Age-Related Performance Issues for PIN and Face-Based Authentication Systems

James Nicholson  
PaCT Lab  
Northumbria University  
Newcastle, UK  
james.nicholson@unn.ac.uk

Lynne Coventry  
PaCT Lab  
Northumbria University  
Newcastle, UK  
lynne.coventry@unn.ac.uk

Pam Briggs  
PaCT Lab  
Northumbria University  
Newcastle, UK  
p briggs@unn.ac.uk

ABSTRACT
Graphical authentication systems typically claim to be more usable than PIN or password-based systems, but these claims often follow limited, single-stage paradigm testing on a young, student population. We present a more demanding test paradigm in which multiple codes are learned and tested over a three-week period. We use this paradigm with two user populations, comparing the performance of younger and older adults. We first establish baseline performance in a study in which populations of younger and older adults learn PIN codes and we follow this with a second study in which younger and older adults use two face-based graphical authentication systems employing young faces vs. old faces as code components. As expected, older adults show relatively poor performance when compared to younger adults, irrespective of the authentication material, but this age-related deficit can be markedly reduced by the introduction of age-appropriate faces. We conclude firstly that this paradigm provides a good basis for the future evaluation of memory-based authentication systems and secondly that age-appropriate face-based authentication is viable in the security marketplace.

Author Keywords
Graphical codes; usable security; authentication; older adults

ACM Classification Keywords

General Terms
Human Factors; Security.

INTRODUCTION
User authentication is a requirement for secure environments that provide access to confidential information or offer personalised services. The most common authentication systems are knowledge-based and typically rely upon the users ability to commit a password or PIN to memory. Problems arise because the user typically needs to generate passwords and PINs for multiple accounts. The memory load becomes onerous, with the result that poor authentication practices prevail. Users duplicate passwords, choose family names or keep stronger codes written down in a drawer or wallet [1,20,23]. Authentication can be challenging to all users, but is particularly problematic for older adults who show a notable decline in memory capacity past the age of fifty (e.g. [10,11,32]).

Graphical authentication (GA) systems have been developed as a potential solution to the usability problems posed by PIN and password systems. In a typical recognition-based GA system, the user is presented with a set of images, a subset of which form that user’s authentication code. To log in the user must successfully respond to a series of challenges in which they must select their target image from a set of distractors or decoys. One well-known example of such a system is Passfaces [39], in which the authentication codes comprise a set of faces that the user must recognise from a grid of distractor faces.

There have been a number of evaluations of GA systems in the research literature, but these evaluations are often conducted with sets of relatively young users, who could arguably represent the most able population in terms of memory performance. In addition, usability testing has often been conducted in single laboratory sessions, sometimes but not always involving the requirement to learn multiple codes. This is important, as one of the key memory demands in this space is the requirement to remember which code is associated with which account and interference effects are rife.

In this paper, we have three objectives: Firstly, we wish to introduce a robust test paradigm for new, purportedly ‘usable’ authentication systems. To this end we introduce a test procedure in which users learn multiple codes in two batches and are repeatedly tested (weekly) on these codes over a period of three weeks. Secondly, we wish to explore the extent to which GA can provide an inclusive solution for older as well as younger adults. Finally, we wish to make design recommendations that are realistic in their technical demands and could legitimately be rolled out to a large population.
BACKGROUND RESEARCH

Passwords and PINs are generally assumed to be more difficult to remember than recognition-based GA codes because they rely on the more cognitively demanding process of recall as opposed to recognition [24]. GA systems typically ask users to recognise a set of target stimuli when they are set within a context of multiple distractors, although some utilise a click-based approach where users are required to recreate the selection of specific location-based points within a single or series of images, one example being PassPoints [8,42].

Recognition-based GA systems usually employ pictorial stimuli, including photographs [13] or digital art [15], playing to another known aspect of human cognition: a picture superiority effect in which concepts are more likely to be remembered if they are presented as pictures rather than as words [30]. In this paper we are particularly interested in those GA systems, like Passfaces, that use faces as stimuli, taking advantage of humans’ innate ability to effectively recognise large numbers of known faces [7,35]. For a comprehensive review of GA systems see [5].

In user testing, GA systems have been found to be more memorable than password or PIN based systems in a variety of different studies (e.g. [13,15]). However, usability testing in this field suffers from a number of methodological weaknesses: Firstly, only a few of the published studies have presented a rigorous challenge to the user, for example by including the requirement to learn multiple GA codes and reproduce these codes over longer time periods (e.g. [8,19,27]). Secondly, user evaluations in this field have relied rather heavily on university students and so often fail to reflect the needs of older users. Finally, tasks are relatively simple and so ceiling effects are common and these can mask differences between systems.

Those studies that have explored the cognitive demands of using more than one graphical system typically show that the advantages of using GA codes may be short-lived as performance dropped off over time, although it is worth noting that users’ performance with graphical systems in several published studies was at worst equal to their performance with text passwords [8,19,27]. In the study by Chiasson et al., 65 participants were required to remember either 6 textual or 6 graphical user-generated click-based code conditions [8]. During the first trial, participants in the graphical condition were more likely to remember their code than those in the password condition, following only a short 30 second delay (99% recall versus 88%). About two weeks later, 70% could recall their textual password and only 57% could recall their click-based code but this difference was not statistically significant.

Moncur and LePlatre looked at the memorability of 5 system-generated codes in a study in which 172 university participants were assigned to one of 5 conditions: 4 digit PIN code; graphical code (photograph of an object); graphical code + colour; graphical code + mnemonic; graphical code + colour + mnemonic [27]. Participants were required to return twice, each time after a 2-week delay. The researchers had an extreme problem with drop out, with 64% of participants not returning and they suggested that drop out for the PIN condition was greatest because it was the most difficult task. The initial success rate for the 5 PIN condition was approximately 10% while it was over 50% for graphical codes. However, after the first two-week period all success rates dropped to approximately 10%.

Face-based graphical authentication

In recent years, a number of face-based authentication systems have been proposed, the most widely reported being Passfaces a commercial systems developed by RealUser. Brostoff and Sasse conducted an evaluation of the Passfaces system in a field study that found that participants using Passfaces experienced approximately 95% login success rate compared to 85% for alphanumeric password users, despite accessing the system less frequently [6]. Other evaluations conducted by Valentine discovered that users had low expectations that they would remember Passfaces codes over a long period of time [39,40], citing, amongst other things, the difficulties they would experience in trying to write down their codes. However, Valentine found that users’ short-term memory (99.98% accuracy) and long-term memory (84% accuracy) for Passfaces was significantly better than for conventional passwords. It should be noted that in the three studies reported above users were required to only remember one Passfaces code. Interestingly users were allowed to personalise their code, by selecting the faces presented on the subsequent system, Such personalisation can lead to security breaches, as reported by Davis et al. [12].

These face-based systems offer an interesting domain for authentication with older adults because they do well in a recognition task provided that faces are drawn from an appropriate age group (e.g. [2,3,21,35,37]). It makes sense to assume that a face-based GA system may work well with older adults, provided that they are presented with age appropriate faces.

In this study we explore for the first time the performance of older and younger adults on a GA system that uses different types of face stimuli (young and old faces), in a new evaluation paradigm in which multiple authentication codes are learned over several weeks. We would predict that older adults are likely to perform less well than younger adults at memory-based authentication but that this deficit could be overcome with the use of age appropriate faces. Further, we also seek to understand baseline performance for this paradigm and for these user groups (both older and younger adults) with a PIN-based
STUDY 1: PIN AUTHENTICATION

Experimental Design
Participants were recruited into two groups of young adults and two groups of older adults. Participants in each age group were randomly assigned to a high or low cognitive load condition and were tested over a three-week period. In the low load condition participants acquired a set of two PINs in week 1 (labelled SET 1 for analysis purposes) which were tested during weeks 1, 2 and 3 and they then acquired two more PINs in week 2 (labelled SET 2) which were tested during weeks 2 and 3. In the high load condition three PINs were allocated in week 1 (SET 1) and three more in week 2 (SET 2). The dependent measure was the number of successful authentications (maximum of five per account) made in each session.

Analysis of SET 1 PINs took the form of a 2 (old vs. young) x 2 (high load vs. low load) x 3 (time of testing: week 1 vs. 2 vs. 3) ANOVA. Analysis of SET 2 PINs was via a 2 (old vs. young) x 2 (high load vs. low load) x 2 (time of testing: week 2 vs. 3) ANOVA. For this and for the subsequent faces study, time taken to complete the attempts was also measured, however there was never any sign of a speed accuracy trade-off and so the accuracy data alone is presented here. Note that design decisions about loading and delay were made following a pilot trial to ensure floor and ceiling effects would be avoided and that the final design was subject to ethical approval by Northumbria University Life Sciences Research Ethics Committee.

Participants
Thirty-six adults were recruited from the area of Newcastle upon Tyne, UK. Participants were either younger adults aged 18-30 (n = 18, 12 female; mean = 20 years; SD = 3.07 years) or older adults aged 65-75 (n = 18, 14 female; mean = 70 years; SD = 3.83 years). Younger participants were recruited using the university’s online recruitment system where the study was publicised and received course credit. Older participants were recruited using our own participant database and various regional charities and were financially compensated for travel expenses to and from the university. Participants were not screened for gender, cognitive ability or educational attainment. However, research shows that neither education nor gender have an effect on memory for numbers [41], and that education does not predict difficulty in remembering PINs [36]. Screening for cognitive ability was ruled out because the known cognitive decline in older adults [11] would lead to a skewed matched selection of younger adults.

Materials
A simple computer-based PIN system was designed that consisted of an initial image depicting the account followed by text instructing the participant to enter their PIN. The system would display on the screen whether the code was correct or incorrect after entering the four-digit PIN code on the number pad. No feedback was given as to individual digits selected, only on the code as a whole. The system recorded all the digits inputted as well as the time taken to input each digit. In this paper we will only report the accuracy data. Participants were given a randomly generated PIN code along with an account name for each of four (low load) or six (high load) accounts. The six accounts used were alarm (1929), credit card (8360), debit card (2040), library (3126), telephone (9984) and television (5088).

![Figure 1: Overview of procedure over the three weeks](image)

Procedure
The study had three phases: enrolment, distraction and authentication. During the enrolment phase, participants learned each PIN code and associated it with a particular account. They were then given the opportunity to practice by entering the code for a particular account 5 times. If they entered it incorrectly 3 times they were shown the code again. Then followed a distraction phase: a ten-minute discussion with the investigator, focussing on their experiences of current authentication systems. Authentication then followed. Participants were required to enter their PIN for each account five times – a realistic number of attempts before dictionary attacks become a serious threat in a real-world scenario. After each attempt...
they were told if they had entered the PIN correctly, but at no time were then reminded of the correct code.

During week 1, participants were randomly assigned to the high or low load condition and were randomly assigned half of their codes (SET 1) passing through the enrolment, distractor and authentication phase. During week 2 participants learned SET 2 codes, with distractor following and were then asked to authenticate on both SET 1 and SET 2. During week 3 participants were asked to authenticate again on SET 1 and SET 2 (order counterbalanced). See Figure 1 for an overview. Note that the one-week delay allows a measurement of long-term retention before a significant decrement in memory affects older adults [33] and also reflects a common interval used in authentication studies (e.g. [13,15]).

Results and Discussion

The number of successful attempts to recall the codes was analysed using two 3-way Analyses of Variances (ANOVAs), with the three variables remaining stable: age, cognitive load, and time. One ANOVA was conducted for SET 1 data (with authentication data from weeks 1, 2 and 3) and a separate ANOVA conducted for SET 2 data (with authentication data from weeks 2 and 3).

For SET 1, a main effect of age was found (F(1,32)=4.759, p<.05), with younger participants (mean: 3.96) performing significantly better than the older participants (mean: 2.61). No further main effects or interactions were found.

For SET 2, a main effect of age was also found (F(1,32)=6.150, p<.05) where younger participants (mean: 3.88) performed significantly better than the older participants (mean: 2.51). There was also a main effect of time (F(1,32)=5.681, p<.05) where participants performed significantly better in week 2 (mean: 3.48) than week 3 (mean: 2.91). No interaction effects were found.

Taken together, then, these data provide empirical evidence for the commonly held belief that multiple PINs provide a memory challenge and that this challenge is particularly onerous for older adults. Based on the mean successful attempts, older adults risk losing access to their accounts when traditional ‘three-strikes’ systems are in place. Note, too, the poorest authentication performance took place when older adults were trying to remember those PINs acquired relatively late in the process (i.e. SET 2 PINs) following a week’s delay between learning and test.

STUDY 2: A COMPARISON OF FACE-BASED SYSTEMS

Experimental Design

As before, participants were recruited into two groups of young adults and two groups of older adults. Participants in each age group were randomly assigned to a young face or old face condition and were again tested over a three week period, learning three system-generated account codes during week 1 (SET 1) and week 2 (SET 2). Note that this was then the equivalent of the ‘high load’ condition in the earlier study. The dependent measure was the number of successful authentications (maximum of five per account) made in each session. Analysis of SET 1 took the form of a 2 (old vs. young) x 2 (old face vs. young face) x 3 (time of testing: week 1 vs. 2 vs. 3) ANOVA. Analysis of SET 2 was via a 2 (old vs. young) x 2 (old face vs. young face) x 2 (time of testing: week 2 vs. 3) ANOVA.

Participants

Participants were screened for adequate vision and prosopagnosia (a face recognition deficit). Seventy-two participants were recruited in total. 36 younger participants (12 female) were recruited from the student population in the University. The mean age for those allocated to the younger faces group was 19 (SD: 1.61) and for those allocated to older faces, the mean age was 20 (SD: 0.86). Younger participants were recruited using an online participation pool maintained by the university and were given participation credits for taking part.

36 older participants took part (12 female). The mean age of those allocated to the younger faces group was 70 (SD: 3.60) and for the older faces group the mean age was 71 (SD: 3.68). These participants were recruited using the lab’s participant database and through regional advertising. All older participants were financially compensated for travel expenses to and from the university.

Materials

As in Study 1, six fictional ‘accounts’ were created and each was paired with four-item authentication codes of either four younger or four older faces. Four-item codes are common practice in GA systems and are thought to be comparable to 4-digit PINs in terms of theoretical security. A prototype face-based graphical authentication system was built to evaluate the two experimental configurations. Initially the system displayed instructions for participants, followed by four challenge grids that were displayed for each code containing an image from the participants code and 8 distractor images (see examples in Figures 2 and 3). After all four selections were made, the system informed participants whether the code was correct or not.

Grid Composition

As faces were being gathered from different sources, all faces were converted to grey scale to avoid any confound effects: all faces should carry the same chance of being selected and no face should stand out due to other selection criteria (e.g. hair colour). Face recognition literature typically uses black & white stimuli and there is no evidence to suggest an associated recognition deficit.

Old faces were obtained from a number of online databases as no one source could supply a sufficient
number of suitable images. The final old face pool consisted of 80 male faces from the Max Planck Institute [18], University of Aberdeen, and Utrecht University. Note that no young faces were used from the Max Planck Institute.

Young faces were obtained with permission from a pool held by Newcastle University. The face pool consisted of 280 young males from which 80 were chosen at random.

From each pool of 80 faces, 24 unique target young and 24 unique target old faces were selected randomly from the bank using a random number generator. There were no faces that appeared in more than one code to minimise potential code-binding problems [29]. Each of the target faces was assigned 12 decoy faces that would appear on the challenge grids. The decoys were initially assigned randomly from the remaining faces but then examined to ensure similarity of the decoys with the target. Some decoy images were present in more than one grid across codes. The grids were constructed to be visually similar to protect against possible observation attacks (see Figures 2 and 3). Previous research by Dunphy et al. demonstrated that participants were able to guess a face code based on a description more easily if the grids were constructed randomly than if they were constructed to be visually or verbally similar [17].

**Procedure**

The procedure for this study matched that of the ‘high-load’ condition in the previous PIN study. The one difference was the addition of a familiarisation phase during the enrolment stage This was thought to be a necessary difference as enforcing a strategy for learning numbers can be detrimental [14] but research has shown that older adults benefit from strategies when learning other items [28]. During the familiarisation phase all participants were given the account name and the 4 target faces that formed the code for that account. They were shown each face individually and asked to engage in an elaboration process that encouraged the user to think more about the face i.e. to think of answers for seven questions about this face: What is this person’s name? How old is [name]? Does [name] remind you of anyone you know? Does [name] look like a friendly person? What is [name]’s occupation? Where does [name] live? What can you tell me about [name]’s family?

**Results and Discussion**

A 3-way ANOVA with repeated measures on one factor (time of testing) and the independent factors of participant age and face age was carried out on both SET 1 and SET 2 with the dependent variable being the number of successful authentication attempts (max = 5).

For the first set of faces learned (SET 1), a main effect of participant age was found (F(1,68)=17.154, p<.001) with younger participants (mean: 4.84) achieving more successful attempts than older participants (mean: 3.91). A main effect of time was also present (F(2,67)=8.059, p=.001) with pairwise comparisons showing a significant difference in accuracy between week 1 (mean: 4.60) and week 2 (mean: 4.30), and week 1 and week 3 (mean: 4.16), but no difference in accuracy between week 2 and week 3 (see Figure 4). There was no main effect of face-age.

A significant interaction effect was found between participant age and time (F(2, 67)=8.783, p<.001) where no significant difference was found for the younger group in accuracy between weeks 1 (mean: 4.82), 2 (mean: 4.84), and 3 (mean: 4.84) (F(2,107)=0.031, p>.05) while a significant difference was found for the older group in accuracy where older participants were significantly more accurate in week 1 (mean: 4.40) when compared with week 3 (mean: 3.50) (see Figure 4). No further interactions were found.

For the second set of codes, a main effect of participant age was found (F(1,68)=1092.447, p<.001) with younger participants achieving more successful attempts (mean: 4.71) than older participants (mean: 3.60). A main effect of time was also present (F(1,68)=14.216, p<.001) with
participants achieving more successful attempts in the second week (4.38) compared to the third week (3.92).

Figure 4: Interaction between Week and Participant Age for the first set of faces learned (SET 1).

An interaction effect was found for SET 2 between participant age and face age (F(1,68)=12.642, p=.021) where the younger group showed no significant difference in accuracy between the young faces and the old faces (t(70)=1.749, p>.05) while the older group were significantly more accurate with the old faces (mean: 4.05) than the young faces (mean; 3.14; t(70)=2.455, p<.05) (see Figure 5).

Figure 5: Interaction between Participant Age and Face Age for the second set of faces (SET 2).

There was a 3-way interaction between participant age, face age and time (F(1,68)=4.146, p=.046; see Figure 6) where the same-age face advantage was seen at its strongest in week 3. Independent samples t-tests showed that young adults experienced no face-age effect – i.e. by week 3, there was no significant difference between their recall of old vs. young faces (t(34)=1.484, p>.05). The same could not be said of older adults – who showed a clear recognition advantage for older faces by week 3 (t(34)=2.237, p<.05).

This is an interesting finding for those involved in the design of face-based GA systems, as it suggests that the use of older faces would produce significant benefits for an older population without creating undue problems for younger users, although we should be cautious here: it is possible that our results are demonstrating a ceiling effect for younger users and there may still be some mileage in considering age-appropriate faces as optimal stimuli. We discuss this later, following a comparison, below of relevant PIN and face data from this study.

Figure 6: Three-way interaction between Participant Age, Face Age and Time of testing (SET 2).

STUDY 1 AND STUDY 2 COMPARED

In Study 1 we presented an empirical comparison of younger adults and older adults tasked with recalling multiple PIN codes over the course of three weeks. The results demonstrated that older adults were disadvantaged when compared with younger adults and their overall accuracy was below an acceptable level. These results did not come as a complete surprise due to previous self-report evidence suggesting older adults struggled remembering their PIN codes (e.g. [36]), but nonetheless is the first empirical study to demonstrate the extent of the problem. In study 2, we presented two face-based systems which compared the performance of the younger and older adults on authentication systems that comprised young faces or old faces.

Our findings suggest that a face-based GA system can be useful for older adults provided age appropriate faces are used as stimuli. However it would be useful to understand the relative performance of our two age groups on PIN vs. face systems and so we conducted a secondary analysis, comparing PINs with the most inclusive of the two face systems we tested – i.e. that based on older faces.

PIN vs. Face Set 1

Figure 6: Accuracy of younger and older participants with PIN and Faces (SET 1).

A 3-way mixed ANOVA was carried out on a subset of the data from Studies 1 and 2. Firstly we used the high-load condition from Study 1, as participants in that condition were asked to learn six account codes, the same as in Study 2. Then we selected the old face condition as
a point of comparison as this design was shown to be most effective for older adults. Both studies used a common research methodology, however we acknowledge that secondary analyses across separately conducted studies can result in increased type 1 error and so we have utilised a stricter Alpha level of 0.01 in our analysis.

For SET 1, we found a main effect of age (F(1,50)=5.518, p<.001) where the younger group (mean: 4.13) was significantly more accurate than the older group (mean: 3.26). A main effect of system was also found (F(1,50)=17.625, p<.001) where participants were more accurate with Faces (mean: 4.47) than with PIN (mean: 2.91). No interaction effects were found.

For SET 2, we found a main effect of age (F(1,50)=19.535, p<.001) where the younger group (mean: 4.42) was significantly more accurate than the older group (mean: 3.00). A main effect of system was also found (F(1,50)=4.31) than with PIN (mean: 3.12). A main effect of time was also present (F(1,50)=15.271, p<.001) showing that participants were significantly less accurate during week 3 (mean: 3.42) than during week 2 (mean: 4.01).

An interaction was found between age and system (F(1,50)=7.721, p>0.1) where the difference in accuracy between the two systems was not significant for the younger group (PIN: 4.28, Faces: 4.57) (t(52)=1.027, p>0.1) while the older group were significantly less accurate with PIN (mean: 1.96) when compared with Faces (mean: 4.05) (t(52)=4.987, p<.001). In Figures 6 and 7 we plot these results, allowing a visual comparison between performance on the PIN vs. older-face system with younger and older adults.

**PIN vs. Face Set 2**

Figure 7: Accuracy of younger and older participants with PIN and Faces (SET 2).

If we look firstly at SET1 (Figure 6) performance, the older adult disadvantage shows up clearly, as does the performance advantage for the face-based GA system – present for both younger and older adults. If we look at SET 2 (Figure 7) data we can see that the initial PIN problems experienced by the older adults have been exacerbated with the addition of a second set of PIN codes, such that their accuracy levels by week 3 are very poor. In contrast, older adults performance with the second set of face codes seems more robust.

**GENERAL DISCUSSION**

Age is an important issue for designers of authentication systems and in this paper we have provided data on the ways in which specific authentication stimuli can differentially affect performance in younger and older adults. This is a key issue, as the proportion of older adults in the general population is rising and older adults have become the fastest-growing group of Internet users [22]. It is therefore important to consider this population in the design of future, more inclusive authentication systems. We recognise that the total elimination of an age-related performance decrement for authentication will be difficult, given the well-known decline of long-term memory with age [11], but our intention is to explore ways in which we can ameliorate this problem.

Specifically, in this paper we have shown that a face-based GA system using old faces was associated with a significant improvement in older adults’ authentication performance. Further, we showed that we could achieve this improvement without unduly penalising performance in the younger adults. Our first conclusion, then, is that the existing Passfaces’ database of young faces does not provide an optimal, fully inclusive authentication solution.

In exploring the design space for face-based authentication, we are studying only the tip of the iceberg. In this study we took advantage of a known same-age effect in face recognition [37]. However, it should be noted that the impact of the same-age effect on authentication may be greater than that reported here as the majority of our participants were females and yet the faces in our system were males. Own-gender effects also exist for females where they are better able to recognise female faces over male faces – an effect not present for males [25] and so it is arguable that the best form of inclusive design would either be based on female faces, or would include a gender-matched set of stimuli. This opens up the question of how realistic it may be to employ personalised design in face-based GA involving authentication stimuli that are matched on gender, age or ethnicity.

**Personalised Authentication**

Let us consider an authentication design solution where face-based stimuli are tailored to the specific characteristics of the user. This would arguably be the most appropriate solution if our sole concern is to improve performance across the largest possible user base but we need to consider whether it is also a more secure solution and also consider whether it is realistic as a design aim.
The characteristics of a personalised face-based system would rely on three key user characteristics: the user’s age, the user’s ethnicity, and the user’s gender. In this paper we have focused on an own-age effect that seems to be effective in an authentication context and that is backed up by a vast psychology literature [37]. This literature is also in agreement that people are better able to recognise faces from their own ethnic origin than from other ethnicities [26] and that females are better able to recognise female faces than male faces, suggesting a limited same-gender effect [25]. Note, however, the absence of a recognised same-gender similar effect for males, suggesting that only females would benefit from gender-specific face assignment.

Security
How would such personalised systems impact on the security of an authentication system? By assigning users faces based on personal characteristics that they are particularly sensitive to, there is an argument that the system would be more, rather than less secure, as discrepancies between the face images would be notable and highly salient to that class of user, but not particularly salient to other potential attackers. This would be akin to assigning a mechanic an image of a car engine and having that user select the correct engine from a grid consisting of nine car engines. While the mechanic will be able to differentiate between the nine engines, an untrained user (e.g. a doctor) will not necessarily notice the subtle differences [9]. From a security standpoint, then, a personalised system may disadvantage untrained attackers who would be less likely to pick up on key cues in the stimulus array. Such a personalised system would have the advantage of maximising image discriminability for the individual while maximising perceived image similarity for the general population. Certainly there is an important security issue in ensuring that images in a GA grid are sufficiently similar – Dunphy et al. have shown that visually similar grids are much less vulnerable to description attacks [17].

Are such personalised systems realistic as an inclusive design aim? We believe they are provided that the stimuli are system-generated. This is important for two reasons: firstly it creates a much more tractable design problem in which the system needs to know relatively little about the user in order to generate an appropriate set of GA codes. Secondly, system-generated grids are actually more secure. The literature has many examples of the problems that can follow user-selection of GA codes. For example, example, 90% of black males would choose a black face as part of their security code, which is obviously a problem if the array offers mixed ethnicity. Further, the same author found that the worst 10% of the codes generated by men were available to potential attackers in only two guesses [12]. Users’ preferences have been shown to be affected by gender, race and attractiveness, but all of these could be controlled in a system-generated array.

Note that a similar user-generated effect has also been found in click-based graphical authentication schemes. PassPoints has vulnerabilities that arise from structural aspects of visual perception – i.e. users may be strongly attracted to certain “hotspots” within an image which are predictable. Thus the entire system is rendered vulnerable to security attacks [16,31,34]. Security studies are starting to use volunteers via Mechanical Turk in their studies in order to learn more about these common behaviours [38] and one could imagine a scenario where such volunteers inadvertently help criminals by identifying hotspots, faces or images most likely to be selected as user codes.

It is possible, of course, that user privacy concerns may be greater for personalised systems, where the authentication codes themselves could reflect sensitive race, gender or age information about the user. However, we should note that the use of authentication codes in public spaces (such as with an ATM) can lead to observation attacks where both codes and user attributes can be acquired directly.

While user selection may be preferable for memory, it does not solve memory problems such as associating the code with the item to be accessed and may be too big a security cost. For security, system generated codes are required to assure the optimum entropy of a system and we must understand how to maximise the memorability of such codes. With the personalisation of codes, it may be possible to improve security without sacrificing usability.

The evaluation paradigm
In this study a new evaluation paradigm was developed that allowed us to explore the effects of age (young vs. old), system (old faces vs. young faces), time (week 1 vs. week 2) and cognitive load (high vs. low) for more than one set of learned codes and with testing taking place over several weeks. This paradigm can be applied to other usable security studies with the purpose of evaluating the effect of multiple codes and we would recommend it, as some of the strongest effects that we observed were absent with the first set of authentication codes learned, but only became manifest once users had to learn a second set – an effect that can be attributed to proactive interference [4] where early learning can have a negative impact upon the subsequent learning of similar material.

Looking at the history of testing new systems, at first new authentication systems were tested only using a single code [6,13,15,39,40]. These studies suggested extraordinary performance gains over existing systems, yet they did not address the major problem affecting current systems – multiple codes. More recently, multiple systems have been tested to provide a more realistic measure of performance [8,19,27]. However, very different methodologies have been used to test both single and multiple codes, making it difficult to compare results across studies. Additionally, many of those recent studies that have employed multiple codes have lacked ecological validity in the way the codes have been assigned.
Participants have been given or have chosen all codes during the same session, while we have split the acquisition of codes to multiple sessions, representing the progressive real-world way users gradually acquire codes to their different accounts.

As we have noted, one advantage of using the paradigm presented here lies in the separation of different code sets in analysis, where it yields short-term and long-term performance data for both ‘old’ and ‘new’ codes. This is useful for understanding where the performance decrement is most prominent. Hence, rather than reporting an age effect with a system, we are able to determine where exactly the problem is occurring and relate this more closely to our understanding of the cognitive attributes of a particular user population.

**Design Guidelines and Conclusions**

In addition to informing more realistic and comparative evaluations of graphical authentication codes, our results have direct implications for the design of graphical authentication systems. To maximise usability of graphical authentications systems and to achieve a balanced approach to security the following guidelines should be considered:

- Use independent non-overlapping codes to remove the association deficit problem (i.e. [29]), although it may only be possible if a central organisation distributes the images;
- Personalise images to maximise memorability (e.g. age, gender and race appropriate faces) as suggested by psychology literature (e.g. [37]) and demonstrated here with age-appropriate faces;
- Use elaboration techniques when learning the graphical image (i.e. ask questions about the image) similar to [6,39];
- Use system generated images to maximise entropy and reduce guessing as previously suggested by [12];
- Employ similar images to reduce guessing and giving away (i.e. [17]).

In this paper we have presented a face-based GA system that gives superior performance for both younger and older adults when compared to a PIN system when the requirement to learn multiple codes was tested over a three-week period. Further, we have shown an effect of the face stimuli themselves, such that older face stimuli would be the better selection for a more inclusive design as they were associated with much more successful authentication performance in older adults. We argue that age-appropriate faces could be a useful development in this design space, but suggest other ways in which personalised design could be used to maximise performance across the entire user population.

**REFERENCES**

