Are Passfaces[1] More Usable Than Passwords? A Field Trial Investigation

Sacha Brostoff & M. Angela Sasse

Department of Computer Science, University College London, London, WC1E 6BT

Tel: *+44 20 7679 3462*
Fax: *+44 20 7387 1397*
Email: *s.brostoff@cs.ucl.ac.uk*

The proliferation of technology requiring user authentication has increased the number of passwords which users have to remember, creating a significant usability problem. This paper reports a usability comparison between a new mechanism for user authentication - Passfaces - and passwords, with 34 student participants in a 3-month field trial. Fewer login errors were made with Passfaces, even when periods between logins were long. On the computer facilities regularly chosen by participants to log in, Passfaces took a long time to execute. Participants consequently started their work later when using Passfaces than when using passwords, and logged into the system less often. The results emphasise the importance of evaluating the usability of security mechanisms in field trials.

Keywords: task performance, evaluation, passwords, security, human memory.

1 Introduction

Most computers contain and process data which needs to be protected, and many other technologies – such as mobile phones - require some sort of access control. On most computer systems, this is done through a process of user identification and authentication (Garfinkel and Spafford, 1996). Through *identification,* the user's right to access a system is established. Once a user's identity is established, the *authentication* mechanism verifies that the user is who he says he is.

There are three types of user authentication: examining what the user *knows, possesses* or *is* (Menkus, 1988). *Knowledge-based authentication* uses a secret word or phrase shared between the user and the computer system, with the user revealing the secret to the computer to prove their authenticity. *Token-based authentication*

[1] Passfaces™ have been used by kind permission of the patent and trademark holding company, Id-Arts (http://www.id-arts.com).

uses a physical token that is difficult to obtain or forge. *Biometric authentication* relies on the uniqueness of details in a person’s anatomy or behaviour - a user whose characteristics match the electronic equivalent of those characteristics recorded in the computer is accepted as valid. Examples of such characteristics used currently include fingerprints (Roddy and Stosz, 1997), retinal patterns (Arthur, 1997), signatures, keystroke dynamics in typing (Obaidat and Sadoun, 1997), and voice properties (Kim, 1995).

Today, knowledge-based authentication is the most widely used mechanism, in the form of the *password*. Many companies require multiple computer systems throughout their businesses. Business is using an ever increasing number of computer systems, and so more users are acquiring more passwords. However, there is plenty of evidence that passwords are neither usable nor secure. Many users forget their passwords (Zviran and Haga, 1993), and with the number of passwords per user increasing, the rate of forgetting increases further (Adams et al., 1997). A visible consequence is that password users require extensive support (Murrer, 1999). Support takes the form of a password reset, estimated at up to £40 a reset (Brennan, 2000, personal communication). With a typical support requirement of 1 password reset per 4-5 users per month, this represents a considerable cost: a company with 100 to 120 thousand employees would have 25,000 password resets a month.

Despite the large amount of money invested in it, password mechanisms often are not as secure as expected. The passwords chosen by most users are relatively easy to crack (Davis and Ganesan, 1993; Adams and Sasse, 1999). The continuing increase of networked systems introduces an additional risk, since passwords sent across networks in plain text can be intercepted through mechanisms such as packet sniffing (Garfinkel and Spafford, 1996).

Given the number and quality of problems associated with passwords, why are they still so widely used? In our experience, staff responsible for computer security – system administrators and IT managers – are generally reluctant to change existing security mechanisms. Despite mounting evidence of password problems, they feel that sticking with “the devil you know” is safer than experimenting with new mechanisms. A closer examination of the alternatives explains that reluctance – other mechanisms have their own problems.

1.1 Token-based authentication

Token-based authentication requires token construction and distribution, which is far from trivial and has led to documented financial loss (Anderson, 1994). The token must be physically presented to the computer system, which requires additional hardware for reading the token. Both token and token reader cost money, and a reader must be available at every point a user might be authenticated. As costs of tokens and readers fall, this will be less of an issue. However, presentation of a valid token does not prove ownership - the token may have been stolen. And although a token may be hard to forge, it does not mean it is impossible or uneconomic to do so (Svigals, 1994). For these reasons, tokens are mostly used for identification only as part of a two-factor procedure (see below); the user still needs to authenticate him/herself through some other means, usually a password.

1.2 Biometric authentication

Biometric authentication raises issues of trust among many users, who fear it could be used to track them constantly, as in the Big Brother scenario. But there are problems beyond potential mis-use by an unaccountable entity. Since users' biometric characteristics (such as the shape of face or fingerprints) cannot be easily changed, it is paramount that the security of the characteristic is protected. There are, however, many points at which a description characteristic may be illicitly gained without maiming the actual owner: digital representations of the characteristic must be stored somewhere to compare against the user being authenticated. If somebody else obtains the digital representation of the characteristic, the user can be impersonated with impunity (Kim, 1995). The digital representation may have to be transmitted across a computer network during authentication, and so could be intercepted (using mechanisms such as packet sniffing – see above). Finally, analogue copies of a biometric characteristic may be left behind by the user from which the digital representation can be replicated, such as fingerprints on a beer glass, speech on an answering machine, or a signature on a form. Unlike the U.S. president, most users are not trailed by secret service agents who systematically break their beer glasses for them; thus, the chore of safeguarding of these characteristics falls on the user.

As with token-based authentication, structural or physiological biometrics requires special hardware, which is expensive. Behavioural biometrics - such as keystroke dynamics - do not necessarily require costly hardware, but are not popular with users since they can be used to monitor productivity as well (Deane et al., 1995).

To make tokens or biometrics sufficiently secure, they have to be combined with another mechanism into a two-step procedure, using a second mechanism to shore up the weaknesses of the first. A combination of two mechanisms requiring special hardware would double the already high cost associated with these methods. Therefore, a combination of tokens or biometrics with a knowledge-based mechanism is likely to remain the most common form of access control – such as the cash card and Personal Identification Number (PIN). Using tokens or biometrics for user identification reduces the cognitive load of traditional computer login procedures, since the user no longer has to recall the specific user-id or account name for a particular system. Adams (1996) found that for users with many different systems and varying account names, recalling the user id presents a significant load in itself. But overall, the introduction of tokens and biometrics will lead to a further increase in the total number of knowledge-based items users have to recall for access procedures.

1.3 Improving knowledge-based mechanisms

The majority of users in Adams and Sasse's (1999) study reported they could not cope with the number of passwords they had; consequently, they wrote passwords down and/or disclosed them to others, breaking the most elementary rule of knowledge-based authentication. Many security professionals would regard this “remedy” as unacceptable. However, the cost of resetting forgotten passwords has reached such proportions in some organisations that their security staff regard “writing passwords down and storing them in a safe place” as the lesser evil. Given that the number of applications requiring user authentication in some form is increasing rapidly – consider mobile phones, Personal Digital Assistants (PDAs), remote access to services and encryption – many individual and corporate users will face serious security problems unless usability of knowledge-based mechanisms is improved.

What are the options for improving the usability of knowledge-based authentication mechanisms? The biggest problem with passwords is that users forget them easily (Zviran and Haga, 1990; Zviran and Haga, 1993). Recall is one of many routes to remembering which have been assessed in psychological experiments (Baddeley, 1997). To be secure, i.e. not be guessable, a password must be a random combination of numbers, symbols and letters (Garfinkel and Spafford, 1996). Unfortunately, these types of passwords are more difficult for people to recall than meaningful – guessable – ones, such as names. It has also been established that *cued recall* leads to better remembering than recall alone, and *recognition* has better accuracy than cued recall (Parkin, 1993; Baddeley, 1997). As well as using more powerful modes of remembering, it is possible to use authentication items (in place of passwords) that are more memorable, without being guessable. These include pass-sentences (Spector and Ginzberg, 1994; longer strings of meaningful words), associative passwords (Zviran and Haga, 1993; a form of cued recall), and Passfaces, which utilise recognition of images rather than recall of words.

1.4 Passfaces

The enrolment procedure allows users to first select whether their Passface set is male or female. They then select 4 faces, and are directed to consider the characteristics of their selections, and why they selected them. The users are then twice taken through the Passfaces login procedure, with their Passfaces indicated to them. They complete enrolment by correctly identifying their 4 Passfaces twice in a row with no prompting, then (in this field trial only) entering an enrolment password.

To log in, users select their Passfaces from a grid of faces displayed on the screen. This study uses the standard implementation of the Passfaces demonstration toolkit, requiring participants to memorise 4 faces, and correctly select all 4: one in each of 4 grids of nine faces (see Figure 1 for an example grid). The grids are presented one at a time on the screen, and the order of presentation remains constant, as do the faces contained in each grid. However, no grid contains faces found in the other grids, and the order of faces within each grid is randomised. These features help secure a user's Passface combination against detection through shoulder-surfing and packet-sniffing.

Passfaces were shown to be memorable in a study involving 77 staff and students of Goldsmiths College (Valentine, 1998). All participants went through the Passfaces enrolment procedure, and 3 conditions were tested. The first condition had 29 participants logging in every working day for 2 weeks. Participants correctly recalled their Passfaces in 99.98% of logins. The second condition had 29 participants log in approximately 7 days after enrolment. On their first attempt, 83% logged in successfully. Everyone in this condition logged in by his or her 3rd attempt. The third condition had 19 participants login only once approximately 30 days after enrolment, with 84% of participants remembering their Passfaces at the first attempt, and the remainder remembering their Passfaces by the third attempt.

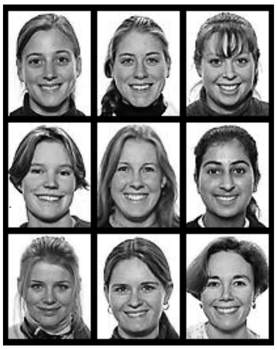

Figure 1: Example Passfaces grid

Passfaces have also proved to be memorable over long periods without use[2]. The participants were contacted and asked to log in again on average 5.4 months after they had last used their Passfaces. 56 participants completed the follow up study. Overall, 72% of participants remembered their Passfaces on the first attempt, and 84% had remembered their Passfaces by the third attempt. Participants who had originally been in the everyday use condition remembered their Passfaces the best, with 87% remembering them at the first attempt and 100% by the third attempt (Valentine, 1999). There have been similar studies of password memorability. In most cases the password is selected by the participant, who is then asked to recall it after an interval that varies between studies. The intervals and the resulting memorability are shown in Table 1.

A comparison of the results (Table 1) suggests that Passfaces are more memorable than traditional passwords, and hence a solution to the usability problem described above. However, there has been no direct comparison of Passfaces and

[2] Password resets are most often required after holidays. Internal helpdesks and those of Internet Service Providers experience a surge of calls after the Christmas break and the end of the holidays. In one case, more than 60% of calls to the help desk were due to forgotten passwords.

passwords. The participants were different and the intervals over which the words and faces were recalled were of different lengths. A final concern is that the situation under which the mechanisms were tested was somewhat artificial – users were prompted by experimenters to log in, rather than observed using the mechanism to access the systems in the context of their normal activities.

Interval	Passwords (% remembered)	Passfaces (% remembered)	Study
1 day	-	99.98 (1st attempt)	Valentine, 1998
1 week	-	100 (by 3rd attempt)	Valentine, 1998
2 weeks	77 (1st attempt)	-	Bunnell et al, 1997
1 month	-	100 (by 3rd attempt)	Valentine, 1998
3 months	35 (1st attempt	-	Zviran and Haga, 1990
3 months	27.2 (1st attempt)	-	Zviran and Haga, 1993
5 months	-	72 (by 3rd attempt)	Valentine, 1998

Table 1 - Memorability of passwords and Passfaces over different intervals

The goal of the study reported in this paper is to compare passwords and Passfaces with the same participants in a field trial, with participants using authentication mechanisms to gain access to a real system as part of a real task.

2 The Field Trial

2.1 Participants

Thirty-six first year undergraduate students in Information Management taking a one-term course in Systems Analysis participated in the trial. Thirty-four students logged in frequently enough to be included in the study.

2.2 Trial context

As part of the course, the students had to complete 6 assignments on-line on the Web over a period of 10 weeks; the coursework was authored and managed through the TACO system (Sasse et al., 1998). To interact with TACO, students used computers at the University (mostly PCs running Windows 3.1 with 486 processors, and some Macs running MacOs 8.1 with 601 processors) or from home. Students can practise each question set as often as they like before submitting an assessed version; since they receive scores and feedback, frequent practice tends to result in better grades. Users of TACO are required to go through authentication before being allowed to interact with courseworks. Logging in usually consists of being identified by entering a username, and then authenticated by entering a system-generated password (both of which had been previously supplied to the user through a secure channel). A facility allowing participants to change passwords and a facility to select and use Passfaces were added to TACO. Each of these facilities required participants to supply the system-generated password before they could select their user-generated password or Passfaces set. In addition, the modified authentication system allowed logging of interactions taking place during the login (i.e. keystrokes), thus making it possible to count successful and failed logins, and reconstruct what participants typed.

2.3 The System

The password mechanism of TACO is executed at the server side, and appears instantaneous to users. However, the Passfaces mechanism is executed at the user's computer. There are two versions, one using *Active X* and the other using *Java* technologies. Both require users initially to download the Passfaces mechanism and the Passfaces themselves. *Active X* allows these to be stored on the user's computer, such that subsequent uses of Passfaces appear instantaneous, but is not supported by the versions of the web-browsers (Netscape 3 and 4.03) available on the computers commonly used by this cohort of students. Since the *Java* version does not support local storage, the Passfaces set and mechanism must be downloaded for each log in. The user must wait for this download across busy university networks. In addition, for each initial use of the Passfaces system in a session, the user must wait for a software package to load that converts the *Java* to a working Passfaces mechanism. On the slowest computers available to participants, a Passface login took up to 3 minutes, while a password login was completed in seconds[3].

Passface enrolment has been described in section 1.4. In contrast to Passfaces, the password enrolment procedure is relatively brief. Guidance about the selection of cryptographically strong password content is displayed on screen, and users are required to submit a password of their choosing twice, and their enrolment password. If the enrolment password is correct, and both submissions of the chosen password match, then enrolment is complete.

TACO was further changed to offer participants reminders of their passwords/Passfaces. On their request, participants were emailed a copy of their password, or sent the address of a web page where they could view their Passface. TACO log files were enhanced so that the failure or success of login attempts could be determined, and all requests for reminders were recorded.

2.4 Procedure

A repeated-measures design was used, with each student using both passwords and Passfaces. The design was counterbalanced to take account of order effects, with half the participants using the mechanisms in order PW-PF (passwords then Passfaces) and the other half in order PF-PW (Passfaces then passwords). This maximised power for the test of difference between Passfaces and passwords (a simple between-groups design would have insufficient power to detect differences between Passface and passwords given the relatively small sample size).

[3] Participants were pointed towards more powerful computers on which the Passfaces login was much faster, but the majority of students continued accessing the system from the old machines. The specification of machine used not only affected the speed of the login, but also the time it took to complete the coursework exercises.

Participants were pseudo-randomly assigned to PW-PF or PF-PW with the aid of a random number table. Participants were given enrolment passwords (on paper slips) at the start of term, to authenticate them for their subsequent selection of passwords or Passfaces. Participants used the web-based coursework system as normal to complete course-works.

Halfway through the term (marked by the lecture-free Reading Week) the authentication mechanisms used by students were changed over (those using passwords were now using Passfaces, and vice versa). Participants were required to re-enrol, and *new* enrolment passwords were distributed by email to all participants and on paper slip by request.

3 Results

Logins and login problem rates were analysed, followed by reminders, time taken before first use of the system, and number of logins. This paper will describe each variable's effect using the effect size indicator *d*, which is the distance in standard deviations between the means of two groups. The following sections may describe differences between groups that are small in absolute terms, counter-intuitively, as being due to large effects. In the context of effect sizes, *large, medium* and *small* have technical definitions (see Rosenthal and Rosnow, 1991 or Clark-Carter, 1997, for further information about effect size and statistical power).

3.1 Problem rates

Task performance is an important part of usability, and is often measured by time and errors. In the context of passwords, an easy type of error to record is a failed login attempt. Failed login attempts are user costs, and so should be minimised where possible. If two people have the same number of failed logins but different numbers of successful logins then counting the absolute number of failed login attempts is misleading. We will therefore use login failure *rate* as one of our measures of usability.

The numbers of successful and unsuccessful logins were used to calculate each participant's failure rate for logging in (problems÷(problems+successes)) for both passwords and Passfaces. Table 1 shows descriptive statistics for login failure rate, with authentication mechanism and order of presentation of authentication mechanism as the independent variables.

Passwords had a login failure rate of 15.1%, while Passfaces for the same participants produced a login failure rate of 4.9%. Thus, the number of login problems occurring with Passfaces was approximately a third that of passwords.

A mixed ANOVA was performed on the data, testing authentication mechanism (repeated measure) and order of presentation (between groups measure) as the main effects and the interaction between them (Table 2). The test of authentication mechanism achieved a power of .924 (better than the recommended level), and showed that the difference between Passfaces' and passwords' error rate was highly significant ($F_{(1,31)}$=12.31, p=.001), and that authentication mechanism had a large effect on login problem rate (d=1.26).

Group	Mean	N	95% CI for Mean	Std Err	Std Dev	Min	Max
Mechanism							
Passwords	0.15	33	.09/.21	0.03	0.16	0	0.57
Passfaces	0.05	34	.02/.08	0.02	0.09	0	0.38
Order							
PW-PF	0.06	34	.03/.08	0.01	0.07	0	0.29
PF-PW	0.14	33	.08/.20	0.03	0.17	0	0.57
Order x Mechanism							
PW-PF							
• Passwords	0.07	17	.03/.11	0.02	0.08	0	0.29
• Passfaces	0.04	17	.01/.08	0.02	0.07	0	0.20
PF-PW							
• Passwords	0.24	16	.12/.33	0.05	0.18	0	0.57
• Passfaces	0.05	17	.00/.11	0.01	0.11	0	0.38

Table 2 - Descriptive statistics of login error rates. Mechanism, Order, and Mechanism x Order are significantly different

The ANOVA achieved a power of .861 in testing order of presentation, slightly bettering the recommended value. Order effects were not predicted but there was a highly significant difference ($F_{(1,31)}$=9.92, p=.004) between the login problem rates of those who were presented with passwords first (PW-PF) and those who were presented with them last (PF-PW). Order of presentation had a large effect (d=1.13).

Source of Variation	SS	DF	MS	F	Sig of F
(Between groups measure)					
WITHIN+RESIDUAL	.44	31	.01		
MECHANISM	.17	1	.17	12.31	.001
ORDER BY METHOD	.09	1	.09	6.68	.015
(Repeated measure)					
WITHIN+RESIDUAL	.42	31	.01		
ORDER	.13	1	.13	9.92	.004

Table 3 - ANOVA table for login error rates

The ANOVA, operating at .704 - slightly less than recommended power, also showed that the effect of order of use was different for each authentication mechanism ($F_{(1,31)}$=6.68, p=.015). This difference was large (d=.93). The PF-PW group had a login error rate of 5.5% with Passfaces, whilst the PW-PF group had an error rate of 4.3%. Thus, PF-PW had an error rate more than 25% higher than PW-PF.

The difference was in the other direction for passwords. The PW-PF group had a password login error rate of 7.1%; while PF-PW had an error rate of 23.6%. The error rate for PW-PF was less than a third of that of PF-PW. Using PF-PW had a detrimental effect on both their password and Passfaces login failure rate, but much more so for passwords.

As explained in 2.2, the experimental apparatus captured failed password login attempts. By comparing the failed attempt with the participant’s correct password, problem types could be inferred. This helps to diagnose the causes of password login problems, and to prioritise them. Table 4 shows the relative frequencies of password problems encountered during the experiment. The most frequent problem was entering a previous TACO password in place of the current one. The next most frequent problem was substituting a password-like sequence for the correct password.

Problem type	Proportion
Previous (TACO) password used	37%
Other password used	15%
‘ENTER’ only	9%
Character missing	6%
Additional character	5%
Part of password only	5%
Admin problem	4%
System problem	3%
Wrong character	2%
2 passwords mixed	1%
Capitals not used correctly	1%
User ID entered instead of password	1%

Table 4 - Password problems encountered by participants

Separate analyses of login error rates for each error type would be preferable to the lumped together measure employed here, as would analysis of Passface error types. Such analysis would suffer from the small data set available to this study, and technical issues prevented the recording of Passface login errors. These analyses will, however, become feasible when data sets from studies in progress are added to the data presented here.

3.2 Reminders

The previous section measured the numbers and types of errors made in task completion – logging in. This section looks at the prevalence of not being able to do the task – giving up on logging in and calling in the helpdesk. This is an important measure of authentication mechanism usability, because it is such a large cost to industry (see section 1).

Participants in this experiment were offered a facility to have a reminder of their password or Passfaces sent to them by e-mail. Automatic reminders are now widely employed in many E-commerce systems as a means of reducing the number of password-related calls to helpdesks, even though sending passwords in unencrypted email is not secure.

Descriptive statistics of password/Passface reminders are shown in Table 5. A mixed ANOVA was performed on the data, with dependent variable being the number of reminders per participant, and the independent variables authentication mechanism (repeated measure) and order in which the mechanisms were used (between groups measure). The results are shown in Table 6.

When using passwords, users requested .14 reminders on average, approximately two thirds more than when using Passfaces (mean of .09). However, the difference is not significant ($F_{(1,32)}$=.27, p=.605). Power analysis showed that the ANOVA only achieved a power of .055 in testing the effect of mechanism, and therefore had only a 5.5% chance of detecting a real effect. Further analysis showed that mechanism had a small effect on number of reminders (d=.18). Rosenthal and

Rosnow (1991) show that 400 participants would be required for a test to achieve significance at the .05 level (2 tailed) for an effect of this size. If this small effect were real, then multiplied by the large scale of corporate computer use, Passfaces could make an appreciable difference in helpdesk costs.

Group	Mean	N	95% CI for Mean	Std Err	Std Dev	Min	Max
Mechanism							
Passwords	0.15	34	-.05/.34	0.10	0.56	0	3
Passfaces	0.09	34	-.01/.19	0.05	0.29	0	1
Order							
PW-PF	0.00	34	0/0	0.00	0.00	0	0
PF-PW	0.24	34	.02/.45	0.10	0.61	0	3
Order x Mechanism							
PW-PF							
• Passwords	0.00	17	0/0	0.00	0.00	0	0
• Passfaces	0.00	17	0/0	0.00	0.00	0	0
PF-PW							
• Passwords	0.29	17	-.10/.69	0.20	0.77	0	3
• Passfaces	0.18	17	-.03/.38	0.10	0.39	0	1

Table 5 - Descriptive statistics of password/Passfaces reminders

Order of use of the authentication mechanisms had a large effect on participants' mean number of reminders (d=.85). The PF-PW group had a mean number of .24 reminders, whereas PW-PF required none. The ANOVA reached a power of .65 in testing order of use. This difference between the groups was unexpected.

Whereas login error rates showed an interaction between the effects of order of use and authentication mechanism, reminders did not ($F_{(1,32)}$=.27, p=.605). As with the effect of method on participants' mean number of reminders, the test for order x mechanism achieved a power of only .05, having a small effect (d=.18) that if real would have required 400 participants to detect.

Source of Variation	SS	DF	MS	F	Sig of F
(Within groups effects)					
WITHIN+RESIDUAL	6.88	32	.22		
METHOD	.06	1	.06	.27	.605
ORDER BY METHOD	.06	1	.06	.27	.605
(Between groups effects)					
WITHIN+RESIDUAL	5.12	32	.16		
ORDER	.94	1	.94	5.89	.021

Table 6 - ANOVA table for password/Passfaces reminders

3.3 Time before first use

The popularity of a system may be measured by the speed with which users adopt it. In a work-related piece of software such as a login mechanism, usability is a good starting point for popularity. The speed with which people took up each authentication mechanism could be viewed as an indicator of usability. This is particularly the case in a domain where time is limited – students have coursework deadlines which must be met.

Participants' coursework consisted of multiple-choice and free-response questions that were distributed, responded to by participants, marked and corrections displayed all via web pages. Practice questions were made available (practice coursework), which participants could use whenever and as often as they wished and for which marks were not formally recorded. We observed that Passface users were waiting longer before submitting practice or assessed coursework than password users. Data regarding the date of first use of the system were collected from system logs, and descriptive statistics for these are shown in Table 7. These same data were added to a mixed ANOVA, the results of which are shown in Table 8.

Group	Mean	N	95% CI for Mean	Std Err	Std Dev	Min	Max
Mechanism							
Passwords	16.36	33	13.98/18.75	1.17	6.72	5	36
Passfaces	20.00	34	17.84/22.16	1.06	6.20	11	35
Order							
PW-PF	17.38	34	14.94/19.83	1.20	7.00	5	32
PF-PW	19.06	33	16.83/21.29	1.10	6.30	11	36
Order x Mechanism							
PW-PF							
• Passwords	11.18	17	9.47/12.88	0.81	3.32	5	15
• Passfaces	23.59	17	22.11/25.07	0.70	2.87	17	32
PF-PW							
• Passwords	21.88	16	19.40/24.35	1.16	4.65	15	36
• Passfaces	16.41	17	13.01/19.81	1.60	6.61	1	35

Table 7 - Descriptive statistics for day of first use of target application

As in previous sections, the independent variables were authentication mechanism (repeated measure) and order of use of authentication mechanisms (between groups measure). As authentication mechanism is a repeated measure, each participant experiences first use of a system twice, once for passwords and once for Passfaces.

Passfaces were first used on average 4 days later than passwords. The ANOVA achieved better than recommended power (.859) and showed that this difference was highly statistically significant ($F_{(1,63)}$=9.55, p=.003), and was equivalent to a medium sized effect (d=.78). This finding shows a usability advantage for passwords, where previous sections gave the advantage to Passfaces. A synthesis of these apparently contradictory results can be achieved by examining evidence from the number of logins made (section 3.4 below) and from participants' anecdotes.

The ANOVA did not detect a significant difference in first use dates due to order of use ($F_{(1,63)}$=2.45, p=.112). However, the observed effect size was small (d=.39). Due to the limited number of participants available, the test achieved a power of only .338 (a one in three chance of detecting a real effect) for the order of use effect. Assuming this small effect does exist, it would require 400 participants to detect (cf. Rosenthal and Rosnow, 1991).

Because each participant contributed data for 2 first courseworks (one for each authentication method) and the two deadlines were on different days (days 15 and 24) we would predict a strong interaction effect between authentication mechanism and order of mechanism use. This is in fact the case. As this is merely an artefact of the experimental design, it will not be further reported.

Source of Variation	SS	DF	MS	F	Sig of F
WITHIN+RESIDUAL	1332.46	63	21.15		
ORDER	51.91	1	51.91	2.45	.122
MECHANISM	202.04	1	202.04	9.55	.003
ORDER BY MECHANISM	1337.05	1	1337.05	63.22	.000
(Model)	1602.62	3	534.21	25.26	.000
(Total)	2935.07	66	44.47		

Table 8 - ANOVA table for day of first use of target application.

3.4 Number of login attempts

To help interpret the results of the time before first use analysis, descriptive (Table 9) and inferential statistics (Table 10) were calculated for the number of login attempts for each participant. The experimental apparatus counted a login attempt as a successful or unsuccessful submission of passwords/Passfaces. It could not record logins interrupted before a password or Passface was entered. For example, a login attempt was not recorded if cancelled by a participant while Passfaces were downloading.

Group	Mean	N	95% CI for	Std	Std	Min	Max
			Mean	Err	Dev		
Passwords	33.91	34	27.00/40.83	3.40	19.81	0.00	92.00
Passfaces	12.32	34	9.92/14.73	1.18	6.88	2.00	29.00

Table 9 - Descriptive statistics for number of login attempts

Overall, the authentication mechanism had a large effect (d=2.6) on the number of logins attempted. Participants attempted to use the coursework system with Passfaces approximately a third of the amount they attempted to use it with passwords ($F_{(1,32)}$=53.92, p=.000, highly significant; observed power=1.0).

Source of Variation	SS	DF	MS	F	Sig of F
(Within groups effects)					
WITHIN+RESIDUAL	4702.18	32	146.94		
METHOD	7922.88	1	7922.88	53.92	.000
ORDER BY METHOD	52.94	1	52.94	.36	.553
(Between groups effects)					
WITHIN+RESIDUAL	9691.00	32	302.84		
ORDER	72.06	1	72.06	.24	.629

Table 10 - ANOVA table for number of attempted logins recorded (dependent variable) x order x authentication mechanism (independent variables)

There were no significant order effects ($F_{(1,32)}$=.24, p=.63, not sig.; observed power=.06), or interaction effects between order and authentication mechanism ($F_{(1,32)}$=.36, p=.55, not sig.; observed power=.36). Should these small effects (d=.17 and .21 respectively) exist, 800 participants would be required for a similar ANOVA to class them as statistically significant.

Not only was there a delay before Passfaces were used, they were used less frequently. The *time* part of our analysis suggests a usability problem for Passfaces in this field trial, but the *errors* analysis shows them to have a usability advantage. Passfaces could be said to trade some login speed for greater memorability. It is argued below that several factors greatly exaggerated this trade off, causing one usability problem whilst solving another.

4 Discussion

4.1 Performance of the authentication mechanisms

There was no significant difference between the number of reminders asked for by participants when using passwords or Passfaces. This measure of usability is relevant to helpdesks – where forgotten passwords would need to be reset. From this perspective the mechanisms appear to be equal (for users similar to those in the study). From the participants' point of view however, they are not equal.

Passfaces had a login problem rate of less than a third of the login problem rate of passwords in this study. In particular, the PF-PW group of participants experienced nearly a 1 in 4 password login failure rate, 3 times higher than PW-PF's. This finding is unexpected - how might it be explained?

Participants were randomly assigned to PW-PF or PF-PW groups, so the differences should not be due to participant differences between groups. Since every participant used both passwords and Passfaces, every participant also necessarily underwent a transition from using one mechanism to the other. It is likely that something related to this transition is responsible.

The protocol used to move participants from the mechanism they first used to the second was different to the protocol used to start them on the first mechanism at the beginning of the experiment. During induction, participants were informed of enrolment details via paper slips handed out in the first few lectures. During changeover, participants received emails with their changeover information, in addition to verbal and hard copy notification. It seems, however, unlikely this difference in procedure could have caused such severe problems for password users.

We know that the largest problem encountered by password users was attempting login with defunct passwords (Table 4). Why should participants in the PF-PW group be more susceptible to password confusion than participants in the PW-PF group?

It is a counter-intuitive finding. A simple hypothesis would be that people have problems changing passwords, because they confuse the password they *previously* used for the one they should now be using. However, this would not explain the large difference between the groups, who both had opportunities to make the slip.

Using this hypothesis we may even predict a difference in the *opposite* direction to the one found. Assuming a schema model of human performance such as Reason' s(1990) (and that passwords are schemata), participants who changed from passwords to Passfaces (PW-PF group) would have had their self-selected passwords at high levels of activation due to frequency and recency of use and links to related schemata, and so these passwords should offer high levels of interference to recall and use of the current password (but do not). In contrast, the (PF-PW) group which suffered extreme interference in password recall and use suffered it at the expense of an enrolment password that had been used once more than a month previously, and which being a non-word would have little relation to other contents of participants' memories.

A second possible explanation of the finding is that password and Passfaces use are competing skills, and that Passface use de-skills the participant in password use. This would effect all passwords systems participants may use. To test this hypothesis, each group should undergo the suspect transition and contribute data from more than one password protected system. To assess the de-skilling effect's duration might require repetitions of the suspect transition, and observation over longer periods.

A similar but alternate interpretation is that Passface use inhibits password use on the system in which Passfaces were previously used. For example, assuming Passfaces are easier to remember than passwords, participants may use the same effort in processing passwords that they used for the Passfaces – leading to an insufficiently deep level of password processing, and so poorer memory for the password. We feel that this explanation is unlikely, as participants were often alarmed at the prospect of having to remember the faces – and so would be likely to process the Passfaces to a deep level.

Another explanation of the finding is possible: that Passfaces are simply more resistant to confusion at changeover time than passwords. The Passfaces patent holding company intends to control their distribution to minimise possible confusion between different sets of Passfaces (Barratt, 1999, personal communication). However, to properly support this explanation of the results would require a new experiment in which both authentication mechanisms are equivalently used; with system chosen Passfaces protecting participants' Passface selection (as passwords protect password selection). This would allow comparisons of mistaken use of defunct secrets in the two systems. In the wider context, this would assess the untested claim that Passfaces are resistant to confusion with previous/other Passfaces.

Overall, login failures are user-costs, and so should be minimised. There were no restrictions on login attempts in this field trail. In industrial contexts the consequences may be more severe – authentication systems may enforce delays between login attempts, or "*3 strikes*" policies to reduce the password guessing opportunities of hackers. The present findings have a more detrimental effect in such settings.

4.2 Login frequency

The strengths of the present study have been to provide detailed observations of both authentication mechanisms in longer-term use with real users and real tasks in a real system environment. However, the reality of the study environment (many of the machines available to the participants were old and underpowered) led to significantly longer login times with Passfaces. When using Passfaces, participants attempted to login with a third of the frequency with which they did using passwords. They also started their attempts a mean of four days later than when using passwords. When using Passfaces participants, therefore, had less practice for coursework and had less opportunity for practice. This did not reduce their final mark (when using Passfaces participants scored 6% higher though the difference was not significant, $F_{(1,202)}$=3.71, p=.56, d=.27 small effect, power=.48). Whilst the detailed impact of this usage bias needs to be explored in a future study, its existence also demonstrates the importance of evaluating the usability of security mechanisms in context.

Combining our knowledge of the study environment with anecdotal evidence (several participants commented unfavourably to their course lecturer on the time taken to login with Passfaces) suggests an explanation for the delayed and reduced Passface use. On finding Passface use to be slow on college facilities (anecdotal evidence), participants abandoned their attempts to use them (reduced use, see section 3.4) until close to the deadline for submitting the coursework (delayed used, see section 3.3).

Anecdotal evidence suggests that the use of more up-to-date computing facilities can lead to dramatic gains in user acceptance: a participant who had strenuously objected to using Passfaces because of slow and erratic system response during enrolment attempted it with faster equipment (PII 300/Win95) after a discussion with the experimenter. The participant withdrew all objections.

4.3 Psychology of authentication mechanisms

The primary component of any knowledge-based authentication system is human memory. Psychology has much substantive knowledge that can be used to explain password problems, and in intervention. Being a real world activity, password use involves a complicated knit of contextual factors as well as the laboratory capabilities and mechanisms of human memory. Understanding the current field trial would involve partitioning the effects of at least: levels of processing (Craik and Lockhart, 1972), pro-active (Baddeley, 1997), retro-active (Tulving and Psotka, 1971) and within-list interference (Wickens, 1992), free and cued recall (Parkin, 1981) and recognition (Parkin, 1993), whether the item being remembered is a word, picture (Nelson et al., 1977) or face (Bahrick et al., 1975), group working practices and perceptions of threat (Adams and Sasse, 1999), the use of prompts (Cohen, 1996), and individual differences such as absent-mindedness (Reason, 1990). A thorough understanding of the psychology underlying the remembering of secrets in knowledge-based authentication will require a research program spanning these and more topics.

5 Conclusions and Further Work

5.1 Passfaces

Passfaces showed a third the login failure rate of passwords, despite having users with a third the frequency of use (less frequent means the memory task was more difficult). This performance difference was partly due to the password confusions of participants who had recently changed from Passfaces to passwords. While Passfaces' low error rate may be due to their superiority over passwords, there are other explanations that need to be ruled out.

Passfaces have been shown to be very memorable over long intervals in previous studies (Valentine, 1998; Valentine, 1999). Implemented appropriately (with more powerful computers and in *ActiveX* rather than *Java*), we predict that Passfaces would offer better performance than passwords for users who log in infrequently (less than once every two weeks).

Passfaces are a security mechanism designed with many theoretical advantages over passwords. They have been tested in previous studies under laboratory conditions and shown to perform well. This study tested Passfaces and passwords in a group of real users' work contexts, and with a number of unpredicted results. Consideration of task and environmental context in which a system is used is a fundamental part of human-computer interaction methods. However, security research and implementation do not often concern themselves with user costs, nor consider the context of system use as their source (Adams and Sasse, 1999). Developers ignore contextual factors at their peril; this study reminds us that evaluators do also. Security mechanisms designed and *tested* outside users' work contexts may shine on paper and in laboratory settings, yet may behave unexpectedly in practice.

5.2 Passwords

In this study, password users experienced substantial login failure rates (in one condition as high as 1 in 4 attempts failing). Passwords can therefore have user costs beyond the resets observed by computer helpdesks. Whilst user report data has identified similar problems to the present study (Adams and Sasse, 1999), the extent of failure had not been quantified in the security or HCI literature to date. This study therefore represents a step forward in the evaluation of user-authentication mechanisms, and computer security mechanisms more widely.

Research of this kind, however, is likely to remain rare. Security personnel and systems administrators are duty bound to prevent dissemination of data that might aid attackers. This makes even the collection of security system usability data, such as the capturing of failed login attempts, possible only in unusual circumstances.

5.3 Further work

Passfaces implementation with older computing services may have led to reduced and delayed system use. The increasing power of computing infrastructure is inevitable. As the increase occurs, the resources that Passfaces require will become ubiquitous. Passfaces should therefore be tested with up to date hardware and software facilities – recent CPUs and web-browser *Active X* support. If these facilities are not available, speed of the authentication mechanisms' responses to user input should be measured and included in analyses, and response times made similar by retarding the password mechanism.

This experiment raised the possibility that Passface use interfered with password use. To assess this possibility an experiment is needed that authenticates Passface selection with Passfaces, and which repeats the transition from Passfaces to passwords.

The study raised the issue of confusing previous authentication secrets with new, when this was found to be participants' largest source of password error. Although Passfaces have been designed to reduce similar confusions, their ability to do this has not been tested. Studies are required of the relative effects of transition from previous to new secrets in both passwords and Passfaces. More widely, studies should be made of the contributions of different psychological phenomena to authentication mechanism usability.

Every user has at least one password story, and user reports are easy to gather. Future studies should augment user reports with objective data, even though it is hard to obtain.

Acknowledgements

Sacha Brostoff's PhD research is funded by a grant from British Telecom Plc. The authors would like to thank Charles Brennan for his contribution to the research project of which this is a small part. We would also like to thank Id-Arts Ltd. for permitting us to use Passfaces, and for their support in adapting and integrating the Passfaces demonstration toolkit into the experimental apparatus.

References

Adams, A. (1996). *Reviewing Human Factors in Password Security Systems.* Unpublished M.Sc., University College London, London.

Adams, A. and Sasse, M. A. (1999). "Users Are Not the Enemy: Why Users Compromise Security Mechanisms and How to Take Remedial Measures." *Communications of the ACM* **42**(12), 40-46.

Adams, A., Sasse, M. A. and Lunt, P. (1997), *Making Passwords Secure and Usable*, *in* H. Thimbleby, B. O' connailland P. Thomas (Eds.), HCI ' 97- People and Computers XII, pp.1-20. Springer-Verlag, Bristol

Anderson, R. J. (1994). “Why Cryptosystems Fail.” *Communications of the ACM* **37**(11), 32-40.

Arthur, C. (1997, Tuesday 2nd December). Your Eye. The Ultimate Id Card. *The Independent*, p. 1.

Baddeley, A. (1997), *Human Memory: Theory and Practice*, Revised edition, Psychology Press.

Bahrick, H. P., Bahrick, P. O. and Wittlinger, R. P. (1975). “Fifty Years of Memory for Names and Faces: A Cross-Sectional Approach.” *Journal of Experimental Psychology: General* **104**(1), 54-75.

Clark-Carter, D. (1997). “The Account Taken of Statistical Power in Research Published in the British Journal of Psychology.” *British Journal of Psychology* **88**(1), 71-83.

Cohen, G. (1996), *Memory in the Real World*, second edition, Psychology Press.

Craik, F. I. M. and Lockhart, R. S. (1972). “Levels of Processing: A Framework for Memory Research.” *Journal of Verbal Learning and Verbal Behavior* **11**(6), 671-684.

Davis, C. and Ganesan, R. (1993), *BApasswrd: A New Proactive Password Checker*, Proceedings of the National computer security conference '93, the 16th NIST/NSA conference, pp.1-15., September,

Deane, F., Barrelle, K., Henderson, R. and Mahar, D. (1995). “Perceived Acceptability of Biometric Security Systems.” *Computers and Security* **14**(3), 225-231.

Garfinkel, S. and Spafford, G. (1996), *Practical Unix and Internet Security*, second edition, O'Reilly & Associates.

Kim, H.-J. (1995). “Biometrics, Is It a Viable Proposition for Identity Authentication and Access Control.” *Computers and Security* **14**(3), 205-214.

Menkus, B. (1988). “Understanding the Use of Passwords.” *Computers and Security* **7**(2), 132-136.

Murrer, E. (1999). “Fingerprint Authentication.” *Secure Computing* (March), 26-30.

Nelson, D. L., Reed, U. S. and Walling, J. R. (1977). “Picture Superiority Effect.” *Journal of Experimental Psychology: Human Learning and Memory* **2**(5), 523-528.

Obaidat, M. and Sadoun, B. (1997). “Verification of Computer Users Using Keystroke Dynamics.” *IEEE transactions on systems man and cybernetics part B-Cybernetics* **27**(2), 261-269.

Parkin, A. J. (1981). “Determinants of Cued Recall.” *Psychological Research* **1**(4), 291-300.

Parkin, A. J. (1993), *Memory: Phenomena, Experiment and Theory*, Blackwell.

Reason, J. (1990), *Human Error*, Cambridge University Press.

Roddy, A. R. and Stosz, J. D. (1997). “Fingerprint Features - Statistical Analysis and System Performance Estimates.” *Proceedings of the IEEE* **85**(9), 1390-1421.

Rosenthal, R. and Rosnow, R. (1991), *The Essentials of Behavioural Research*, second edition, McGraw Hill Book Co.

Sasse, M. A., Harris, C., Ismail, I. and Monthienvichienchai, P. (1998), Support for Authoring and Managing Web-Based Coursework: The TACO Project. *in* R. Hazemi, S. Hailes and S. Wilbur (Eds.), *The Digital University: Reinventing the Academy*, Springer-Verlag, pp.155-175.

Spector, Y. and Ginzberg, J. (1994). “Pass Sentence - a New Approach to Computer Code.” *Computers and Security* **13**(2), 145-160.

Svigals, J. (1994). “Smartcards - a Security Assessment.” *Computers & Security* **13**(2), 107-114.

Tulving, E. and Psotka, A. (1971). “Retroactive Inhibition in Free Recall: Inaccessibility of Information in the Memory Store.” *Journal of Experimental Psychology* **87**(1), 1-8.

Valentine, T. (1998). *An Evaluation of the Passface^{tm} Personal Authentication System* (Technical Report). London: Goldmsiths College University of London.

Valentine, T. (1999). *Memory for Passfaces^{tm} after a Long Delay* (Technical Report). London: Goldsmiths College.

Wickens, C. (1992), *Engineering Psychology and Human Performance*, second edition, Harper Collins.

Zviran, M. and Haga, W. J. (1990). “Cognitive Passwords: The Key to Easy Access Control.” *Computers and Security* **9**(8), 723-736.

Zviran, M. and Haga, W. J. (1993). “A Comparison of Password Techniques for Multilevel Authentication Mechanisms.” *The Computer Journal* **36**(3), 227-237.