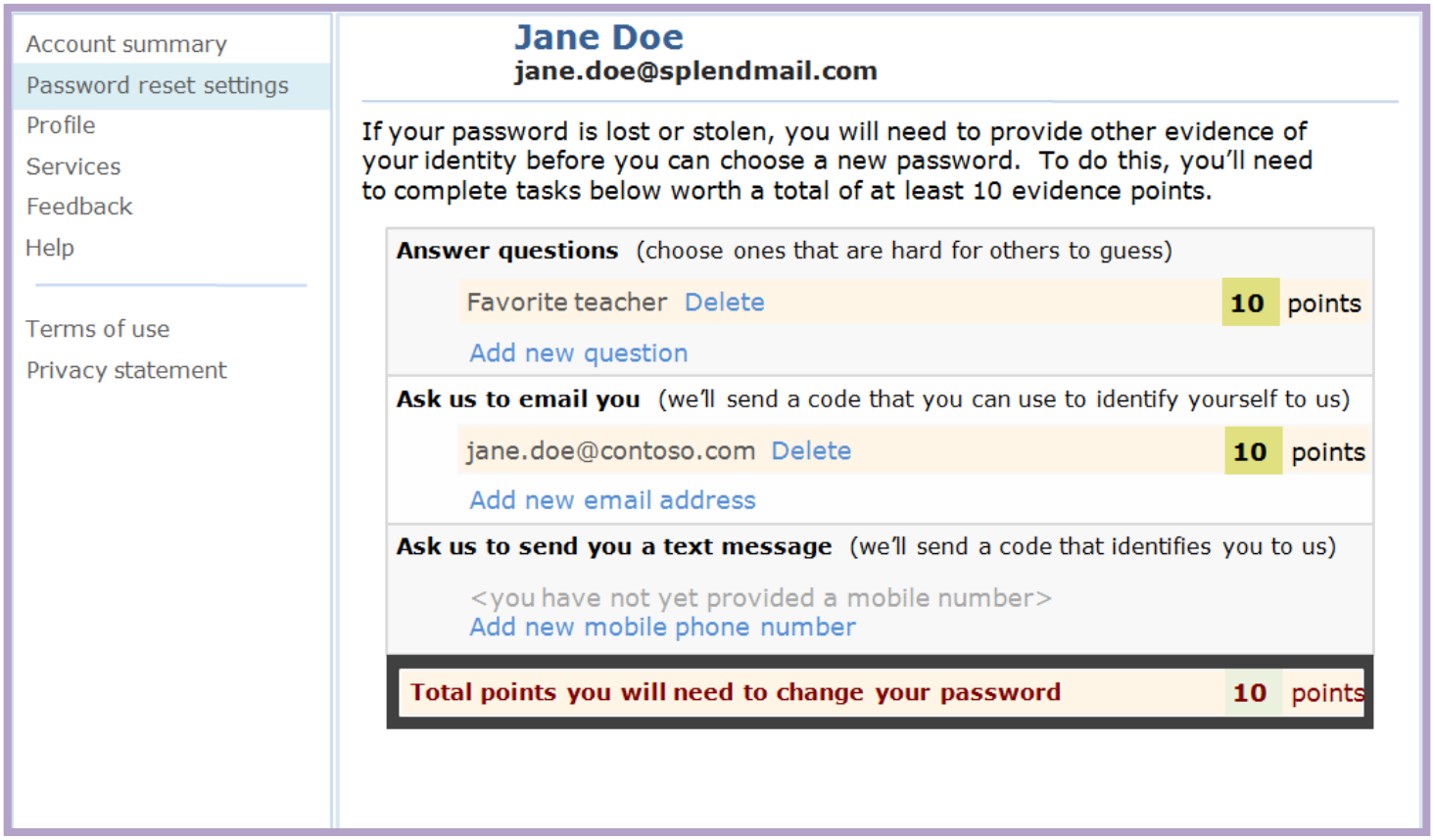
37. If SplendMail were to send e-mail to jane.doe@contoso.com, would it reveal that jane.doe@contoso.com was the email address to which the mail was sent or keep this information private?

Circle one:

keep private	probably keep private	not sure	probably reveal	reveal
4	4	8	1	1

SplendMail Password Reset Settings 2

Consider the following settings page for Jane Doe's SplendMail account.



You may refer to the information on this page, as well as any existing knowledge you may already have about SplendMail, to answer the following questions about how SplendMail will verify Jane's identity should she need to reset her password.

38. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly

Circle one:

yes	probably	not sure	probably not	no
11	1	0	2	4

39. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Asking SplendMail to email her a code to jane.doe@contoso.com and using that code to identify herself to SplendMail

Circle one:

yes	probably	not sure	probably not	no
11	2	0	0	5

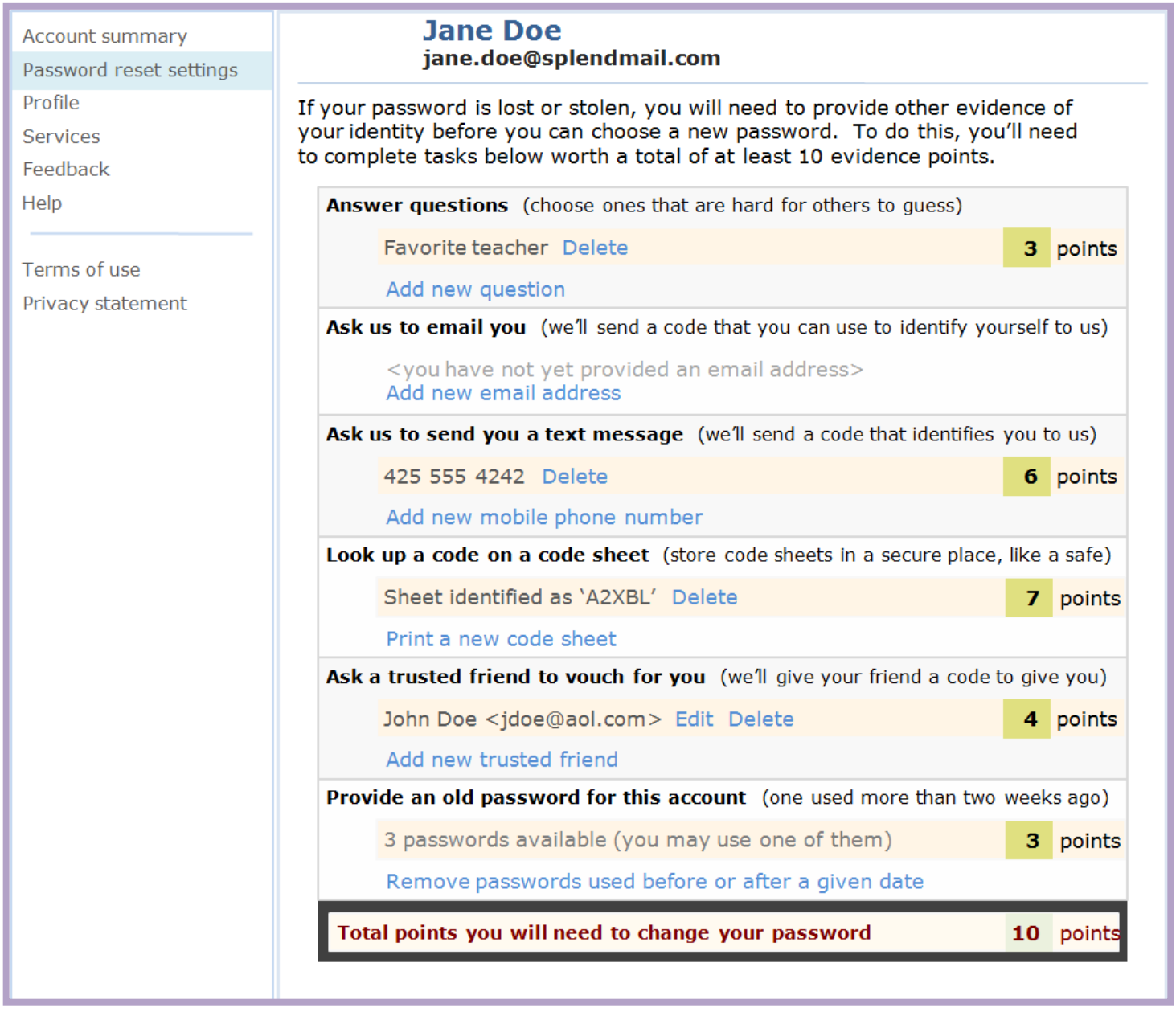
40. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Asking SplendMail to email her a code to jane.doe@contoso.com and using that code to identify herself to SplendMail

Circle one:

yes	probably	not sure	probably not	no
12	3	0	2	1

SplendMail Password Reset Settings 3

Consider the following settings page for Jane Doe's SplendMail account.



Using SplendMail's advanced settings, Jane has configured a question ("favorite teacher"), added a mobile phone number (425 555 4242), printed a code sheet (labeled 'A2XBL'), and identified a trusted friend (John Doe). She has changed passwords three times since setting up her SplendMail account and SplendMail still has records of her three old passwords.

You may refer to the information on this page, as well as any existing knowledge you may already have about SplendMail, to answer the following questions about how SplendMail will verify Jane's identity should she need to reset her password.

41. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly

Circle one:

yes	probably	not sure	probably not	no
0	0	0	0	**18**

42. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Looking up a code from the code sheet identified as ‘A2XBL’ and providing it to SplendMail

Circle one:

yes	probably	not sure	probably not	no
0	1	0	0	**17**

43. *Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Asking SplendMail to send her a code via text message to 425 555 4242 and using that code

Circle one:

yes	probably	not sure	probably not	no
2	0	0	0	**16**

44. *Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Looking up a code from the code sheet identified as ‘A2XBL’ and providing it to SplendMail, **and**
    - Answering the question “favorite teacher” correctly

Circle one:

yes	probably	not sure	probably not	no
17	0	0	0	1

45. *Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Providing one of the three old passwords she had previously used for her SplendMail

Circle one:

yes	probably	not sure	probably not	no
1	0	0	1	**16**

46. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Asking her friend John Doe (jdoe@aol.com) to vouch for her by giving her a code and providing that code to SplendMail (as described in the SlendMail password reset settings page) , **and**
    - Providing one of the three old passwords she had previously used for her SplendMail

Circle one:

yes	probably	not sure	probably not	no
17	0	0	1	0

47. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Asking SplendMail to send her a code via text message to 425 555 4242 and using that code, **and**
    - Providing one of the three old passwords she had previously used for her SplendMail

Circle one:

yes	probably	not sure	probably not	no
17	1	0	0	0

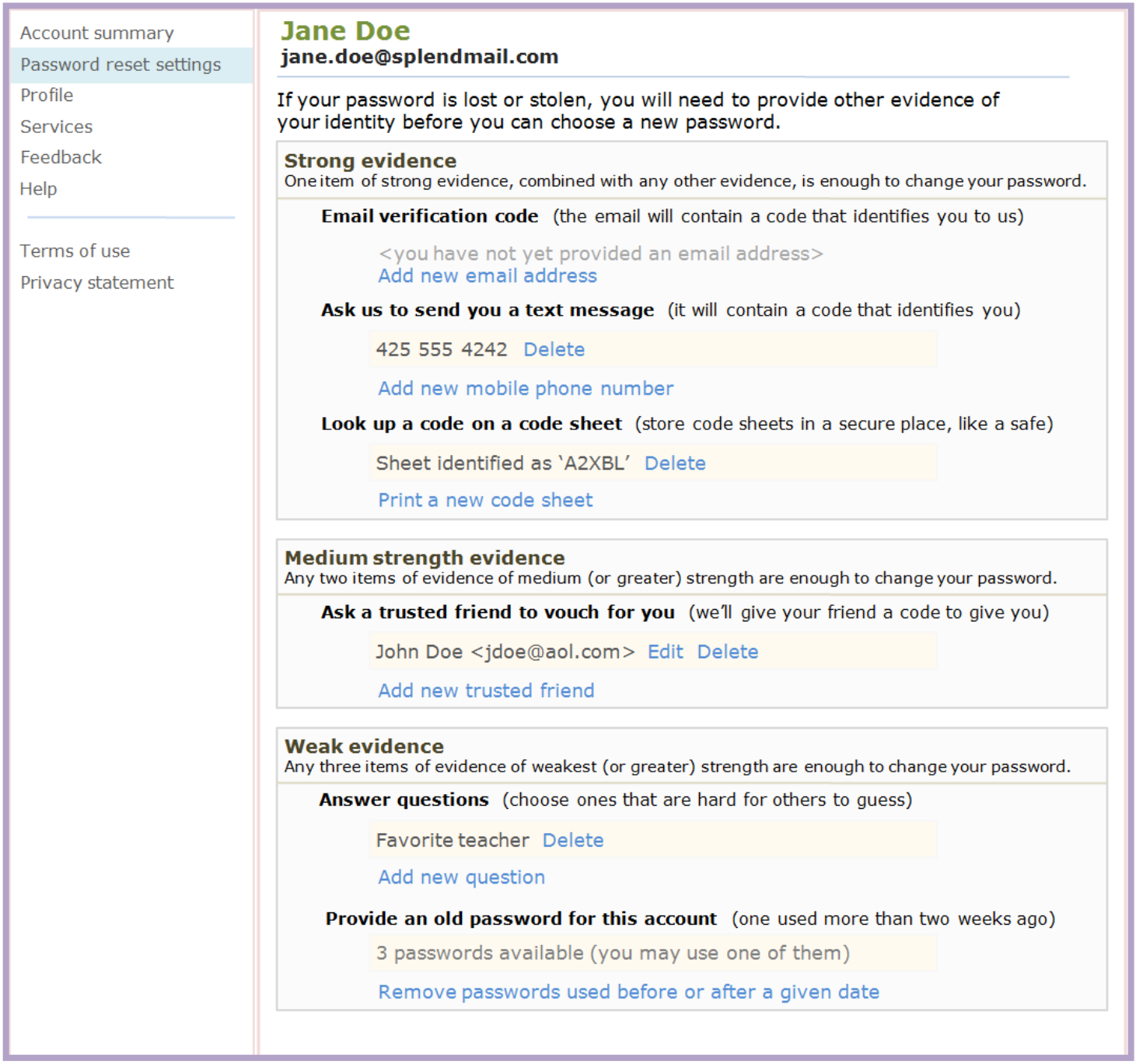
48. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Providing two of the three old passwords she had previously used for her SplendMail

Circle one:

yes	probably	not sure	probably not	no
0	0	0	1	**17**

Old SplendMail Password Reset Settings

SplendMail recently updated their advanced password reset settings page. Consider the settings page for Jane Doe's SplendMail account from before this update.



You may refer to the information on this page, as well as any existing knowledge you may already have about SplendMail's old password reset settings page, to answer the following questions about how SplendMail will verify Jane's identity should she need to reset her password.

49. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly

Circle one:

yes	probably	not sure	probably not	no
1	0	0	1	**16**

50. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Looking up a code from the code sheet identified as ‘A2XBL’ and providing it to SplendMail

Circle one:

yes	probably	not sure	probably not	no
5	1	0	0	**12**

51. *Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Asking SplendMail to send her a code via text message to 425 555 4242 and using that code

Circle one:

yes	probably	not sure	probably not	no
17	0	0	0	**1**

52. *Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Looking up a code from the code sheet identified as ‘A2XBL’ and providing it to SplendMail, **and**
    - Answering the question “favorite teacher” correctly

Circle one:

yes	probably	not sure	probably not	no
17	0	0	1	0

53. *Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Providing one of the three old passwords she had previously used for her SplendMail

Circle one:

yes	probably	not sure	probably not	no
2	0	0	2	**14**

54. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Asking her friend John Doe (jdoe@aol.com) to vouch for her by giving her a code and providing that code to SplendMail (as described in the SlendMail password reset settings page) , **and**
    - Providing one of the three old passwords she had previously used for her SplendMail

Circle one:

yes	probably	not sure	probably not	no
13	1	0	2	2

55. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Asking SplendMail to send her a code via text message to 425 555 4242 and using that code, **and**
    - Providing one of the three old passwords she had previously used for her SplendMail

Circle one:

yes	probably	not sure	probably not	no
16	2	0	0	0

56. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Providing two of the three old passwords she had previously used for her SplendMail

Circle one:

yes	probably	not sure	probably not	no
3	0	0	3	**12**

57. Will Jane be able to change her password after performing all of the following actions (and only those actions) to prove her identity to SplendMail?
    - Answering the question “favorite teacher” correctly, **and**
    - Asking her friend John Doe (jdoe@aol.com) to vouch for her by giving her a code and providing that code to SplendMail (as described in the SlendMail password reset settings page)

Circle one:

yes	probably	not sure	probably not	no
1	0	0	3	**14**

SplendMail Comparison

58. Do you prefer the old interface (with strong, medium, and weak evidence) used for these questions, or the new interface that used points (which you used for questions 41 through 48 above)?

Circle one:

prefer old	slightly prefer old	indifferent	slightly prefer new	prefer new
2	1	2	2	11