Usability Problems with Password Creation Systems: Results from Expert and User Evaluation

Saja Althubaiti
Department of Computer Science
University of York
United Kingdom
saaa505@york.ac.uk

Helen Petrie
Department of Computer Science
University of York
United Kingdom
Helen.Petrie@york.ac.uk

1 INTRODUCTION

Textual passwords are still extremely widely used and continue to be a problem for users and a major concern for online security. Password creation systems (PCSs) present their own usability problems, and if users are struggling to understand how the PCS works, they will have less cognitive effort available to create a strong password. We investigated the usability problems of six typical PCSs using two different evaluation methods, to assess the number and severity of the usability problems users might encounter and to begin development of a method specifically for the evaluations of PCSs.

2 METHOD

2.1 Password Creation Systems evaluated

Six PCSs were selected based on the Alexa ratings of their websites (ratings taken 26 May 2014). Additional criteria for inclusion were: (1) PCS should be in English; (2) website should have a dedicated PCS (i.e. not use Google or Facebook for login); and (3) PCS should not automatically generate passwords. The PCS websites, with their Alexa ratings, are shown in Table 1.

Table 1. Six Password Creation Systems (PCSs) evaluated

Website	Description	Alexa rating
Apple	Online retailer	33
Daily Mail	Online newspaper	89
Netflix	Internet streaming media	78
Stackoverflow	Q&A service for programmers	50
Wikipedia	Online encyclopedia	6
Wordpress	Blog web hosting	27

2.2 Method for expert-based evaluation

A group expert evaluation method based on collaborative heuristic evaluation [1] was used. A group of three to five experts worked through each PCS, agreeing a set of usability problems, and rating them privately using Nielsen's usability problem severity rating scheme [2].

2.2.1 Experts

Seven usability experts, who all work or study at the University of York, participated. Three were women and four were men, ages ranged from 25 to 59 years (mean 34.5). All have at least five years experience with usability evaluations, including expert evaluations. Experts were not compensated for their participation, but were offered coffee and tasty cookies during the evaluations.

2.2.2 Equipment and Materials

Two PCs were used, both connected to projectors. One displayed the PCS to the experts and one displayed the evolving list of proposed usability problems.

Experts were given a summary of the severity rating scheme, which asked them to rate problems on a four point scale, from 1 ("cosmetic problem, would be nice to fix") to 4 ("catastrophic problem the user would not be able to proceed) [2].

For each PCS, a set of appropriate passwords was prepared, designed to show strengths and weaknesses of the particular PCS; these were based on the authors' extensive exploration of the PCS. The experts were encouraged to use these passwords, but were also free to try any other passwords, to see their effects on the PCS.

2.2.3 Procedure

Four evaluation sessions were conducted with three to five experts participating in each session, depending on the availability of the experts. Each session lasted approximately two hours, with three PCSs evaluated in each session. Each PCS took approximately 30 minutes to evaluate, with short comfort breaks between each PCS evaluation. Each session was guided by a facilitator (one of the two authors).

At the beginning of each session, the facilitator introduced the aim of the study and briefed the experts on the procedure. One expert acted as "driver" of the PCS, interacting with the PCS as requested by all the experts. The facilitator acted as scribe, recording potential usability problems proposed by the experts. The PCS and the list of potential problems were displayed on large screens, so the experts could view and discuss them easily.

For each PCS, the experts completed one task: create a new password. They were asked to explore as many possibilities in this task as they wished, trying out different passwords from the list provided and any others they wished to try, to see the effects in the PCS. Any expert could propose a potential usability problem and discussion was allowed about the precise nature of the problem. If an expert believed it was not actually a problem, they were asked not to air this opinion publicly. When the problem description was agreed, each expert rated its severity privately. If an expert did not think it was actually a problem, they rated its severity as zero. This procedure was repeated until the experts felt there were no more problems to be identified in the PCS.

2.3 Method for user-based evaluation

2.3.1 Participants

24 participants took part, who all work or study at the University of York, 22 students and 2 administrative staff. 11 were women

and 13 were men, ages ranged from 18 to 33 years (mean 21.8). All participants (except one for one PCS) had not created an account in the last month for any of the PCSs in the study. Participants were remunerated with £15 Amazon gift vouchers.

2.3.2 Equipment and Materials

A MacBook Pro laptop running MacOS v10.10 and Mozilla Firefox v33.1 were used. For recording the computer screen and the participant's voice, ScreenFlow software (v4.5) was used. Lists of passwords and summary sheets for the severity ratings were the same as those used in the expert evaluation.

2.3.3 Procedure

Participants were run in individual sessions lasting approximately 80 minutes, taking 15 minutes to evaluate each PCS. The order of evaluation of the PCSs was counterbalanced between participants. For each PCS, participants used the same set of appropriate passwords to create a password as used by the experts.

Participants were briefed about the study, the facilitator emphasizing that the study did not test their password creation skills, but the usability of the PCSs. Participants then signed an informed consent. Participants were instructed to think aloud while doing the password creation tasks, mentioning any problems they encountered. The facilitator prompted the participants if they appeared to be having a problem but not articulating it. Whenever they encountered a problem, they were asked to rate its severity using the four-point scale as used by the experts. The procedure was repeated for each PCS.

3 RESULTS AND DISCUSSION

The two evaluations produced a pool of 121 usability problems, 40 (33.1%) found by the expert evaluation only, 43 (35.5%) found by both the expert and user evaluation and 38 (31.4%) found by the user evaluation only. Figure 1 shows the breakdown of the 6 PCSs, and indicates that there was no clear overall pattern of either expert or user evaluation revealing a greater proportion of problems.

Table 2 shows the number of problems identified by expert and user evaluation per PCS, as well as their mean severity rating. The range in the number of problems was large for both expert and user evaluations; this was due to the very different levels of functionality of the PCSs. In general, the number of problems found by expert and user evaluations were similar, except for WordPress, for which the users encountered far more problems than the experts. Table 2 also shows that the users were more severe in their ratings than experts.

Table 3 shows the distribution of usability problems into three main categories as identified by the expert evaluation only, the user evaluation only and by both evaluations. Both experts and users reported numerous problems with the clear statement of the features of PCS. The experts reported more problems in the interface/interaction category than users. Whereas users encountered more problems in the Feedback category than the experts.

These results show that there are numerous usability problems in current PCSs, which users in particular, rate as quite severe. We have begun to explore the different types of problems identified by experts and users and are now creating an method to guide experts in the evaluation of PCSs.

Table 2. Number of problems identified by expert and user evaluation with mean (standard deviation) severity ratings

	Expert Evaluation		User Evaluation	
	N	Severity	N	Severity
Apple	23	2.0 (0.59)	16	2.0 (0.50)
DailyMail	20	2.1 (0.49)	11	3.0 (0.28)
Netflix	7	1.5 (0.49)	8	3.0 (0.84)
Stackoverflow	17	2.2 (0.49)	19	3.0 (0.68)
Wikipedia	8	2.3 (0.65)	6	3.0 (0.80)
WordPress	8	1.9 (0.63)	21	3.0 (0.69)
Total	83		81	

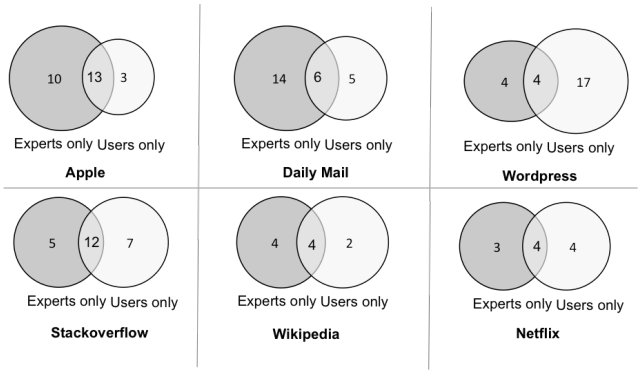

Figure 1. Number of usability problems found by expert and user evaluation for the six PCSs

Table 3. Number (%) of usability problems in three categories

	Experts Only	**Users Only**	**Experts and users**	**Total**
Statement of features	10 (25.6%)	9 (23.1)	20 (51.3)	39
Interface/ interaction	17 (47.2)	11 (30.6)	8 (22.2)	36
Feedback	6 (20.7)	11 (37.9)	12 (41.4)	29
Misc	7 (41.2)	7 (41.2)	3 (17.7)	17

4 REFERENCES

[1] Petrie, H. & Buykx, L. (2010). Collaborative Heuristic Evaluation. Proceedings of UPA 2010 International Conference. Omnipress. Available at: http://upa.omnibooksonline.com/index.htm

[2] Nielsen, J. (1995). Severity ratings for usability problems. Available at: http://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/