2. Study Methodology

The typical tasks that users would need to accomplish with password managers fall into four categories:

- Migrate user accounts (passwords) to use the password manager
- Log in to protected user accounts from a primary computer
- Change passwords for user accounts
- Access user accounts remotely, i.e., from a computer other than the primary computer, such as on a public or friend's machine.

Each participant completed a one-hour session, where they completed a set of five tasks designed to simulate the real tasks that users would accomplish with the password managers. The set of tasks was repeated so that each participant completed them with both PwdHash and Password Multiplier. The order in which the tasks and the programs were presented was balanced to avoid bias. Throughout the session, the experimenter observed the participant and recorded their actions. Additional user feedback was gathered through questionnaires.

Participants

Twenty-seven adults participated in the study. Most were students at our university, from various faculties and degree programs; none were students specializing in computer security. A few had technical backgrounds: four were from Computer Science, one studied Information Systems, and none were from Engineering. Data from one participant was eliminated as a language barrier coupled with very little computer experience hindered their ability to understand the tasks. Of the remaining 26 participants, 21 were between the ages of 18 and 30 and five were over 30 years old.

The participants were familiar with using the web and logging on to web sites requiring a username and password. All but two reported visiting the web daily, and these two said they were online several times a week. The participants were fairly comfortable with using computers; 24 of the participants self-rated their general computer skill level at 6 or higher on a scale of 1 to 10.

We chose not to screen participants based on experience using Firefox. Typical Firefox users are more technically sophisticated than average users so pre-selecting on this criterion would have biased our pool of participants. Additionally, interaction with the browser's interface was minimal; participants simply had to enter URLs and navigate within web pages. These tasks are accomplished in the same manner in Firefox and Internet Explorer.

A pre-task questionnaire was used to gain insight into the participants' initial attitude towards web security and passwords.

Apparatus

Participants completed a set of tasks using two different computers during the session. Both computers were running Windows XP and Mozilla Firefox. One system had the PwdHash plug-

in (version 1.0 for Mozilla Firefox) installed while the second computer included the Password Multiplier plug-in (version 0.3 for Windows, Linux, and Mac OS).

The tasks are described in the following list. The *Second Login* task is dependent on the *Update Pwd* task, i.e., users must have successfully changed their password before they are able to log on to the site a second time with their new protected password. All other tasks are independent of each other. We did not include a "delete password" task because neither system supports this functionality.

The tasks are:

Log In: Logging on to a web site that already has its password protected by the plug-in. This simulates how users log on once their passwords have been converted to protected passwords.

Migrate Pwd: Logging on to a web site with an unprotected password then changing the password so that it becomes protected. This is required by users to initially migrate each of their passwords.

Remote Login: Logging on to a web site with a protected password from a remote computer that does not have the plug-in installed. This models how users would log on to their accounts from a computer other than their primary machine.

Update Pwd: Logging on to a web site with a protected password then changing it to a new protected password. This situation would arise if users had to change their password once it is already protected.

Second Login: Logging on to a web site a second time, once the user has changed the password to a protected password. This task tests whether users understand how to log on to their account once they have changed to a protected password.

The tasks were set up using popular web sites (Hotmail, Google, Amazon, and Blogger) that users may encounter in real life. Test accounts were created so that participants did not use their personal accounts or passwords at any point during the experiment.

Participants completed the set of tasks with both plug-ins; the order was balanced so that each plug-in was seen first the same number of times. The order of the tasks within a set was also shuffled but an individual participant saw the tasks in the same order for both plug-ins. The *Update Pwd* and *Second Login* tasks were ordered so that they were always separated by exactly one task (for example, a participant completed the tasks in the order of *Log In, Remote Login, Update Pwd, Migrate Pwd, Second Login*). This ensured that participants changed their focus for a time before logging on to the web site a second time with their new protected password. One participant quit after completing only the tasks with PwdHash, but the remaining participants completed all tasks.

One of the difficulties with testing the usability of these plug-ins is that they initially have no visible interface. Even during the interaction, only Password Multiplier has a visible pop-up

window. So simply giving the tasks to participants without instructions on how to use the plug-ins would have been futile. To preserve ecological validity, we tried to keep the instructions to a minimum; giving them written details of how to activate the plug-in, a brief explanation of how to change a password, and a short description of how to log on to a web site using a remote computer. The entire set of instructions was approximately half a page long for each plug-in. Participants were also given a list of the usernames and passwords that they would require to complete the tasks. To minimize the effect of learning new passwords, a simple, one-word password was given for all tasks within a system ("alphabet" and "carleton"). These passwords were also written on a sheet in front of participants throughout the session.

Participants were given the instruction sheet for the particular plug-in and told that they could refer to it whenever necessary. They were directed to a computer with a Firefox browser window open and the appropriate plug-in pre-installed. They were instructed to pretend that this was their home computer and they should use Firefox as the browser for these tasks. Participants completed all tasks with a plug-in before switching computers to repeat the tasks with the second plug-in. No participant expressed any concern over using Firefox instead of the more popular Internet Explorer and no difficulties were observed due to using this alternative browser. Firefox was selected as the browser because the stable versions of the plug-ins were not available for Internet Explorer. Firefox was used in the original PwdHash usability study as well.

Each task was described on an index card. The card also included two questions asking participants to rate the difficulty of the task and their satisfaction with the software for this particular task. Participants could take as long as they needed to complete the task and were told that if they felt they had spent enough time on a task and could not complete it, they could quit. At the end of each task, they circled their responses to the two questions and were provided with the next index card.

When participants reached the task where they had to log on to a web account from a remote computer (*Remote Login* task), they were instructed to change computers and pretend that they were now at their friend's house where the software was not installed. This proved problematic for Password Multiplier since the authors' solution to remote access is to install the plug-in. Participants could not install the plug-in on the second computer because it had PwdHash installed and the combination of the two crashed the computer. The *Remote Login* task was therefore eliminated for Password Multiplier. Judging from participants' reactions as they read from the instruction sheet that they had to install software for remote access, they would not have been pleased with this solution even if they had been able to complete the task. Using a third computer would have been a better experimental design, allowing participants to complete the task, but we do not expect that this would have led to different results.

After completing a set of tasks, participants answered a paper questionnaire about their experience with the particular plug-in. The entire process was repeated for the second plug-in. A final post-task questionnaire asked participants to compare the two plug-ins.

Data Collection

Data was collected in two ways: through observation and through questionnaires. An experimenter sat with each participant throughout the session, recording observations, noting any difficulties, any obvious misconceptions in the participant's mental model of the software, any comments made by the participant, and whether they successfully completed the task. Participants were asked at the beginning of the session to ``think-aloud''. Besides the standard instructions given to all participants, no further explanations were given even if a participant asked for more instructions. In these cases, the experimenter remained cordial, clarifying that we were testing the usability of the systems and needed to see if people could use them without explanations. Occasional prompts such as "what did you expect to happen there?" were used if participants forgot to think-aloud.

The users' goal was to successfully complete the tasks using the given password manager. They were given as much time as they wanted and the observer waited for the participants to signal that they had completed the task or that they had run out of ideas and could not complete it. The outcome of each task was recorded by the observer according to the following possibilities:

Successful: The participant completed the task without difficulty.

Dangerous success: The participant eventually completed the task after several attempts (i.e., had difficulty). The negative impact is that in some cases, the unsuccessful attempts prior to the eventual success expose the password to attack

Failed: The participant gave up on the task without completing it.

False completion: The participant failed to complete the task but erroneously believed that they had in fact been successful.

Failed due to previous: The participant could not complete the task because they had incorrectly completed the preceding task. This only applies to the *Second Login* task, where the *Update Pwd* had to be successful in order to proceed.

The first outcome is considered most positive. The second is somewhat positive but users may have exposed their passwords to danger (e.g., to JavaScript attacks and phishing) as they floundered with the task. They may even have inadvertently exposed multiple passwords, since a typical reaction to being unable to log in is to try all of one's passwords to see if something will work. The fourth outcome is especially dangerous because it leads to a false sense of security on the part of users.

Secondary measures taken in the study consisted of several Likert-scale questions. These ask respondents to choose their level of agreement with the given statement from a set of possible answers, usually ranging from strong agreement to strong disagreement. We used a 5-point scale (strongly disagree, disagree, neutral, agree, strongly agree). Participants answered two of these questions on the index cards after each individual task, then completed a 16-question questionnaire for each plug-in.

The questions from the questionnaire were a priori grouped into four sets that considered different aspects of the interaction: perceived security, comfort level with giving control of passwords to a program, perceived ease of use, and perceived necessity and acceptance. Each set contained four similar questions; the questions were randomly organized on the questionnaire so participants were not aware of the groupings. Participants circled their answer for each question among the five choices. Half of the questions were inverted to avoid bias.