Investigating User Authentication in the Context of Older Adults

James Nicholson

PhD

2012

Investigating User Authentication in the Context of Older Adults

James Nicholson

A thesis submitted in partial fulfilment

of the requirements of the

University of Northumbria at Newcastle

for the degree of

Doctor of Philosophy

Research undertaken in the

School of Life Sciences

November 2012

Abstract

Knowledge-based authentication is almost ubiquitous due to low cost and relatively straightforward implementation. Despite its popularity, there are some well-known problems associated with knowledge-based authentication, such as the cognitive load of memorising multiple codes. As people age and their memory declines, remembering multiple codes is even more challenging.

Due to lack of objective evidence regarding the performance of older adults with existing knowledge-based systems, a study was carried out where younger and older participants were required to learn and remember multiple PIN codes and their performance was evaluated over a three-week period. The results from this PIN study demonstrated a clear age effect where younger participants performed significantly more accurately and faster than the older participants. These results reiterated the need for authentication systems that are inclusive of older users and provided a benchmark performance measure for future evaluations.

In the next phase four graphical authentication systems (GAS) were evaluated with younger and older adults using the same methodology as the PIN study to determine whether any of them were an improvement. The first system, Tiles, was based on a single image and participants were required to recognise segments of their image from segments taken from other images and yielded disappointing results where overall performance was not an improvement over that of PINs. The second and third systems tested were picture-based and face-based recognition systems. The performance of older participants was promising, especially with the face-based system but the systems could be improved to be more suitable for older users.

In the final study, the face-based system was improved by using old faces and ensuring that no two codes shared a face. The results from the final face-based system provide preliminary evidence that a graphical authentication system that is inclusive of older adults may be achievable if designed correctly.

Table of Contents

1. INTRODUCTION TO THE THESIS 13

2. COMPUTER AUTHENTICATION LITERATURE 17

2.1. PURE RECALL 18

2.2. CUED-RECALL 23

2.3. GRAPHICAL AUTHENTICATION SYSTEMS 25

2.3.1. Why do they work? 28

2.3.1.1. Recognition Over Recall 29

2.3.1.2. Visual Superiority 30

2.3.1.3. Face Recognition 31

2.3.2. Graphical Authentication: The Story So Far 32

2.3.3. Testing Paradigms 37

2.4. SUMMARY 41

3. AGE LITERATURE 42

3.1. AGE-ASSOCIATED MEMORY DECLINES RELEVANT TO AUTHENTICATION 42

3.2. AUTHENTICATION AND OLDER ADULTS 46

3.2.1. Graphical Authentication Systems 47

3.2.1.1. Why use GAS with older adults? 47

3.2.1.2. History of GAS with an older adult user base 48

3.2.1.3. Older Adults and Authentication 49

3.2.1.4. Looking Forward 50

4. BENCHMARKING EXISTING AUTHENTICATION: PINS 51

4.1. RATIONALE 51

4.2. METHOD 53

4.2.1. Design 53

4.2.2. Participants 53

4.2.3. Materials 54

4.2.4. Procedure 55

4.2.4.1. Enrolment Stage 55

4.2.4.2. Authentication Stage 56

4.2.4.3. Procedure for Participants 56

4.3. RESULTS 58

4.3.1. Successful Attempts – Accuracy 58

4.3.1.1. Set 1 – Original Codes 58

4.3.1.2. Set 2 – New Codes 59

4.3.1.3. Overall Accuracy 60

4.3.2. Average Time - Speed 61

4.3.2.1. Set 1 – Original Codes 61

4.3.2.2. Set 2 – New Codes ... 62
4.3.3. Order of Acquisition ... 63
4.3.3.1. Low Load ... 63
4.3.3.2. High Load ... 64
4.4. DISCUSSION ... 65
4.4.1. Implications ... 67
4.4.2. Future Work ... 68
4.5. CHAPTER SUMMARY ... 69
5. A NEW APPROACH TO AUTHENTICATION WITH PICTURES: TILES GRAPHICAL AUTHENTICATION SYSTEM ... 70
5.1. RATIONALE ... 70
5.2. METHOD ... 73
5.2.1. Experimental Design ... 73
5.2.2. Participants ... 73
5.2.3. Materials ... 74
5.2.3.1. Grid Composition ... 75
5.2.3.2. Image Overlap ... 77
5.2.4. Procedure ... 77
5.2.4.1. Enrolment Stage ... 77
5.2.4.2. Authentication Stage ... 78
5.2.4.3. Procedure for Participants ... 79
5.3. RESULTS ... 80
5.3.1. Successful Attempts – Accuracy ... 81
5.3.1.1. Set 1 – Original Codes ... 81
5.3.1.2. Set 2 – New Codes ... 81
5.3.1.3. Overall Accuracy ... 83
5.3.2. Average Time ... 84
5.3.2.1. Set 1 – Original Codes ... 84
5.3.2.2. Set 2 – New Codes ... 85
5.3.3. Order of Acquisition Effects ... 86
5.4. DISCUSSION ... 86
5.4.1.1. Speed ... 88
5.4.2. Implications ... 89
5.4.3. Design Implications ... 90
5.5. CHAPTER SUMMARY ... 91
6. GRAPHICAL AUTHENTICATION SYSTEMS RE-ENGINEERED: FACES AND PICTURES ... 92
6.1. RATIONALE ... 92
6.2. METHODOLOGY ... 93

6.2.1. Experimental Design ... *93*
6.2.2. Participants ... *94*
6.2.3. Materials ... *94*
6.2.3.1. Grid Composition for Faces ... 96
6.2.3.2. Grid Composition for Pictures ... 98
6.2.3.3. Image Overlap ... 99
6.2.4. Procedure ... *100*
6.2.4.1. Enrolment Stage & Familiarisation ... 100
6.2.4.2. Authentication stage ... 102
6.2.4.3. Procedure for Participants ... 102
6.3. RESULTS ... 104
6.3.1. Successful Attempts – Accuracy ... *104*
6.3.1.1. Set 1 – Original Codes ... 104
6.3.1.2. Set 2 – New Codes ... 106
6.3.1.3. Overall Accuracy ... 107
6.3.2. Average Time - Speed ... *108*
6.3.2.1. Set 1 – Original Codes ... 108
6.3.2.2. Set 2 – New Codes ... 111
6.3.1. Additional Analysis: Image Interference ... *113*
6.3.1.1. Set 1 – Original Codes ... 113
6.3.1.2. Set 2 – New Codes ... 114
6.3.1.3. Overall Image Interference ... 114
6.3.2. Order of Acquisition Effects ... *115*
6.4. DISCUSSION ... 116
6.4.1. Implications & Improvements ... *118*
6.4.2. Conclusions ... *120*
6.5. CHAPTER SUMMARY ... 120
7. IMPROVING FACE-BASED AUTHENTICATION FOR OLDER ADULTS: YOUNGFACES AND OLDFACES ... 122
7.1. RATIONALE ... 122
7.2. METHOD ... 124
7.2.1. Experimental Design ... *124*
7.2.2. Participants ... *124*
7.2.3. Materials ... *125*
7.2.3.1. Grid Composition for Young Faces ... 126
7.2.3.2. Grid Composition for Old Faces ... 127
7.2.3.3. Image Overlap ... 128
7.2.4. Procedure ... *129*
7.2.4.1. Enrolment Stage + Familiarisation ... 129
7.2.4.2. Authentication Stage ... 130
7.2.4.3. Procedure for Participants ... 130

7.3. RESULTS 132
7.3.1. Successful Attempts- Accuracy *133*
7.3.1.1. Set 1 – Original Codes 133
7.3.1.2. Set 2 – New Codes 134
7.3.1.3. Overview of Accuracy 136
7.3.2. Average Time - Speed *137*
7.3.2.1. Set 1 – Original Codes 137
7.3.2.2. Set 2 – New Codes 138
7.3.3. Order of Acquisition Effects *139*
7.4. DISCUSSION 140
7.4.1. Implications *142*
7.4.2. Future Improvements *143*
7.5. CHAPTER SUMMARY 145
8. FINAL DISCUSSION 146
8.1. INTRODUCTION 146
8.2. RESEARCH QUESTIONS 146
8.2.1. Performance with Current Authentication Systems *146*
8.2.2. Performance with Graphical Authentication Systems *147*
8.3. SYSTEM COMPARISONS 149
8.3.1. Loads of 4 Codes *149*
8.3.1.1. PIN vs. Tiles 150
8.3.1.2. PIN vs. Traditional GAS 150
8.3.2. Loads of 6 Codes *151*
8.3.2.1. PIN vs. YoungFaces 151
8.3.2.2. PIN vs. OldFaces 152
8.3.2.3. Further Analysis of PIN vs. OldFaces 153
8.3.3. Observed Trends *155*
8.4. CONTRIBUTIONS 156
8.5. LIMITATIONS 157
8.6. NEXT STEPS 159
8.7. FINAL CONCLUSIONS 161

List of Figures

Figure 2.1: The Passfaces graphical authentication system (Rose, n.d.). 26
Figure 2.2: The Déjà vu graphical authentication system (Dhamija & Perrig, 2000). 26
Figure 2.3:The VIP 1/2 graphical authentication system (De Angeli et al., 2005). 27
Figure 2.4: The PassPoints graphical authentication system (Wiedenbeck et al., 2005). 28
Figure 4.1: Simplified instructions for participants along with image of account. 55
Figure 4.2: Overview of Average Successful Attempts in SET 1. 59
Figure 4.3: Overview of Average Successful Attempts in SET 2. 60
Figure 4.4: Overview of accuracy for all PIN codes. 61
Figure 4.5: Overview of average time taken to enter PINs in SET 1. 62
Figure 4.6: Overview of average time taken to enter PINs in SET 2. 63
Figure 4.7: Frequency of forgetting PINs for younger and older participants – Low Load condition. 63
Figure 4.8: Frequency of forgetting for PINs for younger and older participants – High Load condition. 64
Figure 5.1: Demonstration of the Tiles system. Participants have to select their segment amongst foils in 4 individual challenges. 74
Figure 5.2: Example Tiles grid. 76
Figure 5.3: Three-way interaction between Age, Grid Type and Week for SET 2. 82
Figure 5.4: Overview of successful attempts for images in both SET 1 and SET 2. 83
Figure 5.5: Interaction between Age and Week for SET 1. 85
Figure 5.6: Frequency of forgetting for Tiles codes for both younger and older participants. 86
Figure 6.1: Example grids for Faces (left) and Pictures (right). 96
Figure 6.2: Target faces used for ATM account. 97
Figure 6.3: Target faces used for Bank account. 97
Figure 6.4: Target pictures used for Email account. 99
Figure 6.5: Target pictures used for NHS account. 99
Figure 6.6: Illustration of image overlap - each grid contains one target image and one foil image used as a target for another account. 99
Figure 6.7: Interaction between age and system for SET 1, showing the older group being more accurate with Faces than with Pictures, while the younger group were more less accurate with Faces than with Pictures. 105
Figure 6.8: Illustration of the interaction between age and week for SET 2, showing how the accuracy of the older group dropped significantly during the second week of the study. 106
Figure 6.9: Overview of successful attempts for both sets of images. 107
Figure 6.10: Illustration of the interaction between age and system for SET 1, showing how the older group took longer to select their Pictures, while the younger group were the other way around. 109
Figure 6.11: Illustration of the interaction between system and week for SET 1. 110

Figure 6.12: Illustration of the 3-way interaction between age, system and week for SET 1, showing how the younger group took significantly longer to select their faces during week 3 while the older group took significantly less time to select their Faces during week 2. 111
Figure 6.13: Illustration the interaction between age and week for SET 2, showing how the older group took longer to select their images in the second week of the study. 112
Figure 6.14: Illustration the interaction between system and week for SET 2, showing how initially Faces took longer to select but during the second week the time taken to select both images evened out. 113
Figure 6.15: Overview of overall image interference for both sets of images (percentage correct). 115
Figure 6.16: Forgetting of image codes for both younger and older participants. 116
Figure 7.1: Simplified instructions to participants. 126
Figure 7.2: Sample grid for young faces. 127
Figure 7.3: Sample grid for old faces. 128
Figure 7.4: Interaction between participant age and week for SET 1. 134
Figure 7.5: Interaction between Participant Age and Face Age for SET 2. 135
Figure 7.6: Three-way interaction between Participant Age, Face Age and Week for SET 2. 135
Figure 7.7: Overview of average successful attempts for both the younger and older groups using old and young faces. 136
Figure 7.8: Interaction between Participant Age and Week for SET 1. 138
Figure 7.9: Forgetting of YoungFaces codes for both younger and older participants (NOTE: no codes forgotten by the younger group). 139
Figure 7.10: Forgetting of OldFaces codes for both younger and older participants. 140
Figure 8.1: Comparison for PIN (left) and Tiles (right). 150
Figure 8.2: Comparison of PIN (left) and traditional GAS (right). 151
Figure 8.3: Comparison of PIN (left) and YoungFaees (right). 152
Figure 8.4: Comparison of PIN (left) and OldFaces (right). 153

List of Tables

TABLE 2.1: METHODOLOGIES USED FOR EVALUATING GRAPHICAL AUTHENTICATION SYSTEMS. L/F = LAB-BASED OR FIELD-BASED STUDY. INTERVAL = DELAY BETWEEN LEARNING AND RECALLING/RECOGNISING. MEASURE = DEPENDENT VARIABLE. ANALYSIS = METHOD OF ANALYSIS FOR THE DEPENDENT VARIABLE. CODES = NUMBER OF DIFFERENT CODES PARTICIPANTS HAD TO REMEMBER AND WHETHER THEY WERE ASSIGNED OR CHOSEN. 38

TABLE 3.1: AUTHENTICATION SYSTEMS TESTED WITH AN OLDER ADULTS POPULATION. L/F = LAB-BASED OR FIELD-BASED STUDY. INTERVAL = DELAY BETWEEN LEARNING AND RECALLING/RECOGNISING. MEASURE = DEPENDENT VARIABLE. ANALYSIS = METHOD OF ANALYSIS FOR THE DEPENDENT VARIABLE. CODES = NUMBER OF DIFFERENT CODES PARTICIPANTS HAD TO REMEMBER AND WHETHER THEY WERE ASSIGNED OR CHOSEN. 49

TABLE 4.1: ACCOUNTS AND CODES USED THROUGHOUT THE STUDY. 54

TABLE 4.2: OVERVIEW OF PROCEDURE. 57

TABLE 5.1: LIST OF QUESTIONS USED EACH OF THE FOUR BASE IMAGES. 78

TABLE 5.2: OVERVIEW OF PROCEDURE. 79

TABLE 6.1: LIST OF QUESTIONS USED FOR BOTH FACES AND PICTURES DURING THE FAMILIARISATION PROCESS. . 101

TABLE 6.2: OVERVIEW OF PROCEDURE. 103

TABLE 7.1: FAMILIARISATION QUESTIONS USED FOR BOTH YOUNG AND OLD FACES. 130

TABLE 7.2: OVERVIEW OF PROCEDURE FOR PARTICIPANTS OVER THE THREE-WEEK PERIOD. 131

Acknowledgements

I would like to thank my supervisors, Lynne Coventry and Pam Briggs, for all their guidance and support.

I would also like to thank all participants who took part in my studies.

Last, but absolutely not least, I would also like to thank Rebecca for her patience and understanding over the past three years.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the School of Life Sciences' Ethics Committee.

Name: James Nicholson

Signature:

Date: 16/11/2012

1. INTRODUCTION TO THE THESIS

User authentication is an important part of computer systems which allows information to be accessible only to authorised users. Knowledge-based authentication (KBA), based on codes such as passwords or PINs, is ubiquitous due to low cost and relatively straightforward implementation. Despite its popularity, there are some well-known problems associated with KBA such as the memorability of multiple codes. As people age and their memory declines, remembering multiple codes could become even more challenging. In the past decade, graphical authentication systems (GAS) have been proposed as alternatives to alphanumeric passwords in part due to humans' superior ability at recognising images over recalling random combinations of characters. These systems have achieved convincing results thus far, but have not been evaluated with older adults despite them being the fastest-growing group of internet users in developed nations (Hart, Chaparro, & Halcomb, 2008). Additionally, the memory literature suggests that older adults could also take advantage of the picture superiority effect.

Two research questions were devised based on the existing authentication literature with the aim of designing a system that is inclusive of older adults. These questions were explored through a series of studies evaluating various authentication systems using younger and older adults:

1. Are older adults disadvantaged by existing authentication systems?
2. Can graphical authentication systems improve the performance of older adults in relation to existing authentication systems to be a more inclusive form of authentication?

To answer these questions this thesis is structured as follows:

The second chapter explores the literature on existing computer authentication systems and discusses their limitations. The literature on authentication makes it clear that a problem exists with regards to the memorability of passwords and other knowledge-based authentication systems. However, a lack of empirical evidence makes it difficult to determine how big the problem is in reality. The literature on graphical authentication systems suggest that they may be better than passwords and PINs, but the lack of

control over the time delays, number of codes, etc. make it difficult to compare results between studies. To move forward, a consistent method should be applied to different systems.

The third chapter covers age-associated memory declines that are likely to impact the performance of older adults using authentication systems. A few studies that explored authentication with older adults are also covered in this chapter, with the main finding being the high success rate that the older adults experienced. However, these studies do not evaluate the systems with younger adults making it difficult to establish the inclusiveness of the systems.

A methodology was designed that would be consistently applied throughout the thesis that would allow for comparison between systems. This methodology included being able to compare results for memorability of two different sets of codes, and a consistent time frame. Chapters four to seven present the usability studies that were carried out using this methodology to evaluate PIN and five different graphical authentication systems. All studies were ethically approved by the School of Life Sciences' Ethics Committee.

Firstly, due to the relative lack of empirical evidence regarding the performance of older adults with existing KBA, a study was carried out where the younger and older participants were required to learn and remember multiple PIN codes and their performance was evaluated over a three-week period. This was done using a robust methodology that divided the multiple codes into two sets – original codes and new codes. Participants were given the original codes during the first week and the new codes during the second week and were tested on the recognition or recall throughout the three weeks. This study was the first to compare the performance of younger and older adults with multiple codes of an existing authentication system – in this case PIN – over the course of a short and long delay. The results showed a clear age effect where younger participants performed significantly better – in terms of both successful attempts and time taken to enter codes – than the older participants. These results reiterated the need for authentication systems that are inclusive of older users and presented a benchmark methodology performance measure that would be used for future evaluations. This study was the first to experimentally record the memorability of multiple PINs for both younger and older adults, rather than relying on self-reporting.

This study also highlights the problem of multiple PIN recall for older adults which leaves them vulnerable in the real world and possibly unable to gain access to systems.

In the next phase, four GAS were evaluated with younger and older adults to determine whether any of them were an improvement over PIN for older adults. This series of studies were the first to evaluate the performance of younger and older adults with multiple GAS codes, and the first to use a consistent methodology that allowed comparisons. The first system, Tiles, was based on principles of cued-recall and recognition-based GAS. Participants were required to remember one picture per account and recognise segments from that image from amongst segments of other images in order to authenticate. This study was the first to evaluate this novel graphical system with the aim of reducing the age effects observed with PINs. The system yielded disappointing results that indicated the inadequacy of this solution for older adults over an extended period of time. The poor results were attributed to the abstract nature of some of the resulting segments which caused confusion. Part of this study was published in a leading international human-computer interaction conference, CHI 2012 (Nicholson, Dunphy, Coventry, Briggs, & Olivier, 2012). See Appendix D for full paper.

The next study evaluated a picture-based recognition system based on an existing graphical authentication system, VIP (DeAngeli et al., 2002) and a face-based system based on Passfaces. These two systems were chosen for their use of full images with the aim of improving the recognition for both age groups while measures were taken to make the systems realistic in terms of implementation. The nature of the images was changed to improve security, where the user had to choose an image from a group of similar images rather than dissimilar images. This study was the first to compare the performance of younger and older adults with multiple codes of face-based and picture-based graphical systems. In terms of accuracy, older participants were better with Faces and Pictures than with PIN, especially with the face-based system. However, a time-based performance decrement was observed for the older participants meaning that the systems in their form did not take full advantage of the visual properties. With this in mind, improvements were planned.

The final study compared the performance of younger and older adults with two face-based systems, one using young faces and the other using old faces to ascertain whether

the age of the faces used in the code affected memorability. This study was the first to evaluate a face-based system that was designed to take advantage of own-age effects and remove overlap between images used in different codes. The results showed a performance decrement over time but it was not age-specific meaning younger and older participants experienced the same decline rate in performance. This study found that when using older faces the older adult performance in the third week was not significantly different from younger adults for the memorability of 6 face-based codes.

The results of these studies demonstrate that simply using graphical images in codes does not guarantee improvements in memorability and that they must be designed appropriately to create a more inclusive solution. One such improvement is the use of age appropriate faces, the other is ensuring that there is no overlap of images used within different codes.

2. COMPUTER AUTHENTICATION LITERATURE

This chapter focuses on the issues in user authentication with a specific focus on graphical authentication systems. The basic mechanisms and the psychological principles behind the authentication systems will be reviewed before discussing the evaluation techniques that have been used in the domain. The principle aim of this chapter will be to highlight the flaws with system evaluations carried out in the past and to suggest improvements which will go on to be implemented in the subsequent experiments in this thesis.

Authentication is the act of proving to a system that you are who you say you are. In the context of computers, authentication involves a person satisfying the computer system that they are the user they claim to be, either by presenting a shared secret, presenting a physical item, or presenting a unique personal characteristic that matches the one stored for that particular username. Renaud and De Angeli (2004) describe authentication as a three-step process where first the user has to be identified – identification: usually involving a username – then provide evidence to prove the identity – authentication – and finally access rights are granted by the system if successful – authorisation. Authentication is an important aspect of computer security that allows users to keep their information private from other users, and in essence allows a computer to serve more than just one person.

Authentication systems typically fall into one of three categories: token-based authentication, biometric authentication or knowledge-based authentication. Token-based authentication relies on the presence of a physical item to authenticate the user – if an appropriate token is presented then the person holding it is granted access. Biometric authentication relies on the users' unique and stable characteristics in order to authenticate the user. If the presented feature matches the 'template' in the database, then the person is deemed to be the user and they are granted access. Knowledge-based authentication (KBA) systems rely on a string of information that is shared between the system and the user. The user is expected to remember a combination of some sort – usually letters for a password or numbers for a personal identification number (PIN) – and present that combination when they wish to access their account. The information string can be either secret – as is the case with passwords and PINs – or public – as is

often the case with challenge questions. Knowledge-based systems are the most common type of authentication – more specifically passwords (Herley, Oorschot, & Patrick, 2009; Renaud, 2005) – due to their relatively simple and inexpensive implementations in addition to being familiar to users. As a consequence, KBA will be the focus of this review.

Knowledge-based authentication encompasses a wide array of systems, but these typically fall into two categories: pure recall or cued-recall. Pure recall systems require users to enter their string with no context provided. Passwords and PINs are examples of pure recall KBA systems – the user is asked to enter the code usually without any prompts. Cued-recall systems give users some contextual information during the login in order to aid the recall of the string. Challenge questions are an example of cued-recall KBA systems – the user is asked a question (the cue) and is then expected to answer that question. This review will firstly cover text-based pure recall systems (see subsection 2.1) and will then follow with text-based cued-recall systems (see subsection 2.2). Graphical systems will be covered in subsection 2.3.

2.1. PURE RECALL

Passwords are the most popular pure recall knowledge systems in terms of implementation (e.g. Herley et al., 2009). Despite their widespread use, there are a number of problems that are associated with this type of system. One of the main problems lies with the constraints that are placed by providers in order to maintain the security of the systems such as update frequency (e.g. Sasse, Brostoff & Weirich, 2001). There are generally two main schools of thought when it comes to the construction of secure passwords: Complexity and Length.

Researchers who believe in password complexity encourage users to utilise and interweave as many character sets as possible to create complex random codes such as f8pRy@7&. Proctor, Lien, Vu, Schultz, and Salvendy (2002) explored the vulnerability of different types of passwords to being cracked – i.e. guessed – by a computer program and found that enforcing users to use a specified length (five characters or eight characters) plus character restrictions (e.g. uppercase and lowercase letters plus numbers) resulted in passwords that were less vulnerable to attack than passwords that

only enforced length. Additionally, they report that the extra enforcements did not affect the login time of participants after a distractor task. However, the authors fail to report on the accuracy of the passwords inputted after the distractor task. Yan, Blackwell, Anderson, and Grant (2004) carried out a similar study investigating the effect of instruction on the quality of the generated passwords and found that the addition of symbols to the combination of uppercase and lowercase letters significantly improved the resistance of the codes to cracking. This finding was supported by a series of studies by Vu et al. (2007) who also recommend using symbols in addition to letters and numbers to create strong passwords. Yan et al. (2004) mention that passwords composed of six or less characters were easy to crack regardless of the complexity, suggesting that length is very important to security. Just as with Proctor et al. (2002) the authors of both studies do not take into account the usability implications of their complexity recommendations.

Researchers who believe in length encourage users to implement the maximum possible number of characters for codes but they do not necessarily have to be composed of complex combinations. They argue that the added password space (e.g. 13 characters over 8 characters) balances the smaller character set (e.g. not utilising symbols). This notion was summed up by Holt (2011) who concluded that password length was more important than password complexity. Komanduri et al. (2011) carried out a study where participants were given various restrictions on passwords they had to create and remember a week later. One of their most important findings was that longer passwords – at least 13 characters long – composed without any specific requirements had more entropy (i.e. more secure) than more 'traditional' passwords – e.g. requiring the use of uppercase and lowercase letters along with numbers and symbols. Additionally participants perceived the longer passwords as more 'usable' although they were also perceived as less secure than complex ones. Experts in this camp tend to take the usability of the system into consideration, although that is not always the case and the main focus remains on the security of the systems.

Regardless of the approach favoured by experts or providers, the focus of security professionals over time has been to make systems more secure without taking into account the repercussions facing the users. Generally speaking, more secure systems involve making the system more complicated for the end user, or requiring the user to perform more steps than before. This is demonstrated by Weir, Aggarwal, Collins, and

Stern (2010) who suggest different methods for evaluating and rejecting 'weak' passwords despite being aware that such actions would result in user annoyance. Another perfect example lies in the suggestion by a security expert that the length of passwords should be a minimum of nine characters as well as being unique to each site (Goodin, 2012a). This change was suggested with the aim of keeping passwords safe from brute force and dictionary attacks – attacks where multiple combinations are attempted consecutively until one finally guesses the password (e.g. Morris & Thompson, 1979) – yet they are likely to cause more problems for users and will most likely encourage users to find a way around the security system rather than comply with the requirements (Inglesant & Sasse, 2010). This observation is backed by previous research that has shown that computer users wish to utilise the machine for their primary task and do not want to deal with added security settings which are perceived to be a hindrance (Dourish, Grinter, Delgado de la Flor, & Joseph, 2004; Sasse, Brostoff, & Weirich, 2001). Research by Weir, Douglas, Carruthers, and Jack (2009) in the context of ebanking authentication confirms that users value usable systems over secure systems if given a choice. In that study three two-factor authentication methods were evaluated. Also, the results found that participants preferred systems that were rated as 'usable' – one of the measures being convenience – over systems that were rated as 'secure'. Additionally, it was found that systems that were rated as 'usable' were also rated as 'insecure', further demonstrating the conflict between the two camps.

A possible reason for the defiance of security implementations amongst computer users is the perception of security as an abstract concept and therefore the impact of poor security is not wholly understood (Borgida & Nisbett, 1977). For example, Sjöberg and Fromm (2001) found that computer users rated potential security risks higher when generalising to all computer users, yet the same risks were considered much less likely when considered as personal risks. Research by West (2008) shows that while some users may take precautions and protect themselves when using computers, that protection will only encourage them to engage in riskier behaviour. Therefore, it is possible that users' lack of perceived threat and the increasing complex methods being employed by experts drive users to engage in insecure behaviours such as sharing and writing down passwords.

However, another potential reason for the circumvention of security comes from various research indicating a large volume of accounts that users have to secure with passwords.

There have been a number of studies investigating the number of passwords and accounts users are expected to remember, but numbers have varied across these studies. Gaw and Felten (2006) carried out a study with undergraduate students where they were asked to log in to all their accounts and record the number of passwords and accounts they had. The researchers found that the participants had an average of eight accounts that required a password, and that users generally did not have more than three unique passwords. Participants were open about their password reuse behaviours explaining that the reuse of codes made them easier to manage. On average, participants reused each password twice. A more recent diary study on password use carried out by Hayashi and Hong (2011) indicated that participants – the majority undergraduate students – needed to protect eleven accounts using passwords. Another password diary study was done by Grawemeyer and Johnson (2011) using a more diverse sample – i.e. not undergraduate students – and found that on average participants managed eight passwords. No details were given about password reuse by individual participants, but their results showed that 69 unique passwords were used for a total of 175 services, indicating a nearly 60% password reuse rate. A large-scale study on password use was carried out by Florencio and Herley (2007) involving half a million users over a three-week period. Participants were required to download a software program that monitored their web password details, distinguishing this study from the others which were mostly self-reported. They found that users on average had 6.5 passwords that are shared across 3.9 different sites – in other words most passwords are reused. This resulted in users having 25 accounts that require a password, a much larger and alarming number than other studies reported. This discrepancy could be due to the sample used in the latter study – the requirement to download and install a piece of software would indicate that more computer literate participants took part, and computer literate people tend to have more accounts and passwords than those that are not. However, it is also possible that the other studies led to an underestimate of account and password numbers due to the methodology employed – self-report diaries. Regardless of the exact number of accounts and passwords, it can be concluded that users are required to remember a considerable amount of information that leads to insecure behaviours.

It is clear from the numerous password habit studies that a very common way of circumventing password demands is by reusing passwords across accounts. From a security standpoint, the problem with reusing passwords is that if one account gets compromised the credentials from that account – username and password – can then be

used with other accounts to access them (Ives, Walsh, & Schneider, 2004). Recently a number of usernames and passwords have been hacked and leaked for a number of online accounts including 453,000 Yahoo! credentials (Goodin, 2012b), 420,000 Formspring passwords (Ragan, 2012) about 8 million LinkedIn passwords (Goodin, 2012c), and 11 million Gamingo passwords (Goodin, 2012d). These security breaches occurred independently of the user's actions, yet further accounts could be vulnerable if any passwords were reused. These breaches are a clear demonstration of the risks of reusing passwords.

Another prevalent way of coping with the cognitive load of remembering multiple codes is writing down the codes (Inglesant & Sasse, 2010). By writing down the different codes, the user is able to maintain unique codes for different accounts. However, the problem then becomes the storage of the documentation as any attacker that gains possession of the written down codes will then have no problems gaining access to those accounts. The comprehensive study on password habits by Grawemeyer and Johnson (2011) found that while users do write their passwords down, the reuse of codes is by far more common.

Using simple codes is another way to get around having to remember multiple codes. These simpler codes, consisting of short length, narrow character set, or both, are vulnerable to brute-force attacks and therefore compromise the security of the system. However, these weak codes are still prevalent as shown by leaked PINs (Bonneau, Preibusch, & Anderson, 2012) and by research in the workplace (Stanton, Stam, Mastrangelo, & Jolton, 2005).

Other pure recall KBA systems include passphrases and mnemonic passwords. Passphrases refer to systems where the user is able to generate longer 'phrases' as a code for a system, usually consisting of multiple words. Although initially viewed as an attractive alternative to passwords due to the increased security space, recent usability studies have demonstrated a similar memorability problem to passwords where users are unable to recall the exact combination of characters (e.g. Keith, Shao, & Steinbart, 2007). Additional problems specific to passphrases included typographical errors that improved over time but affected the perceptions of the users. More recent work by Keith, Shao, and Steinbart (2009) has shown that the typographical errors can be

reduced by adding restrictions, but it is unknown what effect this has on the willingness of people to use the system.

Mnemonic passwords have also been suggested as alternatives to traditional passwords. These types of passwords are formed by abbreviating a long phrase to the first letter of each word and substituting some characters – e.g. the phrase 'We Could Be Heroes Just For One Day' could become 'WcbHj41d'. However, studies have found that users choose popular phrases for their mnemonic passwords, resulting in codes that can be cracked using specially made dictionaries (e.g. Kuo, Romanosky, & Cranor, 2006).

In summary, pure recall-based KBA systems suffer from the reuse of codes, the writing down of codes and the simplification of the codes. These problems occur due to the restrictions that are set by providers that require users to learn and remember essentially a random combination of characters.

2.2. CUED-RECALL

Challenge questions are one of the most common types of cued-recall KBA systems. Challenge questions consist of a pre-established question that users have to provide an answer to. When authenticating, the system asks the question and the user is required to respond with the answer they entered during enrolment. The question can be chosen by the user or implemented by the provider. A common challenge question that is used by several providers is "mother's maiden name" (Just, 2005).

The strength of the system lies in the users' familiarity with their personal information. Providers count on the fact that users will be able to remember their mothers' maiden names, for example, or remember the name of their first pet. As most challenge questions are fairly personal—yet generalisable across the user base—users are expected to remember the information (Just, 2004). However, remembering answers to challenge questions is not always as straightforward as intended. For example, a user's favourite food might change over time, and if they are accessing an account that has been inactive for a significant amount of time, the user might struggle to remember the registered answer—e.g. their favourite food at the time when they registered rather than their current favourite food (Just, 2004).

Another possible problem with challenge questions is regarding the registration method used for the answer. Just (2004) distinguishes between two methods for providing answers: fixed or open. Fixed answers requires user to select an answer to a question from a list of system-provided answers – e.g. a dropdown menu. While this approach controls the quality of the answers, it also has the potential to cause problems for the user: if none of the answers are relevant then there will be problems when it comes to recalling that answer. Similarly, if the system provides more than one answer that is relevant to the question then the user may be confused when asked to select the correct answer at a later date.

Open answers allow the user to enter any text they wish as an answer. This approach guarantees that the user will have an answer to a question, but the problem lies in the repeatability of the answer. For example, a user might have provided an answer as being "Queen Elizabeth II", but then they might struggle to remember exactly how they registered their answer, given that they could have registered it as "QE2", "QEII", "queen elizabeth 2", or any other variation.

One major problem with challenge questions, perhaps more so than with passwords due to the cue, is that people close to the user might be able to answer the challenge questions. For example, a friend of the previous user would probably know that his previous school was "Queen Elizabeth II" and therefore could make a very educated guess when answering the question. Just (2004) argues that the questions chosen should not be in the public domain in order to minimise these instances, but this may not always be possible, especially if it is the provider that is setting the questions.

Just and Aspinall (2009) have suggested combining multiple challenge questions in order to increase the security of the system. In a series of user study with undergraduate students, they found that participants selected questions that lead to guessable answers. When they evaluated the memorability of the multiple question approach after a 23 day delay they found that 18% of participants encountered a problem with at least one question. These results indicate that challenge questions do not provide an adequate level of security as a primary authentication system and they are also plagued by the memorability issues that have been associated with passwords.

In summary, the problem with knowledge-based systems lies in the fact that users are required to remember many complex pieces of information and they are unable to cope with this load. Therefore they engage in insecure behaviours that expose the system to other attacks, such as reusing codes across various accounts, writing the codes down, or creating simple codes. These practices compromise the security of the system, raising the question whether the efforts of security experts in making the systems more robust to attacks are in fact debilitating the systems by forcing users to circumvent these measures.

2.3. GRAPHICAL AUTHENTICATION SYSTEMS

Graphical authentication systems (GAS), also known as 'graphical passwords', are systems that rely on the user remembering visual stimuli instead of traditional text. When the user enrols with a graphical system, s/he is either given various images or is allowed to select various images that have to be remembered for logging in later – the 'target' images. Usually systems require the user to remember four or five target images that form a 'code'. In order to authenticate the user is required to select the target images from amongst foil images in a series of challenge screens. If the user selects all four or five target images correctly then s/he has authenticated. Four key graphical systems are described below followed by the psychology literature detailing their advantages. For a more detailed overview of graphical systems and an evaluation of the methods used see subsection 2.3.2.

Graphical authentication systems have been tested extensively over the past decade and have been proposed as alternatives to traditional alphanumeric passwords due to excellent memorability results (see 2.3.1). The main implementation strategy behind GAS is the hope that by reducing the memory burden of multiple passwords, the resulting fewer passwords will be strong (e.g. Jermyn, Mayer, & Monrose, 1999; Suo, Zhu, & Owen, 2005). The idea of a graphical password was first explored by Blonder (1996) who developed a click-based system and since then a number of graphical systems have been developed and evaluated with the aim of improving the authentication user experience.

Figure 2.1: The Passfaces graphical authentication system (Rose, n.d.)**.**

Passfaces (Valentine, 1998) is an example of a recognition-based GAS that utilises images of human faces for authentication (see Figure 2.1). Users are required to learn and remember five faces that they will then have to select from amongst foil faces in order to authenticate. During authentication, the user will see five challenge grids consisting of nine faces each, with one of the nine being a target face. The user is expected to select the five target face from each of the challenge grids with no other restrictions.

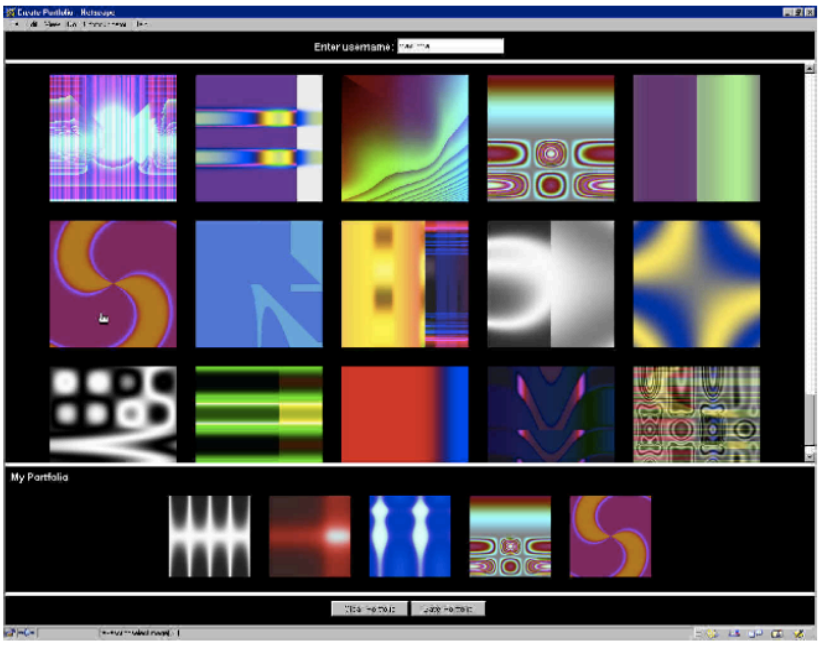


Figure 2.2: The Déjà vu graphical authentication system (Dhamija & Perrig, 2000).

Déjà vu (Dhamija & Perrig, 2000) is similar to Passfaces in that it allows users to select five images during enrolment that they are then required to recognise to authenticate (see Figure 2.2). Unlike Passfaces, Déjà vu uses images of random art with the aim of

limiting sharing and writing down of the codes. Also unlike Passfaces the system presents only one challenge grid during the authentication stage and all target images are present in that grid. The user is then required to select their five images from the single grid, regardless of order.

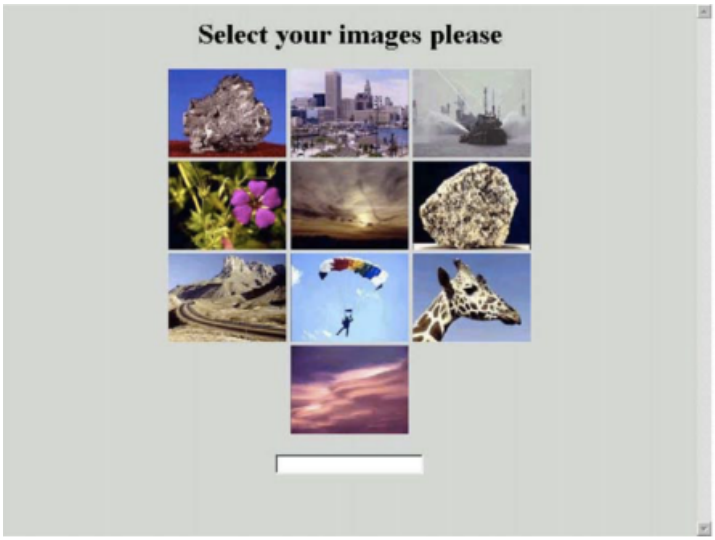


Figure 2.3:The VIP 1/2 graphical authentication system (De Angeli et al., 2005).

VIP (De Angeli et al., 2002) employs a similar setup to Passfaces, with the chief difference being the use of detailed pictures instead of faces (see Figure 2.3). Another difference is the enforcement of order in the selection process – participants must select the pictures in the same order they were chosen or given during enrolment. Three versions of the VIP system have been evaluated, one with the location of the images on the grids varying with each trial and another with the location of the images being fixed for every trial. A third VIP portfolio-based system was tested where users were given eight pictures to learn, but during authentication they were only required to select four random targets from the original eight.

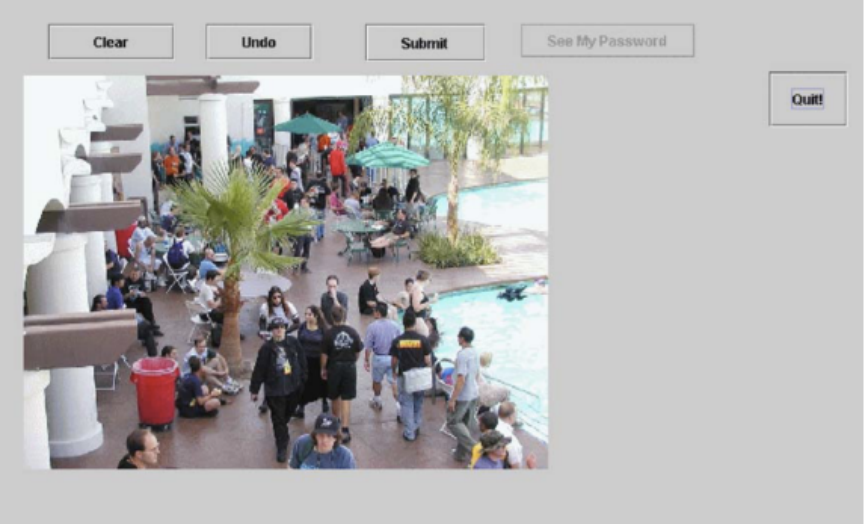


Figure 2.4: The PassPoints graphical authentication system (Wiedenbeck et al., 2005).

PassPoints (Susan Wiedenbeck, Waters, Birget, Brodskiy, & Memon, 2005) implements the graphical element in a different way to the previous three systems (see Figure 2.4). Instead of assigning whole images to participants and requiring them to recognise them to authenticate, PassPoints uses the image as a cue to aid the user in selecting five points within the screen – or within the image. In order to authenticate, the user is required to select those five points in the same order as during enrolment.

Many varieties of graphical system have been designed and evaluated, but the majority are based on the principles from the four systems covered above – face-based (Passfaces) or picture-based (Déjà vu, VIP), multiple challenge grids (Passfaces, VIP) or single challenge grid (Déjà vu), fixed (VIP) or random (Passfaces) image position, full recognition (Passfaces), portfolio (VIP 3), or cued-recall (PassPoints).

2.3.1. WHY DO THEY WORK?

Graphical authentication systems are based on three main principles that elicit high memorability in humans. These are the use of recognition or cued-recall over pure recall, the use of the picture superiority effect for picture systems and the use of face recognition.

2.3.1.1. RECOGNITION OVER RECALL

Graphical Authentication Systems typically rely on either cued-recall or recognition unlike passwords or PINs which rely on pure recall. It is well known that humans are much more accurate when asked to recognise a previously seen item than when asked to recall that item – with or without a cue (Baddeley, 1997; Parkin, 1993). A theory for the superior performance of participants with recognition tasks over recall tasks is the two-process theory supported by Kinisch (1970) amongst others (e.g. Watkins, 1979). This theory states that recall and recognition are part of the same process but recognition is a step before recall. When a participant is asked to recall an item, that item first has to be retrieved from memory using a search process – step 1 – and is then tested using recognition to determine whether the item is correct –step 2. Hence recognition avoids the additional memory load associated with the 1st step.

Another theory, the encoding specificity principle, states that memory is most effective when information that was present during the encoding process is also present during the retrieval process (Tulving & Thomson, 1973). This theory explains the superiority of recognition over pure recall by the fact that the focal item is presented during recognition and the participant is required to recover the context of that item. For recall, however, the context is given and the participant is required to recover the focal item – a process that requires more effort. Tulving and Watkins (1973) further clarify this theory by stating that both recall and recognition rely on the same processes, but during recognition more accurate cues are presented to the participant – i.e. all the information that was present during encoding is present during retrieval and the participant has to make a yes/no judgement – while for recall there are less relevant or no cues resulting in the participant having to generate the item.

In the context of GAS, the use of recall or cued-recall – shown to elicit a better performance than pure recall (Baddeley, 1997; Parkin, 1993) – suggests that they should encourage a better memory performance than existing knowledge-based systems that generally rely on pure recall.

2.3.1.2. VISUAL SUPERIORITY

Graphical authentication systems are based on the well-known phenomenon of the picture superiority effect. This pictorial advantage indicates that people are able to recognise a very large number of pictures even after limited exposure to these pictures. Nickerson (1965) was one of the first to research short-term memory for pictures by using a seen/not seen approach where participants observed a set of pictures and were then shown a larger set containing all of the initial pictures in addition to some new pictures. They were required to respond whether the picture being shown had been seen or not during the initial phase of the study. Results were very good with the majority of participants achieving over a 90% success rate, although a statistically significant decay in performance was observed with larger time lags. These results were later reinforced by Shepard (1967) who carried out a comparison of pictures with words and sentences concluding that short-term memory for pictures was much more accurate than for other stimuli. Despite similar results, the method was marginally different with pairs of pictures being shown to the participant rather than a single picture. Therefore, the response for Shepard was a choice response (a or b) rather than a match like Nickerson's (yes or no). Shepard's testing method – requiring the participant to discriminate between two or more images – is most relevant to GAS as challenge grids present users with a number of choices, rather than asking the user whether a single image has been assigned. However, both types of methodology have yielded the same results: pictures were more easily recognised than verbal stimuli.

Standing, Conezio, and Haber, (1970) further backed the idea of a picture superiority effect with a series of four experiments that found accuracy rates around 90% for a variety of configurations. In these experiments the time gap between exposure and identification varied between 30 minutes and 60 minutes, indicating retention for a longer period of time. Standing (1973) demonstrated memory for pictures after an even longer time delay, with very positive results even after delays of two days between learning and identification. The reported studies used detailed pictures which are thought to contribute to the strength of the effect. This series of studies demonstrated that the superior memory for pictures that was initially observed for a short time delay also applied to extended delays of more than a day. These results are very important for GAS given that a large number of authentication attempts will take place after extended intervals – it is unlikely for a person to continuously authenticate with all of their

accounts or to authenticate after only 5 minutes. More likely, users will authenticate once per day or a few times per week, although there will also be accounts that will be used much less often.

The picture superiority effect has been shown to extend to associative recognition, i.e. remembering pairs of items. In a series of studies by Hockley (2008), participants were shown to have a higher hit rate with picture pairs (line drawings) than with word pairs demonstrating that not only are people able to take advantage of the effect with single pictures, but also when associating multiple pictures. This finding is encouraging in the context of GAS given the associations that will need to be made between the individual items and their respective accounts. However, care must be taken when implementing systems that utilise the picture superiority effect as evidence suggests that it can be reversed – for example by presenting very similar stimuli in cued-recall scenario (Nelson, Reed, & Walling, 1976) or by presenting semantic verbal priming along with the image (Intraub & Nicklos, 1985). Additionally, the presentation of the images can also play a part in the effectiveness of the system, as demonstrated by Weldon (1987) who reported that a cued-recall word fragment completion test yields more accurate responses than a free-recall of pictures. However, participants were clearly superior in the picture identification test – essentially a recognition task – than in any other condition, demonstrating the power of recognition.

2.3.1.3. FACE RECOGNITION

The visual advantage does not appear to be exclusive to pictures as memory for human faces has been shown to be very strong. This is not the same effect as the picture superiority effect, however, as work by Bruce and Young (1986) claimed that the face recognition unit is separate from the picture recognition mechanism. Nonetheless, a clear advantage is present when it comes to recognising faces that people are familiar with. For example, Bruce (1982) demonstrates that known faces are identified faster and more accurately than unknown faces, even when expression and angulation of the face were changed. Additionally, Burton and Wilson (1999) found that familiar faces could be identified accurately even with very poor quality video, a result that was later supported by Bruce, Henderson, Newman, and Burton (2001). Bruce et al. (2001) also found, however, the recognition for unfamiliar faces was not as successful. This finding was consistent with previous work demonstrating the inefficacy in recognising

unknown faces (Hancock, Bruce, & Burton, 2000). These findings are very important for face-based GAS due to the differences in performance that have been observed. While it is obvious that using familiar faces would benefit the users, security implications have to be taken into consideration. Hence, unfamiliar faces should be used for authentication systems despite the potential deficit, but this design decision should be balanced with training to aid the users.

Distinctive faces have been found to be easier to recognise than non-distinctive faces (Bruce, Burton, & Dench, 1994), yet the term 'distinctive' has not been accurately defined. There are currently at least two methods for determining the distinctiveness of a face – by the ease of picking that face out of a crowd and the deviation from an average face (Wickham & Morris, 2003). Depending on the method used, the ease of recognition could vary. Previously it was thought that attractiveness played a part in making a face recognisable (Shepherd & Ellis, 1973), but more recent work fails to support this notion and once again suggests that distinctiveness is the more important characteristic for ease of recognition (Sarno & Alley, 1997; Wickham & Morris, 2003). Therefore, distinctive faces are the obvious choice to be used in face-based graphical systems although the implementation is not as straightforward given that if all faces used for the system are distinctive – including the foils – then no advantage will be present. Similarly, if only the target faces are distinctive, then this will aid the attacker in guessing the correct code.

2.3.2. Graphical Authentication: The Story So Far

Graphical authentication systems have been categorised into two main categories within the computing literature: recall-based and recognition-based (Dirik, Memon, & Birget, 2007; Suo et al., 2005). Recall-based graphical authentication systems require the user to replicate an action – e.g. a drawing or a selection of clicks – that they had previously performed during enrolment. The best examples of recall-based GAS are Draw-a-Secret (DAS) and PassPoint. In DAS, a user is expected to draw a shape using a grid as a guideline (Jermyn et al., 1999). For better security, the shape should cross as many cells in the grid as possible, but should be simple enough to be redrawn later during login. Dunphy and Yan (2007) developed an improved version of the system called Background Draw-a-Secret (BDAS) that presented users with a background photograph

in order to reduce the predictability of the drawings. Participants using BDAS produced more complex drawings without sacrificing the memorability of the system.

The PassPoints system presents the user with an image and the user is required to click on five distinct areas within the image. In order to authenticate the user selects the same five points in order (Wiedenbeck et al., 2005). While the system obtained favourable comments from users, a problem specific to PassPoints is that the choice of the clicks are strongly influenced by the background image that is selected or given. These popular click areas are known as 'hotspots' and they lead to a certain predictability which diminishes the security of the system (Dirik et al, 2007). Simply, every image has a number of areas of interest that draw users' attention and these can be used by attackers to guess the clicks of the users (e.g. van Oorschot, Salehi-Abari, & Thorpe, 2010).

With the issue of hotspots in mind, an improved system called Cued Click Points was developed and evaluated by Chiasson, Van Oorschot, and Biddle (2007) where five images were used and users were required to select one point per image rather than having to select multiple points in a single image. Results were promising, with accuracy being very high and user satisfaction being higher than that of the original PassPoints system. Additionally, the user study found that the points selected by participants using Cued Click Points were less predictable than those using PassPoints while also yielding good point-selection accuracy. These results were based on a thirty second delay which may have been a very short time delay. The system was further improved to aid users in selecting their points by the use of a 'viewport' that focuses their attention on a random section of the image with the ability to shuffle that viewport if necessary (Chiasson, Forget, Biddle, & Van Oorschot, 2008). This iteration of the system, called Persuasive Cued Click Points, was aimed at further reducing the predictability of point selections and the results found that the memorability of the system after 30 seconds was not significantly different from that of Cued Click Points (Chiasson et al., 2007). As hoped, the number of hotspots were significantly reduced with the new version of the system. However, the studies evaluating PassPoints, Cued Click Points and Persuasive Cued Click Points only evaluate a single code, diminishing the applicability of the results in the real world where users are likely to have more than one code. Chiasson, Forget, Stobert, Van Oorschot, and Biddle (2009) addressed this issue by running a study where participants were asked to learn and remember six codes

– either passwords or PassPoints. Participants were asked to recall their codes after a 'short break' and following a two-week delay. Results differed from those reported using single codes, with the most striking finding being that login success rates between passwords and PassPoints did not differ significantly after a two week delay, although PassPoints were more successful after a short break. The results from this study demonstrate the flawed methodology that had been employed with cued-recall based GAS and suggests that although an advantage still exists with graphical systems over traditional text-based systems, that advantage may disappear after an extended delay when multiple codes have to be remembered. However, there are some issues with the methodology that was used. The exact time delay for the first retention phase (short-term recall) is not specified as participants were asked to carry out untimed tasks – i.e. answer two questions. It is unknown whether these tasks were controlled in some way to make sure all participants experienced the same time delay between confirmation and login. Additionally, all six codes were assigned during the same session, potentially overloading participants.

Results with recall-based graphical authentication systems have been generally positive (e.g. Chiasson et al., 2007; Chiasson et al., 2008; Wiedenbeck et al., 2005), but a consideration for these systems is that a successful authentication requires the user to recreate an action to a very accurate degree and therefore mistakes should be expected. The tolerance of the area where the click or selections are to be made plays a big part in the success of the systems, as a margin that is too small will cause errors while a margin that is too big will be exploited by guess attacks (Susan Wiedenbeck & Waters, 2005).

On the other hand, recognition-based GAS require users to learn and remember a number of images and then select them from amongst foils during authentication. Examples of recognition-based graphical systems include Déjà vu, the Visual Identification Protocol (VIP), and Passfaces. In Déjà vu a user selects five different random art images from a large portfolio and is then required to select those five images from a set on the screen. Déjà vu uses random art images due to the abstract nature of the images which aims make it more difficult to both share and write down, features that are perceived to make the system stronger (Dhamija & Perrig, 2000). A user study found that participants encountered less failed logins when using Déjà vu than when using PINs or passwords. However, unlimited attempts at selecting the codes were allowed which influenced misleading results – users are generally not allowed unlimited

attempts to enter their codes in the real world due to security concerns and instead are usually given three attempts before the account is locked (e.g. Brostoff & Sasse, 2003). The most likely reason for the performance found in the study was the testing of a single code which made the task much easier but also less ecologically valid – it is unlikely that only two failed attempts overall would occur if participants had been asked to remember four codes. No statistics were used to compare the results of the GAS to PINs and passwords, so although GAS appear to be superior when comparing the percentages it is not known whether the difference is statistically significant. The lack of inferential statistics is particularly missed in the results after a week's delay where the percentages are close between PIN and passwords and between Déjà vu and Photo. It would have been interesting, and helpful, to determine whether any differences in performance existed between those groups.

VIP is a graphical authentication system that relies on image recognition for authentication (De Angeli et al., 2002). The premise of the system is that users are given four target pictures of detailed and colourful objects to learn and they are expected to select the four target pictures from four challenge grids to log into the system. Each challenge grid consists of ten pictures, but only one is a target picture. In order to authenticate, the user needs to select the target picture from each challenge grid. A number of different configurations of the system were tested: four target pictures with their position in the challenge grids always being the same (VIP 1), four target pictures with their position in the challenge grids being random (VIP 2) and a portfolio-based system where eight target pictures were given to participants but only four were shown in the challenge grids in random order (VIP 3). A PIN system was also evaluated for comparison purposes. The general results found that participants were less prone to making errors when using VIP than when using PIN after a one-week delay. However, the portfolio-based configuration, VIP 3 led to significantly worse accuracy when compared with the other three systems. These results are very important for GAS as they demonstrate that graphical systems can be designed poorly and that in those cases the picture superiority effect can be eliminated. A problem with the study, however, is the lack of detail regarding the assignment of codes. It is clear that participants were given a single graphical code or PIN, but no further details are given regarding how the codes were chosen. Additionally, the use of single codes once again yields unreliable results, as demonstrated by the differences in performance between single and multiple graphical cued-based codes (Chiasson et al., 2009). Moncur and LePlâtre (2007)

evaluated the performance of participants with five codes, all of which were either PINs or VIP 2. The study found that participants were more successful when remembering multiple pictures over a four-week period, although retention rates were relatively poor for both conditions with VIP having about a 25% retention rate. As with the original VIP study (De Angeli et al., 2002) the codes were assigned to participants, but once again no details are provided on how this was done. Another issue lies in the fact that all five codes were assigned to participants in the same session, possibly taxing their concentration and affecting their performance. The researchers attempted to address this by requiring participants to practice their codes only two times before moving on, rather than the ten times employed by De Angeli et al. (2002), but this might have led to poor encoding of the codes due to lack of practice.

Passfaces is a commercial GAS that works in a similar way to VIP 2 – where users are given four target face images to remember and then have to select those faces from a set of foil face images over the course of four challenge grids (Rose, N.D.). The chief difference between the systems lies in the stimuli – Passfaces utilises images of human faces while VIP uses detailed pictures of objects. Additionally, the challenge grids contain nine images rather than VIP's ten. The basis for this system is humans' exceptional ability to recognise known faces (see 2.3.1.3). As the system uses a database of stock face images, the user is guided through a 'familiarisation' process where they have to answer a number of questions about each face with the aim of turning that unknown face into a known face (Rose, N.D.).

In a lab-based study, Valentine (1998) tested frequent and infrequent users of Passfaces with undergraduate students and university staff and found that although frequent users remembered their Passfaces better—with 99.98% of users logging in successfully first time—infrequent users still fared very well, with over 80% of users logging in successfully first-time and 100% logging in within three attempts. It must be noted that in Valentine's study (1998) participants were allowed to select their faces and this resulted in predictable choices, especially for males who chose female faces over 80 percent of the time (Valentine, 1998). The predictability of users' choice of faces was later confirmed by Davis, Monrose, and Reiter (2004) who found that 10% of males' codes could be guessed by just two guesses.

Despite the findings of studies suggesting user-chosen Passfaces codes were insecure, a field trial was run to evaluate the performance of undergraduate students with the system over a period of five months where participants were allowed to select their faces (Brostoff & Sasse, 2000). The study found that participants using Passfaces experienced significantly less problems logging in than when they used passwords. This improvement was offset by longer login times and fewer logins in total with Passfaces. The researchers argue that participants were put off by the long login times, but this may not be a problem with current implementations of the systems running on faster hardware and faster internet connections. A problem with the methodology, however, was the use of a single code and therefore ignoring the problems that are associated with multiple codes, as demonstrated by the other multiple code studies.

Everitt, Bragin, Fogarty, and Kohno (2009) conducted a study where undergraduate students were asked to learn one, two or four Passfaces codes to authenticate with over four weeks. The codes in this study were assigned for security reasons, as detailed above. The purpose of the study was to explore problems that might arise when multiple codes are assigned to participants and to determine whether the training that is used – i.e. when the codes are assigned – affects recognition. The results demonstrate that the addition of codes significantly affected the number of attempts that participants needed to log in as well as an increased failure rate. However, no comparisons were made with existing authentication systems such as passwords or PINs so it is unknown how the performance with multiple Passfaces compares with that of multiple passwords. Due to the differing methodologies and the various measurements that were recorded, it is not possible to accurately compare the performance with other studies that have evaluated multiple codes.

2.3.3. TESTING PARADIGMS

An important observation about all studies that were conducted to evaluate the performance of GAS is the use of different methodologies where the individual differences, system, and intervals varied (see Table 2.1 for a summary of methodologies used).

Table 2.1: Methodologies used for evaluating graphical authentication systems. L/F = Lab-based or Field-based study. Interval = delay between learning and recalling/recognising. Measure = dependent variable. Analysis = method of analysis for the dependent variable. Codes = number of different codes participants had to remember and whether they were assigned or chosen.

Study	n=	L/F	Interval	Measure	Analysis	Codes	Finding
Passfaces (Brostoff & Sasse, 2000)	36 (UG)	F	Varying	Error Rate	Percentage	Single Chosen	PF remembered better than PW but time expense
Déjà vu (Dhamija & Perrig, 2000)	20 (N/A)	L	~10 mins 1 week	Time Failed attempts (from unlimited)	Percentage	Single (per system) Chosen	Users liked system and compares well with existing systems
VIP (De Angeli et al, 2002)	61 (mix)	L	40 mins 1 week	Effectiveness (forgetting code or wrong entries) Time	Chi Square (Effectiveness) ANOVA (time)	Single (per system) Given	Pictures less error prone and more liked. VIP 3 (portfolio) not successful
Never Forget (Weinshall & Kirkpatrick, 2004)	N/A	L	1-3 months	Successful attempts?	Percentage	Single? Given?	Pictures are better
User Choice (Davis et al., 2004)	154 (UG)	F	Varying (~ 5 months)	Successful attempts	Percentage	Single Chosen	Users cannot be trusted to select images (security)
PassPoints (Wiedenbeck et al., 2005)	40 (mix)	L	~10 mins 1Week 4 Weeks	Attempts Time	t-tests	Single (per system) Chosen	Same memorability as password over extended period but PP takes longer
Personal Photos (Tullis & Tedesco, 2005)	14 (N/A)	L	Immediate 30 days	Accuracy (errors)	Percentage	Single (per system) Chosen + Given	Personal photos very memorable over time
CHC (Wiedenbeck et al, 2006)	15 (mix)	L	Immediate '1 week' (just recog.)	Correct attempts Time	Percentage (attempts) ANOVA (time)	Single Given?	Good memorability over time but at expense of time
Handwing (Renaud & Ramsay, 2007)	28 (mostly old)	F	Varied over 2 years	Successful attempts	Percentage	Single Chosen	Better than PIN
BDAS (Dunphy &	46 (UG)	L	5 mins 1 week	Success Rate	Percentage	Single Chosen	Complexity of code increases

Yan, 2007)							and mem. does not decrease
VIP (Moncur & Le Plâtre, 2007)	172 (mostly young)	L	Immediate 2 weeks 4 weeks	Successful logins	Chi Square	Multiple (5) Given	Drop in performance after delays, but performance better than text passwords
CCP (Chiasson et al., 2007)	24 (young)	L	~2 mins	Accuracy (pixels from original)	Percentage	Single independent Chosen	Users found CCP easier to use than PP
PCCP (Chiasson et al, 2008)	37 (young)	L	~2 mins	Success rate	Percentage CHI Square	Single independent Chosen	Better selection (security) without impacting usability
ColorLogin (Gao et al, 2008)	30 (N/A)	L	None - consecutive	Time	Means	Single Given?	Faster than other graphical systems
PassPoints (Chiasson et al. ,2009)	65/26 (N/A)	L	Short Break ~2 weeks	Success Rate	Chi Square	Multiple (6) Chosen	Performance affected after long delay to same extent as text passwords
Passfaces (Everitt et al., 2009)	110 (UG)	L/F	Varying over 4 weeks	Failure Successful Attempts	Chi Square	Multiple (1, 2 or 4) Given	4 week delay flawless for single system, significantly more difficult for multiple
Graphical PIN (Brostoff et al., 2010)	51/54 (6 old total)	F	Varying (1-75 days)	Errors (and types)	Percentage	Single Chosen	Easy to use
Faces vs. Pictures (Hiywa et al, 2011)	20 (N/A)	F	Varying	Time	t-tests	Single (per system) Given	Pictures (objects) preferred to faces
ImagePass (Mihajlov & Blazic, 2011)	211 (UG)	L/F	Var: A: ~1 week B: ~1 month	Login Failure	ANOVA	Single	Frequency of use improves performance
Study	**n=**	**L/F**	**Interval**	**Measure**	**Analysis**	**Codes**	**Finding**

Individual differences refers to the demographics of the participants that were tested. The majority of GAS studies focus on mixed ages or a sample of undergraduate students. The selection of undergraduate students has been an opportunity sample, rather than a strategic sample. Due to these recruitment methods, no studies have looked at comparisions of age-specific performance with graphical systems, or gender-specific

performances meaning that GAS performance results can only be applied to younger users. Few studies have evaluated systems with specific age groups and these studies will be covered in the following chapter.

System differences refers to the system being tested. Studies so far have greatly varied in their approach, with some studies comparing the new GAS to existing systems (e.g. Dhamija & Perrig, 2000; Wiedenbeck et al., 2005), others only reporting the results from that single system (e.g. Brostoff & Sasse, 2000; Wiedenbeck et al., 2006), or testing variations of the same system (e.g. De Angeli et al., 2002; Chiasson et al., 2008). Due to the different approaches researchers take with the testing of systems, it is sometimes not possible to compare the different systems with each other – e.g. Déjà vu with VIP. Additionally, the administration of the codes has varied across different studies, with some researchers assigning the codes to participants and other researchers allowing the participant to select their codes.

Intervals refers to the time delay(s) that participants have to endure between the encoding of the codes to the recalling of the codes. Studies have evaluated systems in short-term memorability, long-term memorability, or both. Yet, even when evaluating short-term memorability the intervals used vary significantly between immediate recall (Tullis, Tedesco, & McCaffrey, 2011; Wiedenbeck, Waters, Sobrado, & Birget, 2006) and over thirty minutes (De Angeli et al., 2002). Similarly, the intervals for long-term memorability vary greatly from one day (Brostoff, Inglesant, & Sasse, 2010) to four weeks (Wiedenbeck et al., 2005). These varying time intervals contribute to the confusion between systems as the memorability of a system after a week cannot be compared to that of another systems' after two weeks.

Moving forwards, it is imperative to control these three factors to make testing more consistent and easier to compare. In this thesis, the focus is also on evaluating how well older adults are able to deal with existing authentication methods and whether the introduction of GAS would improve their current performance. The following chapter will discuss the authentication literature in the context of older adults as well as covering the memory literature for older adults.

2.4. SUMMARY

This chapter presented an overall discussion on the existing authentication literature and the shortcomings of existing systems. An important conclusion from the literature is the problem that users face when remembering multiple passwords, and the fact that this well-studied problem has not deterred providers from a.) implementing passwords or b.) attempting to improve the mechanisms associated with password allocation. Additionally, graphical authentication systems (GAS) were covered due to their potential to improve the authentication experience for users, but a review of the literature demonstrated the lack of consistency in evaluation methods which results in difficulty when comparing results from different graphical systems.

Of more relevance to this thesis, it has been shown that older adults have not been included in the development or evaluation of password mechanisms. There are reasons to believe that older adults would be disadvantaged by existing knowledge-based authentication (KBA) systems, and these reasons will be covered by the ageing literature in the following chapter.

3. AGE LITERATURE

Authentication methods have been developed and tested extensively over the past few years, yet not much is known about the state of authentication in the context of older adults. For years designers have argued that older adults accounted for a very small proportion of users and therefore it did not make sense to cater for them. However, older adults are the fastest-growing group of Internet users in developed nations (Hart et al., 2008), meaning that computer uptake amongst older users is on the increase. There are a number of well known age-associated memory declines (see 3.1 below) that can affect the use of current authentication systems and these will have to be addressed to improve the experience for this user group. While both cognitive and physiological declines are expected with the ageing process, this chapter will focus on cognitive declines. Physiological declines will play a role in the usability of authentication systems, e.g. decline in motor skills making it harder to operate a computer mouse (N. Walker, Philbin, & Fisk, 1997), but the true challenge lies with the cognitive declines, specifically in memory, given the emphasis on learning and recalling codes.

3.1. AGE-ASSOCIATED MEMORY DECLINES RELEVANT TO AUTHENTICATION

It has been well documented that declines in cognitive ability occur as people age (Fisk, Rogers, Charness, & Czaja, 2004). However, it is important to note that the rate of cognitive decline is variable amongst the ageing population and can be very different even within adults of the same age (Fisk et al., 2004). It is generally accepted that these age-associated declines become more apparent in adults aged over 50 (Arbor, 2001) and noticeably so after the age of 60 (Salthouse, 1991).

Perhaps the most prevalent age-associated decline to affect KBA systems is memory. Memory has been traditionally divided into four processes: sensory memory, short-term memory, working memory and long-term memory. Sensory memory relates to the initial process after an item has been perceived – approximately the first 200-500 miliseconds – and is subject to fast degradation and as a consequence information cannot be stored for later retrieval. As such, sensory memory is not relevant for authentication. Short-term memory refers to the limited capacity process that can lead to

information storage with adequate rehearsal and can be relevant to authentication regarding the learning process. Working memory is a system that holds multiple pieces of information with the purpose of further manipulation. Finally, long-term memory provides the storage facility for information and is thought to be infinite in capacity. However, certain memories may become difficult to retrieve and as such they can be 'forgotten'. It is this process – long-term memory – that is arguably the most relevant to authentication.

Short-term memory has been shown to become less reliable with age (e.g. Akatsu & Miki, 2004). The decrement is even more pronounced in older adults when they are required to remember the order of the items. Maylor, Vousden, and Brown (1999) demonstrated this problem when participants were asked to learn letters in the order they were given and then recall the list immediately after the presentation of the last letter. Older participants scored significantly lower than younger participants and made more errors of every type, such as missing out a letter, recalling the letters in the wrong order and simply just forgetting the sequence. Kay (1951) suggested that the reason older adults struggle with serial order recall is due to their inability to adapt their approach when they make an error while recalling. This results in the older participants learning the wrong sequence rather than amending the sequence. Younger adults, on the other hand, are able to recognise when they have made a mistake and take steps to try a new sequence.

Perhaps more importantly in the context of authentication, non-semantic long-term memory has also been shown to be affected (Fisk et al., 2004). Kausler, Salthouse and Saults (1988) demonstrated this long-term memory deficit by testing younger and older adults with memory for single words and word pairs. Results show that older participants struggled significantly more than younger adults in recalling the words over a period of one week. A meta-analysis by Verhaeghen, Marcoen, and Goossens (1993) confirmed the long-term memory problem when they concluded that long-term episodic memory was generally more affected by ageing than short-term memory. The exact reasons for long-term retention decrements are not known, but there are a number of explanations. The first possible explanation for the age-related deficits in long-term memory that have been observed relates to encoding strategies. Mitchell (1986) found that older participants performed on par with younger participants when asked to recall and recognise previously seen objects only when no instructions were given to learn

those items. When participants were told that there would be a test following the presentation of the items younger participants outperformed the older ones. This deficit in expected recall suggests that older adults may employ faulty learning strategies, or at least they are not as effective as those used by younger adults. A study by Grady, McIntosh, Horwitz, and Maisog (1995) found that cerebral blood flow was less prevalent in appropriate areas of the brain in older adults when compared with younger adults. This was not the case during recognition, when cerebral blood flow was equal for both age groups, which again suggests that the encoding strategy that older adults employ is at fault. Koutstaal and Schacter (1997) found that older adults relied on general conceptual or perceptual similarity in picture (object) recognition – e.g. they made a very broad and unspecific association with the object when they were asked to remember it. This strategy led to poorer recognition when similar pictures were used as foils. On the other hand, younger adults did not seem to suffer as much with the task, further supporting the idea that older adults employ a faulty strategy at encoding. Naveh-Benjamin, Brav, and Levy (2007) further reinforced the notion of poor encoding strategies when they found that when older participants were required to use an association strategy to learn word pairs they performed significantly better than when they were allowed to encode the pairs on their own. These findings are very important in the context of authentication as they suggest that older adults need assistance when learning their codes – graphical or otherwise.

Another explanation for the long-term memory decrement found in older adults is that the encoding process is hindered by distractions. A theoretical model by Hasher and Zacks (1988) suggested that older adults were more distracted than younger adults by irrelevant information when encoding items and these distractions were then encoded along with the central information, causing more information than was necessary to be encoded and as a consequence affecting the recall of that information at a later time. An EEG study found that indeed older adults appeared to pay excessive attention to distracting information (Gazzaley et al., 2008). These findings are also important for the design of authentication systems – older adults need to be in a quiet environment where they can learn their codes, but there are also design implications. Codes should be presented using a minimalistic approach to minimise any distractions on screen and direct their attention to the items.

Working memory is another process that is affected by the ageing process. Older adults have been shown to have more difficulties with tasks that require manipulation of information than with those that do not (Dobbs & Rule, 1989). The problems with working memory are likely to lead to the item binding deficit that has been observed in older adults. Short-term memory for individual items is fairly intact, but when older adults are asked to associate two or more items the accuracy of recall declines. A meta-analysis by Old and Naveh-Benjamin (2008) confirmed the strong decrement in item association and old age. This decrement has been observed with face pairs (Bastin & Van der Linden, 2006) object drawings and location (Chalfonte & Johnson, 1996), word pairs (Naveh-Benjamin, 2000) and picture (objects) pairs (Naveh-Benjamin, Hussain, Guez, & Bar-On, 2003). Associations were thought to be easier when they were different types of pairs, such as pictures and words (e.g. Wheeler & Treisman, 2002) but this was countered by Bastin and Van der Linden (2006) who found that performance in associative memory for older adults was equally poor for all types of associations. The binding deficit notion was backed up by an fMRI study (Mitchell, Johnson, Raye, & D'Esposito, 2000) that found greater activation in appropriate brain regions in younger adults when compared to older adults when asked to remember two items – e.g. binding the two items. No difference in activation was observed between the younger and the older participants when a single item was remembered. The binding decrement is very relevant to authentication as users are generally required to remember a number of items together – numbers, letters, or even pictures.

Remembering the location of stimuli is also a problem for older adults, with item memory being negatively affected when location is required to be encoded (Park, Puglisi, & Sovacool, 1983). Similarly, Chalfonte and Johnson (1996) found that memory for object location was significantly poorer in older adults than that for the item itself and the colour of the item. A possible explanation for Chalfonte and Johnsons's (1996) results is that participants were being asked to recall an item and the location, therefore they were required to bind the two items together. As discussed previously, older adults are known to have problems with the binding process, a problem that was also addressed by Chalfonte and Johnson (1996). Nonetheless, it is important to keep this weakness in mind as GAS present images in locations throughout challenge grids. Although it is usually not necessary to remember the location of the image – even though it is meant to help in such systems as VIP 1 (De Angeli et al.,

2002) – it would be interesting to determine whether the presentation of images in set grid positions affects participants' performance.

When taking into account all the declines detailed above it becomes clear that current computer users have every chance of being excluded from the digital world when they age. Given that current authentication systems—most notably passwords—rely heavily on the users' memory, the inevitable decline of memory with ageing will affect the usability of the systems.

3.2. Authentication and Older Adults

The number of age-related memory declines implies that older users may find it increasingly difficult to deal with systems that require them to remember information—passwords for example (Sayago & Blat, 2009). Despite this evident observation, very little research has been done to evaluate the performance of older adults with authentication systems.

Rasmussen and Rudmin (2010) conducted a survey regarding PIN use and habits and found that older adults self-reported more problems remembering their codes than did younger adults. However, the true extent of the problem is unproven as Rasmussen and Rudmin's (2010) findings arise from self-reported measures and thus are subject to bias. Previous experimental evidence of older adults' performance when remembering multiple four-digit numbers found a marked accuracy decline over time which tends to support the self-reports (Derwinger, Stigsdotter Neely, MacDonald, & Bäckman, 2005). The experimental study, however, was designed to investigate the effects of various training methods on forgetting of four-digit numerical codes, rather than evaluating the memorability of PINs. Additionally, the study was not designed or carried out with authentication in mind, despite the use of four-digit numbers (i.e. PINs). This study by Derwinger et al. (2005) is the only study to demonstrate the performance of older adults when learning and remembering multiple PINs, although no data was collected on younger adults making it impossible to determine the extent of the problem or whether it does exist. More importantly, however, is the fact that the time delays do not match those used by studies evaluating the performance of multiple graphical codes, making it impossible to compare the results of the existing system with the new systems.

3.2.1. Graphical Authentication Systems

3.2.1.1. Why use GAS with older adults?

Psychology literature documenting age-specific memory declines seems to suggest that older adults may benefit from using graphical systems for authentication. This literature is reviewed below.

Older adults have been shown to benefit from the picture superiority effect. Winograd, Smith, and Simon (1982) found that when younger and older participants were asked to learn pictures and words over a short period of time there were no age effects present, meaning that older participants were able to match the accuracy of the younger participants. Park, Puglisi, and Smith (1986) demonstrated this effect when learning and recognising complex pictures both immediately after presentation and four weeks following the presentation. No age effects were found when the pictures were detailed – e.g. background as well as object – but a long-term decrement in older adults was observed when the pictures were not detailed. In a follow up study, older participants did not show any decline in recognition until after a one-week delay, further backing the picture superiority effect in older adults (Park, Royal, Dudley, & Morrell, 1988). Smith, Park, Cherry, & Berkovsky (1990) further support evidence for the picture superiority effect when using detailed/complex pictures, and add to the body of knowledge by concluding that abstract pictures elicit problems with recognition at a later time, suggesting that systems like Déjà vu that rely on abstract art may not be suitable for an older user group. However, GAS that utilise detailed pictures may encourage better memorability for older adults.

Older adults have also been shown to be very apt at recognising faces that they are familiar with, but also very inaccurate when recalling faces they are unfamiliar with (Searcy, Bartlett, & Memon, 1999). Smith and Winograd (1978) carried out a study where younger and older participants were asked to learn and later recognise 30 faces and they found no significant difference in the number of faces that were recalled by young and old participants. Searcy et al. (1999) report that older adults are more prone to making mistakes when identifying faces they are unfamiliar with, as demonstrated by three studies in the area of eyewitness identification. The differences in methodology

between these two studies must be noted to explain the conflicting findings. In Smith and Winograd (1978), participants had to recognise the exact face images that they were shown at the beginning, making it a straightforward recognition task. On the other hand, Searcy et al. (1999) showed participants a video containing a number of people and were later asked participants to select a mugshot from a lineup that matched one of the people on the video – e.g. not a straight recognition task. In the context of GAS, Smith and Winograd (1978) is the more relevant methodology as users are given the face images to start with and are then required to select the exact face images from amongst foils – a straight recognition task. What Searcy et al.'s (1999) results show is that users would experience recognition problems if multiple poses would be used for each face.

As with younger adults, older adults experience own-age effects (e.g. Fulton & Bartlett, 1991; Lamont, Stewart-Williams, & Podd, 2005) as well as own-race effects (e.g. Brigham & Williamson, 1979) when recognising faces (see Chapter 7 for more details on these effects).

The implementation of graphical systems that use either cued-recall or recognition for authentication instead of pure recall is likely to benefit older adults greatly (Merriam & Cunningham, 1989). Schonfield and Robertson (1966) present evidence of older adults' memory advantage when asked to recognise word pairs than when asked to recall them. In fact, the older group were shown to be as accurate as the younger group in the recognition condition, but the accuracy dropped significantly with age in the recall condition. Craik and McDowd (1987) discuss how older adults perform better at recognition tasks than cued-recall tasks and suggest this is due to recognition requiring less processing resources than recall. Additionally it could be argued that recognition tasks provide the participant with contextual information and older adults have been shown to benefit from contextual integration when it comes to recalling pictures (Park, Smith, Morrell, Puglisi, & Dudley, 1990).

3.2.1.2. HISTORY OF GAS WITH AN OLDER ADULT USER BASE

Most graphical authentication systems have not been tested with older adults, but when thinking about the processes involved one has to think that the systems have got potential with an older user base. Renaud (2005) evaluated a GAS tailored to older adults called Handwing. Handwing, was aimed at low-security applications such as

forums where users were not protecting valuable information. Older adults were asked to draw a simple picture or write down personal information (e.g. postcode) and then upon login the user was presented with a number of hand-drawn images or bits of information. The user had to select their drawings in order to authenticate, essentially a combination of handwriting and drawing recognition with additional context. Field tests have demonstrated that the usability and memorability of the Handwing system are good, with the majority of older adults being able to authenticate without many issues (Renaud & Ramsay, 2007). Additionally, enthusiasm for the system was high, with all respondents ascertain they would rather use Handwing over a password. Table 3.1 demonstrates the poor state of authentication evaluation with older adults.

Table 3.1: Authentication systems tested with an older adults population. L/F = Lab-based or Field-based study. Interval = delay between learning and recalling/recognising. Measure = dependent variable. Analysis = method of analysis for the dependent variable. Codes = number of different codes participants had to remember and whether they were assigned or chosen.

Study	n=	L/A	Interval	Measure	Analysis	Codes	Finding
Numbers in Old Age (Derwinger et al., 2005)	60 (Old)	L	Immediate 30 min 24 hr 7 week 8 month	Items recalled	ANOVA	Six Given	No mnemonic better for long-term recall
Handwing (Renaud & Ramsay, 2007)	28 (mostly old)	F	Varied over 2 years	Successful attempts	Percentage	Single Chosen	Better than PIN

While the results are encouraging and suggest that GAS may be the way forward in this domain, the system is unlikely to be feasible for large-scale implementation due to security concerns such as guessability.

3.2.1.3. OLDER ADULTS AND AUTHENTICATION

There are a number of key aspects that should be highlighted for future design of authentication systems. First is the clear weakness that older adults exhibit when remembering information in the short-term and the long-term, although procedural long-term memory remains unaffected. Second, psychological literature shows that older adults appear to benefit from remembering visual stimuli more than verbal stimuli both short-term and long-term. Finally, limited evaluation of older adults with graphical

authentication systems shows that this age group does appear to take advantage of the visual nature of the systems.

3.2.1.4. LOOKING FORWARD

Generally, graphical authentication systems have not been tested with older adults with the exception of Renaud's Handwing (2005). Although evidence is weak, the psychology literature suggests that older adults should not suffer a strong performance decrement when using graphical authentication codes (Brown & Park, 2003; A. Smith & Winograd, 1978). Therefore, the age-related performance decay associated with text password systems could be attenuated by the introduction of graphical systems. In order to validate this theory first it is imperative to benchmark existing authentication systems to determine what the extent of the problem is. Following these results other authentication systems should be evaluated using the same methodology in order to allow comparisons between the systems.

First, PINs will be evaluated with younger and older adults over the course of three weeks. The following study will evaluate a novel GAS which aims to use context to aid users in the recognition of segments from one image. Two GAS will then be tested, one picture-based system that utilise full images that are contextually grouped, and a face-based system. Finally, an improvement upon the face-based system will be evaluated.

4. BENCHMARKING EXISTING AUTHENTICATION: PINS

4.1. RATIONALE

Personal Identification Numbers (PINs) are a knowledge-based authentication system that require users to remember several digits – usually four – and enter those digits in the correct order to gain access to an account. PINs are one of the most ubiquitous authentication systems in use today (Weiss & De Luca, 2008). Millions of people use them every day to withdraw cash at Automated Teller Machines (ATMs) or to pay for purchases using credit and debit cards. Additionally, many other everyday systems require PINs for protection – e.g. mobile phones, library cards and house alarms.

Despite their nearly ubiquitous use, PINs have a negative reputation for memorability amongst both the general population and the older population (Sasse et al., 2001; Vines, Blythe, Dunphy, & Monk, 2011). Rasmussen and Rudmin (2010) found through a large-scale survey that older adults reported more difficulty in remembering their PINs than younger adults, although the results were based on self-reporting. In terms of experimental findings, only studies evaluating new authentication systems have utilised PIN as a control system (e.g. Dhamija & Perrig, 2000; De Angeli et al, 2002; Moncur & LePlâtre, 2007; Weiss & De Luca, 2008; De Luca et al., 2010) and none of them have tested the recall of multiple codes. While some limited information can be collated by these studies in relation to PIN memorability, the different goals of the studies and the range of methodologies employed make it very difficult to obtain a reliable benchmark measure. Additionally, PINs – and authentication methods as a whole – have generally been evaluated with single codes thus ignoring the main problem associated with authentication codes – multiple codes for multiple accounts. Therefore, despite a negative reputation associated with PINs, there appears to be little objective evidence to confirm the assumption.

The most relevant empirical research of older adults' memory for PINs comes from the psychology literature, where Derwinger et al. (2005) evaluated three different training programmes for learning and remembering six 4-digit numbers. The study found that participants using a self-generated strategy were more successful recalling the numbers after an extended delay than those either not using a strategy or those using a

mnemonic. Additionally, participants were not required to associate the numbers with accounts which would have added an additional burden to the task, but would have made the results more ecologically valid. Finally, no younger adults were tested therefore it is not possible to say if the results are representative across the lifespan. One can assume the superior performance of younger adults over older adults based on the results of Macdonald, Stigsdotter-Neely, Derwinger, and Bäckman (2006) who predicted an accelerated rate of forgetting for older adults based on remembering multiple numbers over different time intervals.

The purpose of the following study was to produce experimental evidence on the performance of two age groups – a young group and an older group – when learning and remembering multiple PIN codes over the course of three weeks. The aim was to ascertain whether the memorability of PINs was as problematic as depicted, and to explore how older users coped when asked to remember multiple PINs associated with multiple accounts. Additionally, this study would establish a benchmark measurement for performance with authentication systems to be referenced by future authentication studies in this thesis. Participants were asked to learn and remember either 4 PINs (low load) or 6 PINs (high load) consisting of four digits each over the course of three weeks. Half of the PINs were assigned during the first week and the second half of the PINs were assigned during the second week. Participants were evaluated based on the average number of successful recalls and based on the average time taken to recall the PINs. This was the first study to directly evaluate performance of younger and older adults when remembering multiple PIN codes and the first to be interested in multiple codes, hence the introduction of a high load/low load condition.

It is expected that younger adults will be more accurate than older adults when remembering multiple PIN codes, based on the self-reported findings from Rasmussen & Rudmin (2010). It is also expected that participants remembering a low load will be more accurate recalling their codes than those remembering a high load. Finally, it is predicted that accuracy will decline over an extended delay period with the older group being more affected due to cognitive declines associated with the ageing process (e.g. Fisk et al., 2004).

4.2. METHOD

4.2.1. DESIGN

The study consisted of two factorial designs as the codes were separated into 'original' accounts (SET 1) assigned to participants during the first week and tested in weeks 1, 2 and 3, and 'new' accounts (SET 2) assigned to participants during the second week and tested in weeks 2 and 3. The first set of PIN codes were tested in a 2 (participant age: young; old) x2 (cognitive load: high; low) x3 (week of testing: week 1, 2, 3) factorial mixed design. The second set of PIN codes were tested in a 2 (participant age: young; old) x2 (cognitive load: high; low) x2 (week of testing: week 2, 3) factorial mixed design. There were therefore two independent factors – the participants' age (young group, old group) and the cognitive load (four PINs – low – and six PINs – high) – and one repeated factor – the testing week (weeks 1, 2, 3).

Dependent factors comprised the average number of successful code entries (maximum of five per account) and the average time (in seconds) taken to select the four digits in a code.

4.2.2. PARTICIPANTS

36 participants were recruited into one of two age groups, the younger group (18-30 years old, n=18) or the older group (65-75 years old, n=18).

Younger participants (mean age: 21, SD: 3.07) were recruited from the student population in the university using an online participation pool maintained by the university. Given the cultural diversity of the student population this sample was considered adequate.

Older participants (mean age: 70, SD: 3.83) were recruited using the lab's participant database as well as through an advert on the Elders' Council of Newcastle newsletter and various regional charities. All older participants were given £20 to cover travel expenses to and from the university.

Participants were screened for age and computer experience—they were required to have used a computer prior to taking part in the study. Additionally, participants were screened for experience using PINs.

4.2.3. MATERIALS

The materials of the study were designed to reproduce a real life authentication system and the experience of logging into multiple accounts. Six account names were created for the study. The names used were alarm, credit card, debit card, library, telephone and television. Each account was assigned a four-digit code that participants would be required to learn and remember over the course of the study (see Table 4.1). The six PIN codes were generated randomly using a random number generator and were randomly assigned to an account.

Table 4.1: Accounts and codes used throughout the study.

Account	Code
Alarm	1929
Credit Card	8360
Debit Card	2040
Library	3126
Phone	9984
TV	5088

A PIN system was mocked up on the computer using Experiment Builder v1.5.201. Initially, the program displayed a set of simplified instructions for the participants along with an initial image depicting the account (e.g. picture of a TV for the TV account). Participants were required to press the space bar to start the experiment. After entering the four digits comprising the PIN code on the computer number pad, the system informed participants whether they had entered the code correctly or incorrectly, no feedback on individual digits was given. The system carried out the cycle four more iterations, meaning participants entered their codes five times in total.

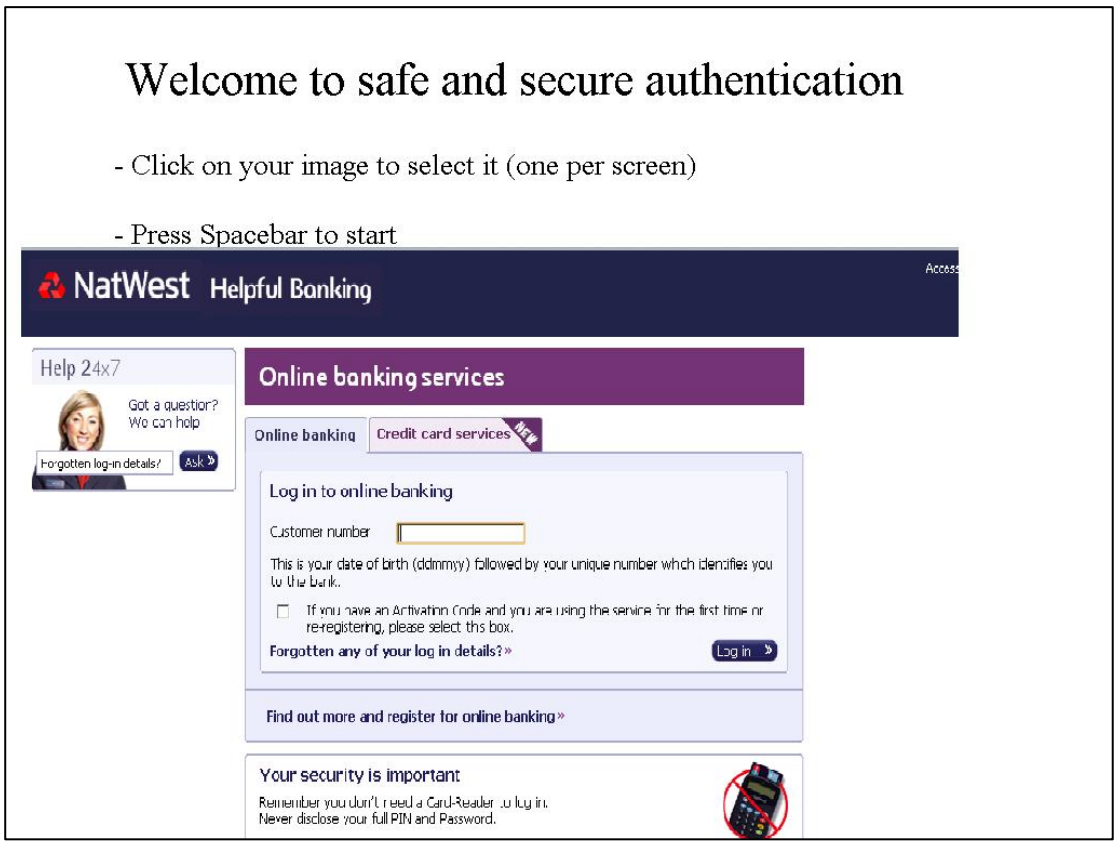


Figure 4.1: Simplified instructions for participants along with image of account.

4.2.4. PROCEDURE

The procedure for this study consisted of two stages: the enrolment and the authentication stage. During enrolment, participants learned their PINs. During authentication, participants attempted to access their 'accounts' by entering the correct PIN for each account.

4.2.4.1. ENROLMENT STAGE

During the enrolment stage participants were given codes to learn in turn and were allowed 60 seconds to learn them using the method of their choice. Participants were allowed to use their own learning method as previous research (Rasmussen & Rudmin, 2010) suggested that older adults are not as successful in remembering PINs if they are required to use a specific learning method.

The PIN codes were given to participants on a sheet of paper which they were allowed to look at for the duration of the 60 seconds. Before handing over the sheet of paper with the PIN, participants were told orally what account the PIN belonged to – "the PIN that you are about to be given will be for your [account name] account". After the allocated time to learn the code, the participants were taken through a mock authentication attempt to make sure they had learned the PIN and to serve as practice. The mock authentication attempt required participants to enter their four digits correctly five times. If participants failed to enter their PIN correctly in at least three consecutive attempts they were required to perform the practice trials once again. They were also given the option of being shown the PIN code again.

4.2.4.2. AUTHENTICATION STAGE

During the authentication stage, participants were required to enter their digits correctly five times, regardless of whether it was correct or incorrect. At first they were presented with simplified instructions (see Figure 4.1) and they were given the full instructions orally. Once they pressed the spacebar to start they were shown an image of the account and were required to enter the four digits using the computer's number pad. After the selection of four digits, participants were told whether they selected correctly or incorrectly. If they selected incorrectly they were not told which one(s) they selected wrong. Participants were then required to enter their codes four more times.

4.2.4.3. PROCEDURE FOR PARTICIPANTS

Participants were asked to attend the lab on three separate occasions (see Table 4.2 for procedure overview). During the first session, participants were given their first set of codes (SET 1) consisting of either two or three PINs depending on the condition (low or high load). The order of the accounts was randomised to eliminate any order effects. Participants were taken through the enrolment stage with the first account and were then taken through the enrolment stage once more with their second account. This was done a third time for the high load condition. After enrolling with the systems, participants were distracted from the encoding task by taking part in a short discussion with the investigator. The discussion focused on their experience of current authentication systems such as passwords and smartcards and lasted approximately 10 minutes.

Following the discussion, participants were taken through the authentication stage with the order of the accounts randomised.

Table 4.2: Overview of procedure.

Session	Activity
1	1. Enrolment with set 1 2. Discussion (distractor) 3. Authentication with set 1
2 (+1 week)	1. Authentication with set 1 2. Enrolment with set 2 3. Discussion (distractor) 4. Authentication with set 2
3 (+1 week)	• Authentication with set 2 • Authentication with set 1 3. Discussion

Participants were asked to return to the lab a week after the first session. Upon their arrival they were greeted and were asked to once again authenticate using the codes they were assigned in the first session in the first week (SET 1) and once again the order of the accounts were randomised. Once participants finished authenticating, they were enrolled with the remaining accounts (SET 2). Following the enrolment with their second set, participants were asked to describe a series of images for approximately 10 minutes. Upon the completion of the description task, participants were asked to authenticate with the new set of codes (SET 2). They were not asked to authenticate with their first set of codes.

Participants visited the lab for a final time one week after the second session. Upon their arrival they were greeted and were asked to authenticate with the codes they were assigned in the previous two sessions. The order of the sets was randomised, as well as the order of the accounts in each set. Once they had authenticated with both sets of codes, participants were asked questions about their experience and about their strategies for remembering the PINs.

The total amount of time taken to complete the study when taking into account the three sessions was approximately 95 minutes for older participants and 70 minutes for the younger participants.

4.3. RESULTS

A 3-way ANOVA with repeated measures on one factor (Week) and two independent factors – age and load – was carried out on both Set 1 and Set 2. Variables measured were the number of successful attempts and the average time taken to enter the codes. For a table of means see Appendix B.

The average number of successful attempts measured the number of times participants entered all four digits correctly for a code – to a maximum of five times (per account) – averaged across the total number of accounts per week. The average time to authenticate, recorded in seconds, measured the average duration of an attempt, described as the time needed to enter the four digits constituting a code – from the press of the space bar to the entry of the final digit.

4.3.1. SUCCESSFUL ATTEMPTS – ACCURACY

4.3.1.1. SET 1 – ORIGINAL CODES

Participants' scores (max=5) for SET 1 codes were collated for each of the three weeks. A 2 (participant age: young, old) x2 (cognitive load: high, low) x3 (week of testing: week 1, 2, 3) mixed factorial ANOVA was carried out.

For SET 1, a main effect of age was present ($F(1,32)=4.759$, $p<.05$), with younger participants (mean: 3.96) being significantly more accurate than the older participants (mean: 2.61). No main effect of load was present ($F(1,32)=1.405$, $p>.05$) and no main effect of week was found ($F(2,31)=0.999$, $p>.05$).

No interaction effects were found for the first set of codes. There was no two-way interaction between age and load ($F(1,32)=0.151$, $p>.05$), age and week ($F(2,31)=1.553$,

p>.05), or load and week (F(2,31)=0.285, p>.05). Additionally, no three-way interaction was found between age, load and week (F(2,31)=0.942, p>.05).

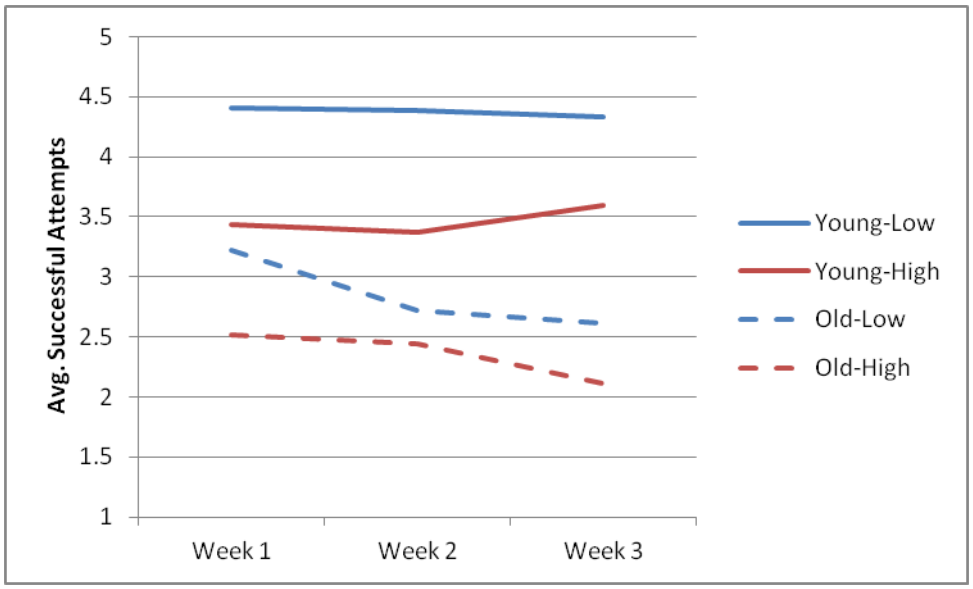


Figure 4.2: Overview of Average Successful Attempts in SET 1.

4.3.1.2. SET 2 – NEW CODES

Participants' scores (max=5) for SET 2 codes were collated for each of the two weeks. A 2 (participant age: young, old) x2 (cognitive load: high, low) x2 (week of testing: week 2, 3) mixed factorial ANOVA was carried out.

For SET 2, a main effect of age was present (F(1,32)=6.150, p<.05) where younger participants (mean: 3.88) were more accurate than the older participants (mean: 2.51) when recalling the SET 2 PINs over the course of three weeks. No main effect of load was found (F(1,32)=0.068, p>.05). A main effect of week was found (F(1,32)=5.681, p<.05), with pairwise comparisons showing that participants were significantly more accurate during the second week (mean: 3.48) than during the third week (mean: 2.91) (p<.05).

No interaction effects were found for the second set of codes. There was no two-way interaction between age and load (F(1,32)=2.970, p>.05), age and week (F(1,32)=1.055, p>.05), or load and week (F(1,32)=1.220, p>.05). Additionally, no three-way interaction was found between age, load and week (F(1,32)=0.235, p>.05).

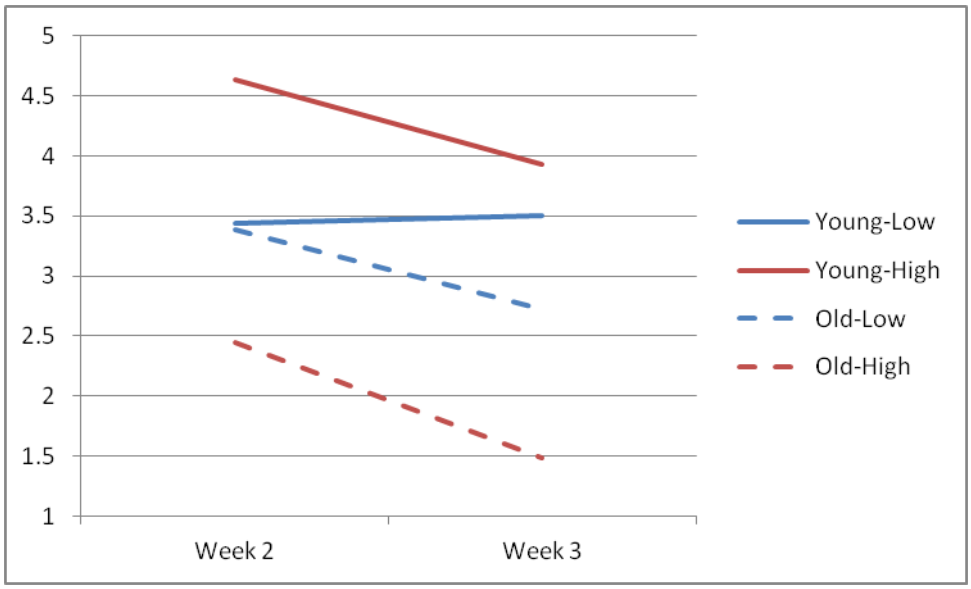


Figure 4.3: Overview of Average Successful Attempts in SET 2.

4.3.1.3. OVERALL ACCURACY

Overall the results show that for both sets of PINs, a main effect of age was present showing that younger participants were significantly more accurate than older participants when remembering PINs. The number of PINs asked to learn and remember did not affect the performance of either age group in terms of accuracy. The most interesting finding is the lack of an age by week interaction in either code set. The lack of this interaction seems to suggest that the younger and older group were equally affected by the task of remembering multiple PIN codes whereas previous research suggests older adults would be more affected. Note, however, that a main effect of week (i.e. performance decrement over time) was *only* shown for the second set of codes learned. Figure 4.4 below illustrates the overall performance of younger and older participants with SET 1 and SET 2 codes.

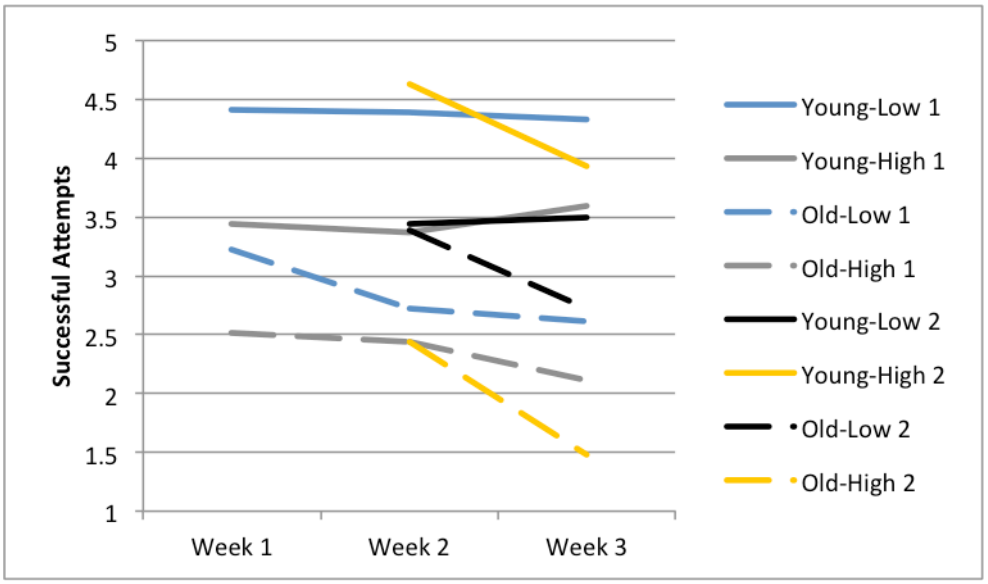


Figure 4.4: Overview of accuracy for all PIN codes.

Examination of the SET 1 results highlights a performance benchmark of 2.6 average successful attempts for older adults remembering a low load (i.e. four codes) and a benchmark of 2.1 average successful attempts when remembering a high load (i.e. six codes). For SET 2, the benchmark for the low load condition was 2.7 average successful attempts while the performance benchmark for a high load was 1.5.

4.3.2. Average Time - Speed

4.3.2.1. Set 1 – Original Codes

Participants' average time taken (in seconds) to enter the digits making up the codes for SET 1 were collated for each of the three weeks. A 2 (participant age: young, old) x2 (cognitive load: high, low) x3 (week of testing: week 1, 2, 3) mixed factorial ANOVA was carried out.

For SET 1 codes, a main effect of age was present ($F(1,32)=40.918$, $p<.001$), with younger participants (mean: 2.10 seconds) entering their codes significantly faster than the older participants (mean: 4.82 seconds). No main effect of load was present ($F(1,32)=0.260$, $p>.05$) and no main effect of week was found ($F(2,31)=1.551$, $p>.05$).

No interaction effects were found for the first set of codes. There was no two-way interaction between age and load ($F(1,32)=0.094$, $p>.05$), age and week ($F(2,31)=0.531$,

p>.05), or load and week (F(2,31)=0.134, p>.05). Additionally, no three-way interaction was found between age, load and week (F(2,31)=2.769, p>.05).

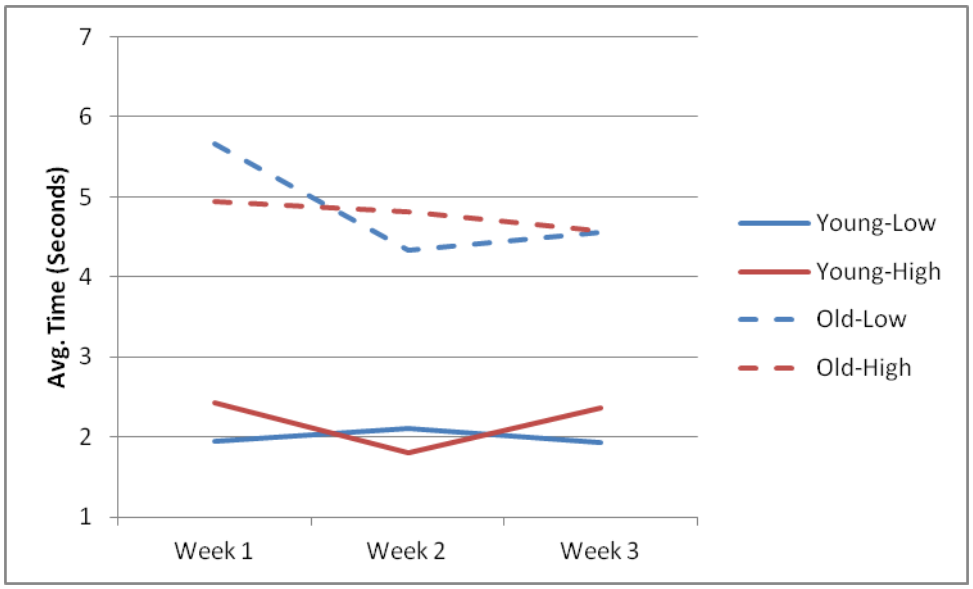


Figure 4.5: Overview of average time taken to enter PINs in SET 1.

4.3.2.2. SET 2 – NEW CODES

Participants' average time taken (in seconds) to enter the digits making up the codes for SET 2 were collated for each of the two weeks. A 2 (participant age: young, old) x2 (cognitive load: high, low) x2 (week of testing: week 2, 3) mixed factorial ANOVA was carried out.

For SET 2, a main effect of age was present (F(1,32)=51.181, p<.001), with younger participants (mean: 2.14 seconds) entering their codes significantly faster than the older participants (mean: 5.21 seconds). No main effect of load was present (F(1,32)=0.321, p>.05) and no main effect of week was found (F(1,32)=0.217, p>.05).

No interaction effects were found for the second set of codes. There was no two-way interaction between age and load (F(1,32)=2.813, p>.05), age and week (F(1,32)=0.860, p>.05), or load and week (F(1,32)=1.146, p>.05). Additionally, no three-way interaction was found between age, load and week (F(1,32)=0.082, p>.05).

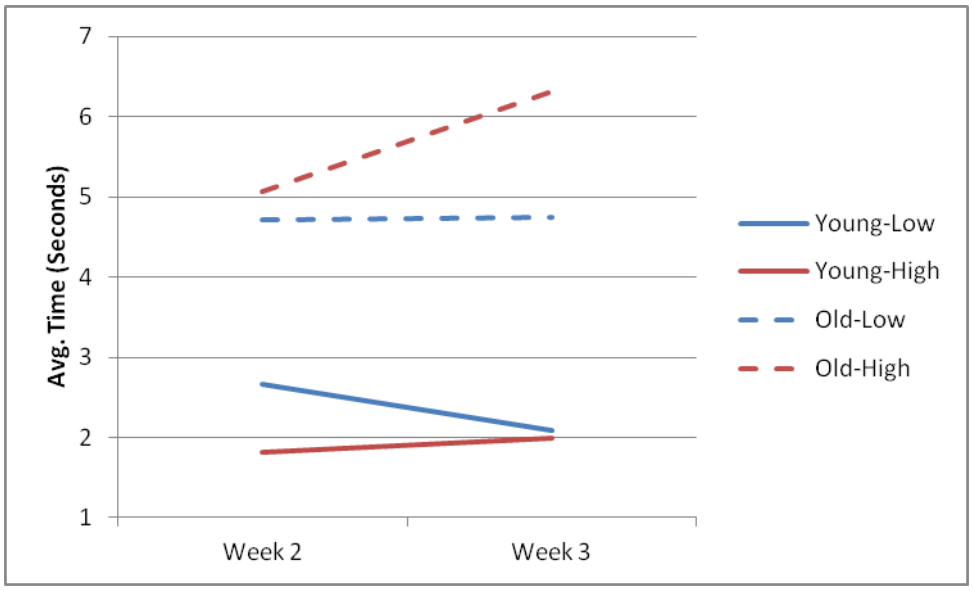


Figure 4.6: Overview of average time taken to enter PINs in SET 2.

In summary, no evidence of a speed-accuracy trade-off was found with the speed data showing the same overall pattern as the accuracy data.

4.3.3. Order of Acquisition

Forgotten codes were further explored – i.e. those codes that participants were totally unable to recall after 5 attempts – as a function of order of acquisition. The underlying question was whether the order of acquisition of the code was reflected in the rate of forgetting. This was mapped out as a function of load.

4.3.3.1. Low Load

In the low load condition the effect of age emerges clearly irrespective of order of acquisition – i.e. older participants were consistently more likely to forget PINs – although there was no sense that memorability was more fragile with those PINs acquired later (see Figure 4.7).

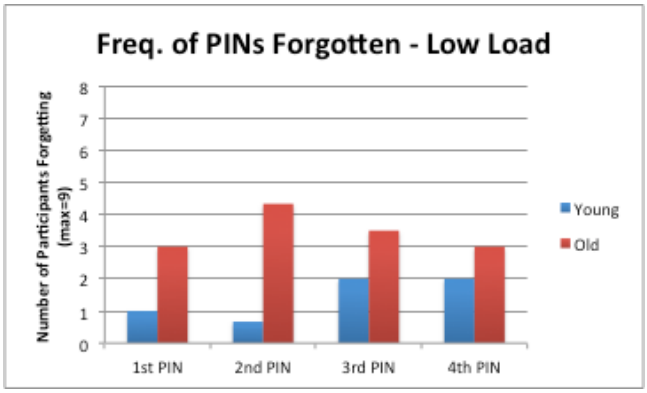


Figure 4.7: Frequency of forgetting PINs for younger and older participants – Low Load condition.

A chi square test using participant age and codes forgotten as factors found a significant association between participant age and forgetting of codes was found when remembering four codes, χ^2 (1)=13.750, p<.001. This seems to represent the fact that older participants were more likely to forget codes than younger participants. Separate chi square tests were run on the younger and older participants with order of acquisition and forgetting as factors and both results found that order of acquisition was not associated with the forgetting of codes.

4.3.3.2. HIGH LOAD

In the high load condition there was also an effect of age across all PINS – i.e. older participants were consistently more likely to forget PINs – although there was a sense that they started to struggle with those PINs acquired later – i.e. they showed particularly poor performance with 5th and 6th PINs (see Figure 4.8).

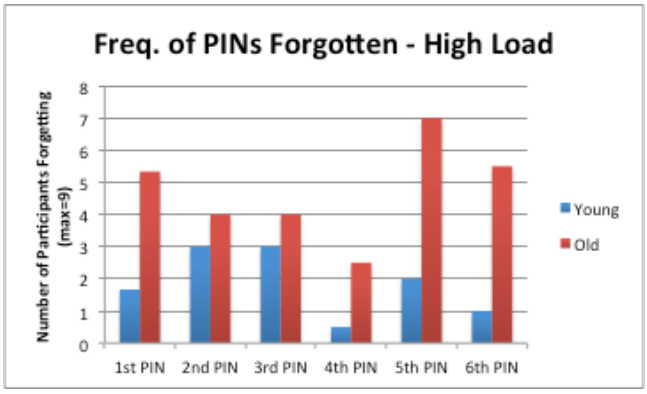


Figure 4.8: Frequency of forgetting for PINs for younger and older participants – High Load condition.

A chi square test found a significant association between participant age and forgetting of codes was found when remembering six codes, χ^2(1)=25.412, p<.001. This seems to represent the fact that older participants were more likely to forget codes than younger participants. Separate chi square tests were run on the younger and older participants with order of acquisition and forgetting. The test for younger participants found that order of acquisition was not associated with the forgetting of codes, however, the test for older participants revealed a significant association between order of acquisition and forgetting of codes, χ^2(5)=11.423, p<.05. This seems to suggest that the fifth and sixth codes were more challenging to retain, with 77.8% and 61.1% of older participants forgetting the codes respectively.

4.4. DISCUSSION

It was predicted that younger participants would outperform older participants in terms of accuracy and time with multiple PIN codes. As expected, the younger group performed significantly better at recalling the PINs than the older group for both sets of codes therefore the prediction was supported by the results: for both SET 1 and SET 2 codes older adults were less accurate and slower to authenticate. The predictions were based on self-reported studies (e.g. Rasmussen & Rudmin, 2010; Vines et al., 2011) and were supported by this experimental study. In the past cued-recall and free recall have been shown to require more cognitive resources than recognition (Baddeley, 1997) and it has been shown that older adults, who do not have as many resources available, will be disadvantaged when asked to recall information (Craik & McDowd, 1987). Most knowledge-based authentication systems, including PINs, rely on free recall – or arguable cued- recall if the code was associated with the account – and thus older adults are expected to be disadvantaged by existing KBA systems.

It is well known that age negatively affects the performance of serial order recall, and research has suggested that older adults are more likely to adhere to a learned sequence, even if incorrect, than younger adults (Kay, 1951). Simply, older participants are unable to learn from their mistakes and instead learn their errors. Maylor et al. (1999) found that older adults are more prone than younger adults to suffer from intrusion errors (forgetting one or more items) and movement errors (remembering the correct digits but in the incorrect order) in serial recall both in terms of volume and proportion, supporting the fact that older adults are consistent with their selections even when they are incorrect. Remembering a PIN code is a form of serial number recall and as predicted older adults' performance in terms of both accuracy and time was poorer than that of younger adults.

Hasher and Zacks (1988) contend that older adults can be more distracted by irrelevant information during the encoding phase and this can affect later performance either by not learning the information in enough detail or learning the unrelated information. This would suggest that older adults must be particularly careful when learning new PINs in real life to ensure they have distraction free time to encode the new PIN. However, having a distraction-free environment in the real world is very difficult to control. In this study it was possible to control for this issue by giving participants quiet time

dedicated to learning the codes yet the results show that older participants were still affected by task. Hence, it could be possible that in the real world the problem of distractions plays a bigger role in the performance decrement.

Naveh-Benjamin (2000) proposed an Associative-Deficit Hypothesis (ADH) where older adults show significant impairment when having to associate multiple items. Part of impairment is explained by using poor strategic behaviour when creating associations, an observation later supported by Naveh-Benjamin et al. (2007). This deficit in associative memory likely plays a role in the binding of the digits together into a code, and in the binding of the code to the account. In the case of PINs, the binding issues could be twofold: binding the individual digits into a code and binding a code to a specific account. From the results it is not possible to determine which, if either, was the cause for the performance decrement. However, it is important to note once again that the older group are being disadvantaged from the outset and it was once again shown to be true by the experimental study.

It was also predicted that participants in the Low Load condition would have more successful attempts at recalling the codes than participants in the High Load condition. This prediction was partly supported by the results, with the main analysis finding no significant difference in accuracy with the two loads but a subsequent chi square analysis suggesting possible load problems with codes acquired last. While this finding initially appears to be confusing, upon further inspection it can be seen that, at least for the older group, accuracy of recall was poor. This means that although the additional two codes did not increase the memorability problem for the older group, the memorability was not good to start with. In other words, a performance decrement was already present with the low load. It is possible that for the younger group the added two codes were not enough to induce a decrement and it would be interesting to establish when this breaking point occurs. Similarly, it would be interesting to establish at what point the recall accuracy for older adults significantly dips again.

Finally, it was predicted that the performance of both age groups – in terms of accuracy and speed – would decline over the course of the three weeks. This prediction was based on previous memory studies which establish that long-term memory for items is sensitive to time (e.g. Verhaeghen et al., 1993). This prediction was partially met, with the accuracy of participants significantly declining with the second set of codes. This

effect was not present with the first set of codes, indicating that participants were able to maintain the accuracy of their code recall for their SET 1 codes. This is not entirely surprising as in reality it should be possible for people to remember two or three four-digit numbers over the course of a week without many problems. However, the accuracy of SET 2 codes was affected – as shown by the main effect of week – suggesting that newer codes are more difficult to retain over a period of one week than original codes.

With regards to the time taken to recall the PINs, younger participants were significantly faster than the older participants as expected and showed no evidence of a speed-accuracy trade-off. This was not a surprise as younger participants were expected to be faster than the older group.

4.4.1. IMPLICATIONS

The implications from the theoretical literature and the empirical findings are discouraging for older adults. Theory implied older adults should perform worse than younger adults when remembering multiple PINs and in practice that was the case. Despite this, PINs are still one of the most common authentication systems.

A main effect of week was found with the SET 2 codes. This effect shows that accuracy of recall is significantly impaired after an extended delay of one week when new PINs are learned on top of existing ones. This was the case when learning a second set of codes and it remains to be seen what the effect is when a third or fourth set is added. This finding demonstrating a decline in accuracy over time with the addition of further codes to an extent validates the common insecure practices of reusing and writing down codes (e.g. Grawemeyer & Johnson, 2011) and indicates that providers must do more to aid users if they do not wish for them to circumvent the security measures.

Although the absence of a statistically significant age-specific performance decrement over time appears to suggest an encouraging result, the overall performance of the older group is far from acceptable. The performance benchmark for the older group with a low load was 2.6 average successful attempts for SET 1 and 2.7 average successful attempts for SET 2. In other words older adults generally would authenticate on their fourth attempt – meaning that the possibility of being locked out of their accounts that

implement the traditional ‘three strikes’ policy becomes a real threat. Perhaps a solution could be to amend the popular ‘three strikes and you are out’ rule that is regularly implemented with PINs and passwords in favour of a higher limit as proposed by Brostoff and Sasse (2003), although this approach would leave accounts vulnerable to attacks.

It should be noted that participants were assigned the codes in this study and they were not able to change those codes. In the real world users are allowed to customise their PINs which allows them to select combinations of digits that may be more memorable than a randomly generated PIN – a generation effect (Slamecka & Graf, 1978). It is understandable that users may choose to re-use their PINs if they are able to choose them – given the fragile nature of new PINs stored in memory on top of a set of old PINs. However, the implications of allowing users to choose their own codes are weak PINs that can be easily guessed (e.g. Bonneau et al., 2012) or the reuse of existing codes (Ives et al., 2004). It should be made clear that the two behaviours are not mutually exclusive and that it is possible – even probable – that users reuse existing weak codes. The reason for assigning the participants with randomly generated codes was to prevent participants from choosing PINs they currently used as this could have artificially aided the learning and recall of the codes. A secondary reason was to evaluate participants in the context of a secure environment where PINs are unique and randomly generated. While this might not reflect the real state of PIN usage, it represents a view of security that will be upheld for the following studies in this thesis (i.e. memorability results where the security of the system is not compromised).

4.4.2. Future Work

Based on the results from this study highlighting the problems older adults experienced remembering multiple PIN codes, it is imperative to think about what other authentication methods can be tested with older users in order to improve their performance. The key is to find a system that is inclusive of older adults, rather than a system aimed solely at older adults. A system aimed exclusively at improving the performance of older adults risks being rejected by the older adult community for singling them out and being rejected by the younger adult community for potentially penalising younger adults.

Graphical authentication systems may hold the key to inclusive KB authentication. Much of the work regarding these systems is based on the premise that the picture superiority effect (e.g. Standing, 1973) and recognition (e.g. Baddeley, 1997) make the systems more memorable than traditional text-based systems that rely on verbal recall. Older adults have been shown to take advantage of the picture superiority effect at least to a similar extent as younger adults (e.g. Winograd & Smith, 1982) so it would be interesting to evaluate their performance and that of younger adults over the course of multiple weeks with multiple graphical codes to see if the relative gap in performance is actually reduced and/or the absolute performance of older adults is improved.

4.5. CHAPTER SUMMARY

This chapter set out to investigate how older adults performed using Personal Identification Numbers (PINs) and to establish a benchmark performance for future authentication systems. It was thought that younger participants would outperform the older participants in terms of successful PIN selections. It was also predicted that participants with a lower PIN load would outperform participants with a higher PIN load. As predicted younger participants outperformed older participants in accuracy and time. A performance decrement was observed over the course of a one week delay, although this decrement was only present for the second set of PIN codes (the new numbers) Finally, there was no difference in performance between participants with a high PIN load and participants with a low PIN load.

5. A New Approach to Authentication with Pictures: Tiles Graphical Authentication System

5.1. Rationale

PIN-based authentication methods are not well suited for older users as demonstrated in Chapter 4. This finding presents a significant problem due to the ubiquity of PIN-based authentication methods and is likely to extend to other KBA such as passwords. A possible alternative approach to authentication is using GAS.

Graphical authentication systems have been extensively tested on younger adults, but we do not know much about their performance with a population of older adults. Previous research suggests that older users may not be penalised as harshly with graphical systems compared to alphanumeric passwords, as recognition has been shown to be less affected by ageing than recall (Brown & Park, 2003; Craik & McDowd, 1987), most likely due to the extra effort and resources that are required for pure recall (Raaijmakers & Schiffrin, 1992). Additionally, previous research also shows that visual memory appears to be less affected by ageing than knowledge-based recall (Brown & Park, 2003), meaning that memory for pictures is likely to be superior than memory for words, and therefore remembering a graphical combination would be less work than remembering an alphanumeric password. Older adults have been shown to benefit from the Picture Superiority Effect (Park et al., 1986; Winograd et al., 1982) to the same extent as younger adults. Moreover, older adults have been shown to be able to remember images after a one-week delay without a significant drop in recognition performance, although a longer delay appears to negatively impact their performance (Park et al., 1988).

When taking all the literature into account, it is evident that GAS should be tested with older adults to determine whether their performance can be improved from the PIN benchmark. Recognition-based systems would be expected to yield better performance than cued-recall systems due to the reduced need for cognitive resources (e.g. Baddeley, 1997). However, recognition-based GAS have the potential to tax cognitive resources in a different way: item load. Users are required to remember multiple items for security purposes as all items are displayed on the screen. This means that for every account a

user has they need to remember at least four items, which can quickly add up if there are multiple accounts to authenticate for. This can be problematic for older adults specifically because of their reduced cognitive resources (e.g. Fisk et al., 2004). Additionally, they are known to experience item binding issues which could complicate the recognition of the codes further – i.e. remembering two features together despite not necessarily struggling with either feature individually (Chalfonte et al., 1996). This item binding problem can be potentially solved with graphical authentication where users are not required to remember four images that form a code, and also remember what account the code is associated with. The Associative-Deficit Disorder (Naveh-Benjamin, 2003) further predicts that older adults will experience more problems when remembering multiple images together.

In order to minimise the item load associated with recognition-based GAS while also taking advantage of the visual aspect, a new GAS was developed. The Tiles system was designed with the aim of facilitating the binding of the codes and reducing the cognitive load of having to remember multiple items per account. These issues were addressed by having participants remember a single image for authentication instead of the traditional four or five (e.g. Valentine, 1998; Dhamija & Perrig, 2000; De Angeli et al., 2002). By cutting down on the number of images that users need to remember per account it is expected that the acquisition of further codes will be easier. For example, remembering four Tiles codes would be equivalent to remembering one Déjà vu (De Angeli et al., 2002) – users would need to remember four images in total. The advantage is clear: while both scenarios require the same cognitive resources, with Tiles the user is able to authenticate with five separate accounts while the Déjà vu user can only authenticate with a single account.

The binding of the code is also facilitated by the Tiles system. Traditional GAS require users to remember four or five different items per account. The user then needs to bind the four items with each other and associate the bound items with their respective account. Tiles addresses the initial item binding problem by adding context to the individual items. The user is required to learn one image per account. That one image is then divided into nine segments. The user needs to identify the correct segment that belongs to his/her image from amongst other image segments. It is expected that the identification of the segments will be aided by the user's knowledge of the overall image. In essence, the user has to associate the image with the account and then identify

the segments that belong to that image during authentication, rather than having to remember four independent images for the account.

The nature of the system means that cognitive load should decrease – with participants only having to remember one image per account rather than four. Tiles aims to address the binding problem by providing context to the individual 'images' (or segments) as all images are linked together as being part of the whole. Park et al. (1990) found that context aids older adults significantly, even if not entirely related. A possible problem with this approach is that some of the segments might result in abstract images which can be problematic for older adults (e.g. Smith et al., 1990), but it is hoped that the addition of context will override this problem.

The purpose of this study was to evaluate the proposed GAS, Tiles, in the context of younger and older adults. This was done by requiring both groups of participants to learn and remember multiple Tiles codes over the course of three weeks. Two different grid types – similar foils and dissimilar foils – were tested with the aim of potentially improving security: while dissimilar images on a grid should make the task easier for the user (i.e. faster to discard incorrect images) it will also make it easier for an attacker to guess the code based on an observation of the authentication attempt. On the other hand, similar images on a grid could increase the difficulty of the task for the user, but will also protect the user from attackers. This security-usability trade-off will be investigated by the inclusion of the two grid types.

It is expected that younger adults will be more accurate than older adults when remembering multiple Tiles codes. It is expected that the gap in performance will be reduced when compared with the PIN benchmark. It is also expected that participants will be more accurate when selecting the segments from dissimilar grids when compared with similar grids as dissimilar grids should reduce the search load and reduce speeds while increasing accuracy (Fisk & Rogers, 1991); It is also predicted that the speed will generally follow the accuracy data (i.e. no speed-accuracy trade-off).

5.2. METHOD

5.2.1. EXPERIMENTAL DESIGN

The study consisted of two factorial designs as the codes were separated into 'original' accounts (SET 1) assigned to participants during the first week and tested in weeks 1, 2 and 3, and 'new' accounts (SET 2) assigned to participants during the second week and tested in weeks 2 and 3. The first set of images consisted of a 2 (age: young, old) x2 (grid type: similar foils, dissimilar foils) x3 (time of testing: week 1, 2, 3) factorial mixed design. The second set of images consisted of a 2 (age: young, old) x2 (grid type: similar foils, dissimilar foils) x2 (time of testing: week 2, 3) factorial mixed design. The factors comprised of one independent – the participants' age (young group, old group) – and two repeated – the grid configuration tested (similar foils, dissimilar foils) and the time period (weeks 1, 2, 3).

Dependent measures comprised the number of successful authentication attempts (maximum of 5 per account) and the average time (in seconds) taken to authenticate.

5.2.2. PARTICIPANTS

36 participants were recruited to fit into one of the two age groups, the young group (18-30 years old, n=18) or the old group (65-75 years old, n=18).

Younger participants (mean age: 19, SD: 1.44) were recruited from the student population in the university using an online participation pool maintained by the university. Given the diversity of the student population this sample was considered adequate.

Older participants (mean age: 71, SD: 3.51), with a mean age of 71 years (SD: 3.51), were recruited using the lab's participant database as well as through an advert on the Elders' Council of Newcastle newsletter. All participants were given £30 to cover travel expenses to and from the university.

Participants were screened for age and for computer experience—they were required to have used a computer prior to taking part in the study. Participants were also screened for adequate vision.

5.2.3. MATERIALS

The materials of the study were designed to reproduce a real life authentication system and the experience of logging into multiple accounts. Four account names were created for the study. The names used were Bank, NHS, Shop, and Email. Each account was allocated a target image that the participants would be required to learn. They would then select segments from that target image in order to authenticate over the course of the study.

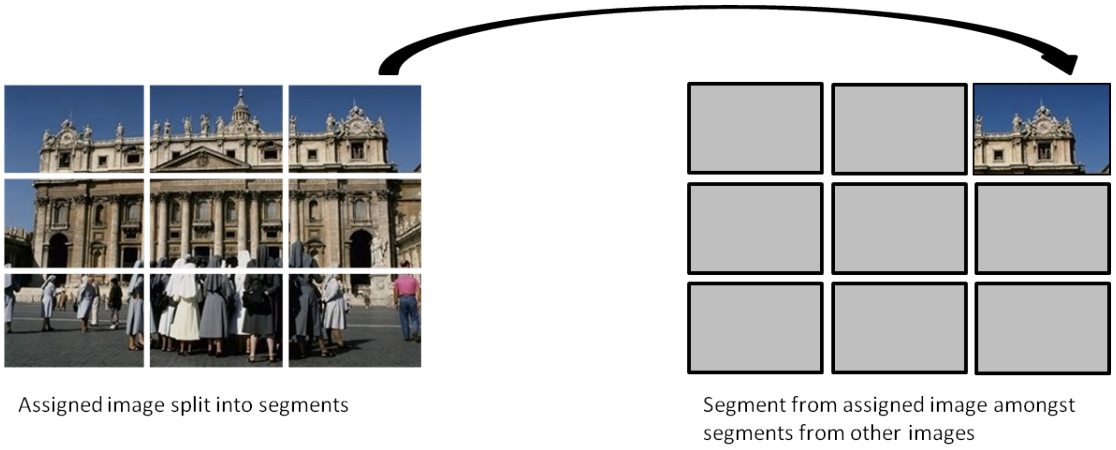


Figure 5.1: Demonstration of the Tiles system. Participants have to select their segment amongst foils in 4 individual challenges.

The main material used was the graphical authentication system, Tiles, that was built for this study (see Figure 5.1). The system was built using Experiment Builder v1.5.201. All participants were tested in the same room and same computer to guarantee the same experience. Two grid compositions were tested: similar grids and dissimilar grids. The aim of testing these two configurations was to determine whether the use of segments that in theory should be more easily dismissed as incorrect could improve the accuracy of participants. Fisk and Rodgers (1991) found that the speed of a visual search can be significantly decreased and the accuracy increased when the search space is reduced. In other words, the lower the number of items the participant has to scan through the faster and the better they can perform the task – both younger and older adults. Based on these findings, participants – both young and old – should benefit from the dissimilar grids as

they should be able to write off a number of segments at first glance and therefore reduce the search space.

In this version of the system, the segments remained in their original position as on the base image – i.e. the upper right segment of the base image was presented on the top right position on the challenge grid (see Figure 5.1). This was done to aid younger and older adults to quickly discriminate foil segments when the target segment is not immediately obvious. For example, if the base image consists of a building with a door on the bottom left and the top right segment on a challenge grid depicts a door, the user can safely exclude that segment from further consideration. Additionally, the fixed location of the target segments makes the familiarisation process a constant mapping (CM) task. Constant mapping instructions have been shown to lead to improved performance in both younger and older adults over varied mapping (VM) instructions (e.g. Fisk & Rodgers, 1991) and therefore a further benefit for using fixed segments.

Initially, the program displayed a set of simplified instructions for the participants and they were required to press the spacebar to start. Four sequential challenge grids were displayed upon the participants' selection. Once four selections were made using the mouse, the system informed participants whether they had selected correctly or incorrectly, no feedback on individual selections was given. The system carried out the cycle four more iterations, meaning participants selected their codes five times in total.

5.2.3.1. GRID COMPOSITION

A collection of 1000 images was obtained from a publicly available image database, purposed for image processing operations (http://wang.ist.psu.edu/docs/related.shtml). The database consisted of 10 natural categories consisting of 100 images each. Example categories include: beach, Rome, buses, elephants, and mountains.

The four base images used for the accounts were chosen from two random categories. Initially two categories were randomly chosen from the database, and from each of the two categories one image chosen at random. The two images that were chosen during this process were used as the base images for the study.

Twelve foil images were chosen for each base image based on visual similarity. The similarity condition (e.g. similar or dissimilar foils) for the base images was determined randomly. An algorithm (Dunphy & Olivier, 2012) was used to select the thirteen most visually similar images from the database (within the base images' categories) or the thirteen least visually similar images – again within the base images' categories. The algorithm compared image signatures in the form of 3D image histograms (in the CIELAB colour space) using *Earth Movers Distance* and produced a list ranking images from most similar to least similar. The least similar image from the list was chosen as the other base image, leaving twelve images as foils for each of the four base images.

Figure 5.2: Example Tiles grid.

Sixteen unique grids were created for each of the two categories. In other words, the two base images from the same category shared the same grids – i.e. images A and B from the same category used the same set of sixteen grids, and images C and D from a different category to image A and B shared the other set of sixteen grids. This setup guaranteed the presence of two possible target segments in each grid, although only one was correct depending on the account. A random number generator was used to construct the grids. Sixteen integer sets were generated with each set containing nine integers. Each integer was mapped onto the corresponding foil (foils were numbered 1-12) and the segment of the image was chosen in accordance with the mapping of a

telephone keyboard. For example, in the list [x,x,4,x,x,x,x,x,x] foil image number four would be placed on the top right corner of the grid. As segments remained in their original locations, the top right segment of the foil image (number four) was used (see Appendix A for sample grids).

A digital voice recorder was also used to record the discussion during the first session.

5.2.3.2. IMAGE OVERLAP

An image overlap was introduced where a target segment for one base image could appear as a foil segment for another target segment. In essence, this meant that a participant had to associate the base image with the account. At any given time after the introduction of SET 2 – i.e. all accounts have been assigned – a participant would be presented with a grid consisting of two possible targets. Only one target would be correct, depending on the account they were being asked to authenticate for.

The overlap was introduced in order to test the system in a more ecologically valid configuration. In the real world it will not always be possible to guarantee separate sets of images per provider, so users may be faced with a situation where a target segment appears as a foil segment for an account. It is acknowledged that this design makes the task harder for participants, but an analysis of the segments chosen by the participants showed that only a small number were affected by the image overlap.

5.2.4. PROCEDURE

The procedure for this study consisted of two stages: the enrolment stage and the authentication stage. During enrolment, participants learned their images. During authentication, participants attempted to access their 'accounts' by identifying the four correct account segments.

5.2.4.1. Enrolment Stage

During the enrolment stage participants were given base images to learn and were guided through a familiarisation process consisting of four questions to encourage the

participant to deeply process the image (see Table 5.1). In total, the familiarisation process lasted for approximately five minutes per image. During the first two questions participants were able to view the whole image. The final two questions were asked for each of the nine segments, with only one segment being shown on the screen. Participants were given each image individually and after having 15 seconds to study the image they were asked the respective questions by the experimenter. The participant was required to reply to the questions out loud.

Table 5.1: List of questions used each of the four base images.

Questions for Complete Image
Could you please tell me a story to go with this picture?
Does this picture remind you of any past experience or some event in your life?
Questions for Individual Segments
How does this segment link with the rest of the image?
What stands out the most in that segment?

Once participants had been through the familiarisation process for the base image, they were asked to select the segments that belong to their base image from amongst foil segments to confirm they had learned them, essentially a mock authentication attempt (see 5.2.4.2 below). If the participant failed to select their segments correctly at least three attempts in a row they were asked to repeat the mock authentication. They were also shown the image again if necessary. Participants were then taken through the same process for their next account.

5.2.4.2. AUTHENTICATION STAGE

During the authentication stage participants were presented with a login screen and asked to authenticate. The login screen contained nine segments in a 3 x 3 grid and participants were required to select the segment belonging to their base image from the nine in four successive screens using the mouse. Once the participant selected four segments, the program told them whether they selected correctly or incorrectly—if incorrectly they were not told which images they got wrong. Participants were required

to do this for a total of five times per set regardless of whether they selected correctly or incorrectly.

5.2.4.3. PROCEDURE FOR PARTICIPANTS

Participants were asked to attend the lab on three separate occasions (see Table 5.2 for procedure overview). During the first session, participants were given their first set of images consisting of one similar grid account and one dissimilar grid account. The order of the accounts was randomised in order to eliminate any order effects. Participants were taken through the enrolment stage with the first account and were then taken through the enrolment stage once more with their second account. After enrolling with both accounts, participants were distracted from the encoding task by taking part in a short discussion with the investigator. This discussion focused on their experience of current authentication systems such as passwords and smartcards and lasted approximately 10 minutes. The discussion was recorded for later analysis.

Following the discussion, participants taken through the authentication stage with the order of the accounts randomised, meaning that it did not necessarily follow the same order as the enrolment.

Table 5.2: Overview of procedure.

Session	**Activity**
1	1. Enrolment with SET 1 2. Discussion (distractor) 3. Authentication with SET 1
2 (+1 week)	1. Authentication with SET 1 2. Enrolment with SET 2 3. Discussion (distractor) 4. Authentication with SET 2
3 (+2 weeks)	• Authentication with SET 2 • Authentication with SET 1 3. Discussion

Participants were asked to return to the lab a week after the first session. Upon their arrival they were greeted and were asked to once again authenticate using the base images they were assigned in the first session a week ago (SET 1) and once again the order of the accounts were randomised. Once participants finished authenticating, they were enrolled with two new accounts, again one from each grid type. Following the enrolment with their second set (SET 2), participants were engaged in another discussion, this time regarding their ideal authentication system. This discussion lasted for approximately 10 minutes. Upon the completion of the discussion, participants were asked to authenticate with the new set of images. They were not asked to authenticate with their first set of images.

Participants visited the lab for a final time one week after the second session. Upon their arrival they were greeted and were asked to authenticate with the accounts they were assigned in the previous two sessions. The order of the sets was randomised, as well as the order of the accounts in each set. Once they had authenticated with both sets of images, participants were asked questions about their experience and about their strategies for remembering the images.

The total amount of time taken to complete the study when taking into account the three sessions was approximately 120 minutes for older participants and 80 minutes for the younger participants.

5.3. RESULTS

A 3-way ANOVA with repeated measures on two factors (Grid Type and Delay) and age as an independent factor was carried out on both SET 1 and SET 2. Variables measured were number of successful attempts and average time taken to select the four segments. For a table of means see Appendix B.

The number of successful attempts measured the number of times participants selected all four segments correctly in an attempt, to a maximum of five times (per account). The average time to authenticate, recorded in seconds, measured the average duration of an attempt between the presentation of the first segment and the selection of the fourth segment.

5.3.1. Successful Attempts – Accuracy

5.3.1.1. Set 1 – Original Codes

Participants' scores (max=5) for SET 1 codes were collated for each of the three weeks. A 2 (age: young, old) x2 (grid type: similar foils, dissimilar foils) x3 (week of testing: week 1, 2, 3) mixed factorial ANOVA was carried out.

For SET 1, a main effect of age was found ($F(1,34)=11.485$, $p<.01$) with younger participants (mean: 4.55) achieving more successful attempts than older participants (mean: 3.42). No main effect of grid type was found ($F(2, 33)=.977$, $p>.05$). A main effect of week was present ($F(2,33)=12.042$, $p<.001$). Pairwise comparisons show participants achieved significantly more successful attempts in the first week (mean: 4.74) compared to both the second week (mean: 3.90) ($p<.010$) and the third week (mean: 3.31) ($p<.001$). There was no significant difference in accuracy was present between week 2 and week 3.

No interactions were found between age and grid type ($F(1, 34)=0.040$, $p>.05$), age and week ($F(2,33)=3.168$, $p>.05$), or grid type and week ($F(2, 33)=0.787$, $p>.05$). There was no 3-way interaction between age, grid type, and week ($F(2,33)=0.256$, $p>.05$).

5.3.1.2. Set 2 – New Codes

Participants' scores (max=5) for SET 2 codes were collated for each of the two weeks. A 2 (age: young, old) x2 (grid type: similar foils, dissimilar foils) x2 (week of testing: week 2, 3) mixed factorial ANOVA was carried out.

For the number of successful attempts on the second set of images, a main effect of age was found ($F(1,34)=5.762$, $p<.05$) with younger participants (mean: 3.92) achieving more successful attempts than older participants (mean: 2.94). No main effect of grid type was found ($F(1, 34)=1.156$, $p>.05$). A main effect of week was present ($F(1,34)=41.570$, $p<.001$) with participants achieving more successful attempts in the second week (mean: 4.38) compared the third week (mean: 2.49) ($p<.001$).

No interactions were found between age and grid type ($F(1, 34)=3.541$, $p>.05$), age and week ($F(1,34)=1.295$, $p>.05$), or grid type and week ($F(1, 34)=0.158$, $p>.05$).

There was a 3-way interaction between age, grid type, and week ($F(1,34)=14.739$, $p=.001$). Looking at Figure 5.3, this would seem to reflect the relatively strong performance in week 3 by younger adults presented with dissimilar grids. In other words, the younger group seem able to take advantage of the dissimilar grids while the older group could not.

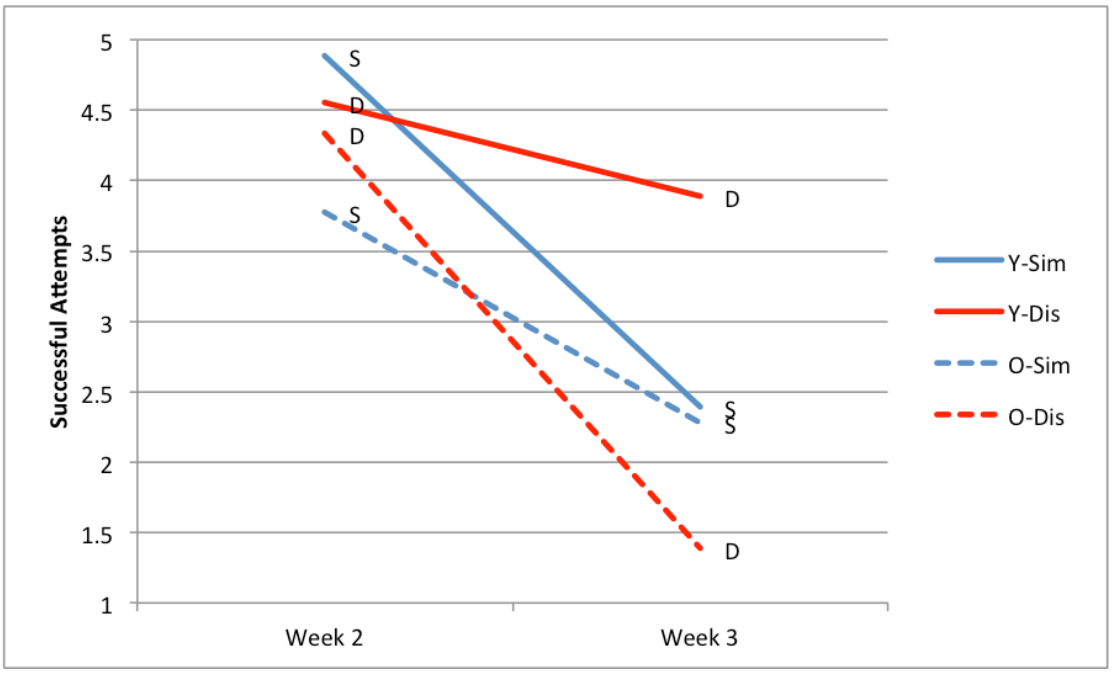


Figure 5.3: Three-way interaction between Age, Grid Type and Week for SET 2.

A one-way ANOVA showed no significant difference in accuracy during week 2 between grids for younger participants ($F(1,35)=1.029$, $p>.05$) while a significant difference in accuracy between grids was found during week 3 ($F(1,35)=4.356$, $p<.05$) where younger participants were more accurate with dissimilar foils (mean: 3.89) than with similar foils (mean: 2.39). Another one-way ANOVA showed no significant difference in accuracy during week 2 or week 3 between grids for older participants (week 2: $F(1,35)=1.496$, $p>.05$; week 3: $F(1,35)=1.440$, $p>.05$). This is supported by subsequent statistical tests, with an independent samples t-tests showing no significant differences in accuracy between similar and dissimilar grids for the older group in week 3 ($t(34)=1.200$, $p>.05$. However, a significant difference in accuracy was observed

between similar and dissimilar grids for the younger group in week 3 ($t(34)=-2.087$, $p<.05$).

5.3.1.3. Overall Accuracy

The study overview (Figure 5.4) suggests that the rate of decay in SET 1 codes from week 1 to week 2 differs from that between week 2 and week 3. The rates of decay between weeks does not appear to follow a fixed pattern, however – memorability appears to decrease less for older participants and for younger participants using dissimilar foils while memorability appears to decrease more for younger participants using similar grids. These trends appear to suggest that additional load does have an effect on memorability, although the effect is variable. A look at SET 2 suggests that accuracy drops off sharply for all but younger participants using dissimilar grids.

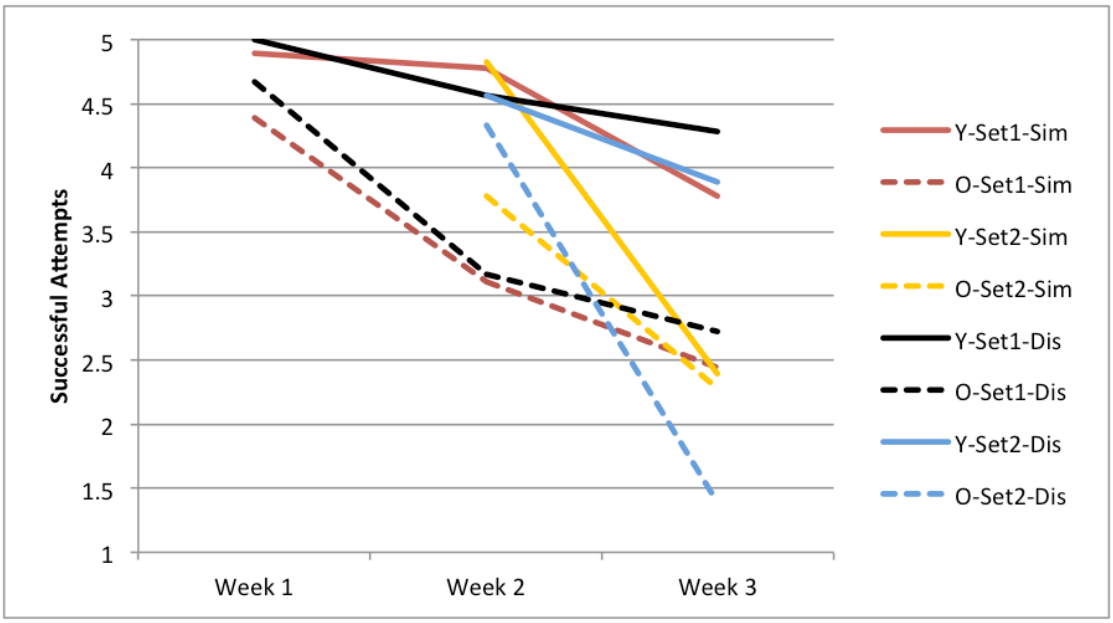


Figure 5.4: Overview of successful attempts for images in both SET 1 and SET 2.

If we consider the performance benchmark established with PIN for the older group of 2.6 average successful attempts for SET 1 and 2.7 average successful attempts for SET 2, the Tiles accuracy results of 2.5/2.7 with SET 1 codes and an accuracy of 1.5/2.4 with SET 2 codes show a similar performance with SET 1 codes but a worse performance with SET 2 codes.

5.3.2. AVERAGE TIME

5.3.2.1. SET 1 – ORIGINAL CODES

For the average time taken to select the segments for SET 1 codes, a main effect of age was found ($F(1,34)=67.773$, $p<.001$) with younger participants (mean: 9.82 seconds) selecting their segments faster than older participants (mean: 20.90 seconds). No main effect of grid type was found ($F(1, 34)=0.100$, $p > .05$). A main effect of week was present ($F(2,33)=16.820$, $p<.001$). Pairwise comparisons show that participants selected their segments significantly faster in the first week (mean: 12.25) compared to both the second week (mean: 15.65) and the third week (mean: 18.18). There was no significant difference in speed between week 2 and week 3. These findings reflect those found earlier (i.e. no sign of a speed-accuracy trade-off).

No interaction was found between age and grid type ($F(1, 34)=0.181$, $p>.05$). An interaction effect between age and week was found ($F(2,33)=4.367$, $p<.05$) (see Figure 5.5). A one-way ANOVA showed a significant difference in time for younger participants ($F(2,107)=4.070$, $p<.05$) with a significant difference between week 1 and week 3, but not between week 1 and week 2 or week 2 and week 3. Another one-way ANOVA showed a significant difference in time for older participants ($F(2,107)=9.332$, $p<.001$) with a significant difference between week 1 and week 3, but not between week 1 and week 2 or week 2 and week 3. It should be noted that the difference between week 2 and week 3 for the older group was borderline non-significant ($p=.053$). Independent samples t-tests show significant age-related differences in time taken to select the segments in weeks 1, 2 and 3. No interaction was found between grid type and week ($F(2, 33)=0.357$, $p>.05$).

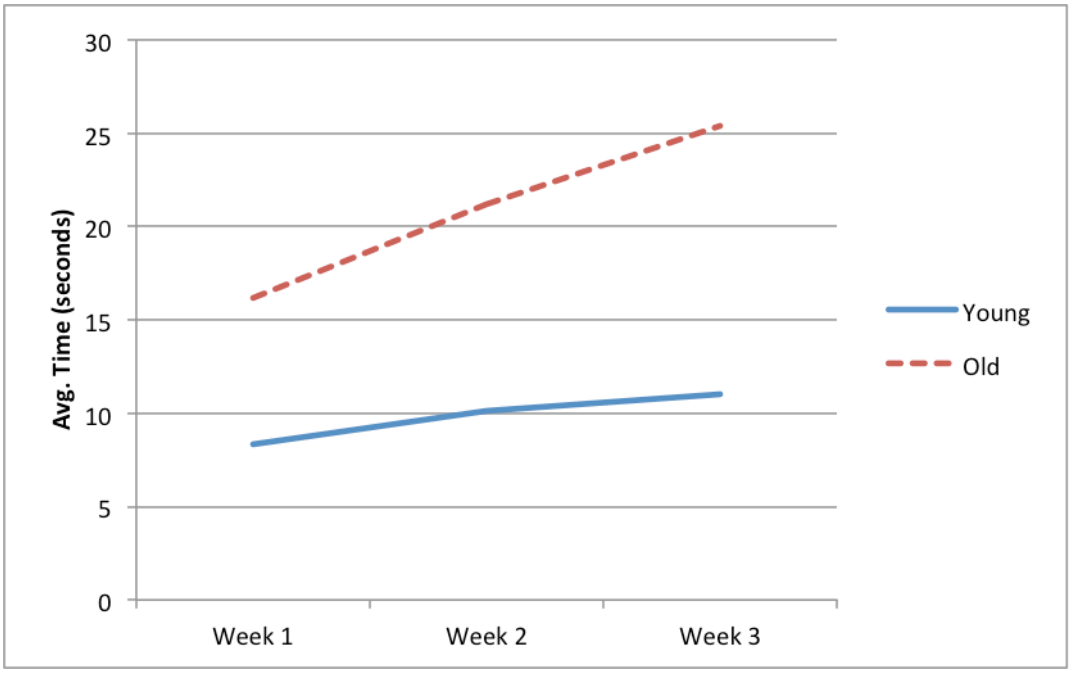


Figure 5.5: Interaction between Age and Week for SET 1.

There was no 3-way interaction between age, grid type, and week ($F(2,33)=0.011$, $p>.05$).

5.3.2.2. SET 2 – NEW CODES

For the average time taken to select the segments for SET 2 codes, a main effect of age was found ($F(1,34)=23.724$, $p<.001$) with younger participants (mean: 11.92 seconds) selecting their segments faster than older participants (mean: 22.75 seconds). No main effect of grid type was found ($F(1, 34)=0.071$, $p>.05$). A main effect of week was present ($F(1,34)=14.513$, $p=.001$) with participants selecting their segments faster in the second week (mean: 14.06) compared to the third week (mean: 20.61). Once again these findings reflect those found earlier for accuracy (i.e. no sign of a speed-accuracy trade-off).

No interactions were found between grid type and age ($F(1, 34)=0.127$, $p >.05$), age and week ($F(1,34)=0.487$, $p>.05$), or grid type and week ($F(1, 34)=0.137$, $p>.05$). There was no three-way interaction between age, grid type, and week ($F(1,34)=2.313$, $p>.05$).

5.3.3. Order of Acquisition Effects

A chi square test using participant age and forgotten codes as factors was carried out on the data. A significant association between participant age and forgetting of codes was found when remembering 4 codes, $\chi^2(1)=21.043$, $p<.001$. This seems to represent the fact that older participants were more likely to forget codes than younger participants (see Figure 5.6).

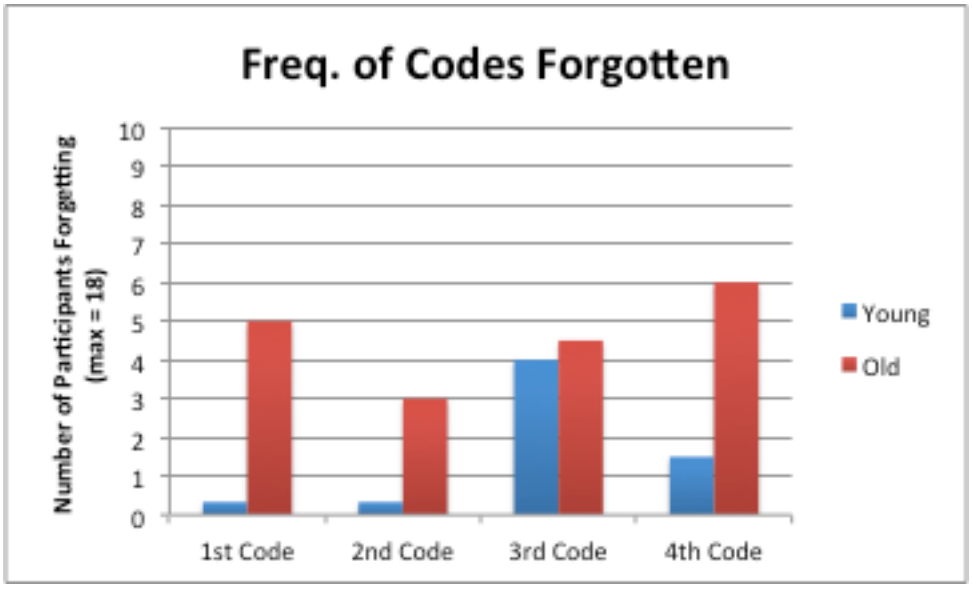


Figure 5.6: Frequency of forgetting for Tiles codes for both younger and older participants.

Separate chi square tests were run on the younger and older participants with order of acquisition and forgetting. The test for younger participants found that order of acquisition was significantly associated with the forgetting of codes, $\chi^2(5)=16.803$, $p=.001$. This seems to suggest that the third code was the most challenging to retain, with 61.5% of younger participants forgetting the third code. The test for older participants revealed no association between order of acquisition and forgetting of codes. The poorer retention of the third code appears to be an anomaly due to code number four not being affected.

5.4. Discussion

Overall, a main effect of age was present for both similar and dissimilar grids where the younger group performed both more accurately and faster than the older group. This age effect was in accordance with the prediction that was made prior to the study. However, the performance for older adults was poor and lower than the PIN benchmark. These

findings are in contrast to what was expected, with the psychology literature suggesting that older adults are able to take advantage of the visual nature of the system to at least a similar extent as younger adults (e.g. Winograd et al., 1982) for concrete images (Smith et al., 1990).

Older adults have been shown to use poor strategies when learning information (e.g. Naveh-Benjamin et al., 2007) and when learning pictures specifically they have been shown to encode the picture as a 'gist' – relying on a general impression of the concept of the picture rather than any specific detail (Koutstaal, 1997). In the context of this study participants were told to study the whole image and that every detail was relevant. In theory this should have resulted in the whole image being classed as primary information for encoding but it was not possible to enforce this on the participants. The second stage of the familiarisation process presented participants with the individual segments, thus emphasizing the different areas of the image and should have further encouraged a more detailed encoding of the picture. In the future perhaps it may be suitable to allow participants more time to observe the whole image to guarantee that the whole picture is being taken into consideration. The implications of the extra time needed to register for an account under this configuration should also be considered.

A possible explanation for the relatively poor performance overall (experienced by the older group – and to some extent the younger group with SET 2 codes) is the image content. The base images that were selected for this study consisted of picturesque scenes utilising the whole picture frame. However, it is possible that the segmentation of the images resulted in segments that did not portray enough detail regarding the original image – or in other words the segment resulted in an abstract image. Work by Smith et al. (1990) has found that older adults' memory for abstract images was not as accurate as that for concrete images, therefore it could be that any resulting abstract target segments were not recognised by the older group. In effect, the segmentation of the images could have removed the meaningfulness of the image. In addition, the resulting segments may have been treated as individual images due to the lack of meaningfulness, leading participants having to remember nine target images instead of the intended single target image – considerably more than the standard four targets for existing GAS.

A further explanation for the insensitivity of the older group to the grid types could be the well-documented problems that older adults experience with the binding of objects and location (e.g. Chalfonte & Johnson, 1996; Park et al., 1983). The purpose for keeping the location of the target segments fixed when mapped onto the base image was to help participants by adding further context to the segment – e.g. if the door was supposed to be on the bottom left of the image and a segment showed a door on the top right of the grid then it could be discarded. While it is apparent that younger adults were able to do this, older adults may not have benefitted from location consistency. Work by Chalfonte et al. (1996) demonstrated that older adults are disadvantaged when having to recall the location of objects on a screen which may partially explain their problems with the Tiles system. Although Tiles did not require participants to remember the location of the segments explicitly, being able to remember the location would have aided the recognition and discrimination of segments. Older participants did not appear to be able to use this feature of the system and as a consequence their performance with both grid types was the same.

A decline in accuracy over time was observed for all participants with both sets of codes. The time-based performance decrement observed for both age groups was supported by previous research into multiple graphical authentication systems. In accordance with Moncur and Le Plâtre (2007) and Chiasson et al. (2009), a performance decrement was observed after a one-week delay when participants had multiple graphical systems to remember. The 3-way interaction found between age grid type and week with the new codes showed a disadvantage for the older group when using dissimilar grids where their accuracy declined at a significant rate from week 2 to week 3 while the younger group demonstrated the same effect but with similar grids. The most important finding from this 3-way interaction is the drastic decline in accuracy for both age groups with SET2 codes, making it clear that the addition of multiple codes causes participants of both ages – albeit with different grid types – to become significantly less accurate after a one week delay.

5.4.1.1. SPEED

As expected, younger participants were significantly faster than older participants in selecting their segments. This was the case across all conditions and presents no

surprises. The speed results demonstrate that there was no accuracy-speed tradeoff and that younger participants were clearly superior to older participants.

5.4.2. IMPLICATIONS

A main effect of age was found where the younger participants were more accurate and faster than older participants. This finding was not a surprise and simply reflects the reality of knowledge-based user authentication, where older adults will always be at a disadvantage when compared with younger adults. However, the lack of a main effect of grid type was surprising. When looking at the third week more carefully, however, it emerges that younger participants were able to take advantage of the dissimilar grids but the older group showed no difference in performance between the two grid types – and generally performed very poorly with both. In other words, older adults underperformed with Tiles regardless of the grid type, while younger participants were able to take advantage of the 'easier' grid configuration as was expected, but their performance with the 'harder' grid configuration was comparable to that of the older adults.

This evaluation of Tiles followed a comparable methodology to that used to evaluate PINs, and as such some comparison can be made between the two studies. In the case of PIN, the benchmark for the older group was 2.6 average successful attempts with SET 1 codes and 2.7 average successful attempts with SET 2 codes. Older adults showed a similar performance using Tiles with SET 1 codes, but noticeably worse performance with SET 2 codes. As such, the accuracy results from the Tiles system were worse than expected. Participants using this graphical system were not able to improve on the accuracy shown with the PIN system. In fact, the accuracy results for the second set of codes shows a significant decline from week 2 to week 3 which was not observed for PIN, making the Tiles system more vulnerable to problems when multiple codes are introduced. It should be noted that short-term performance with Tiles exceeded that of PIN, for both SET 1 codes and SET 2 codes (i.e. week 2). These results imply that Tiles could feasibly be used if the user is guaranteed to authenticate consistently throughout the day – e.g. smartphone authentication. However, it could be argued that PINs serve this purpose already – remembering a PIN that is used regularly should not a problem due to repetition. It is possible to use Tiles as an alternative system to frequently used PINs or passwords in order to prevent the reuse of codes, but the repercussions of

delaying a login beyond a week – especially if multiple codes are used – could be severe. Simply, Tiles does not provide the inclusive authentication that was hoped for.

The results bare some resemblance to those of Chiasson et al. (2009) where GAS showed a marked improvement over existing KBA systems in the short term but long-term recall of the graphical system did not exceed that of the KB system. The similar result shows that the paradigm developed and used in this thesis produces realistic results. This is encouraging for the later evaluations that will be used for comparisons.

The results from the Tiles GAS demonstrate important design considerations for graphical systems. The picture superiority effect has been shown to be present for older adults (e.g. Winograd et al., 1982) but this benefit can be nullified if the graphical system is not implemented correctly. In this case the segmentation of the images is believed to have affected the ability of older adults to recognise the segments. This performance drop experienced by the older group – and the younger group to an extent – suggest that inclusive GAS should only implement full images and not partial or segmented images.

5.4.3. Design Implications

In order to improve the memorability of the Tiles system it is recommended that the base images be selected on the basis that little or no 'dead space' is present. Dead space refers to areas of an image that may contain repetitive similar elements (e.g. water or sky) that could be easily confused by the participant. The screening process would benefit both younger and older adults on the basis that the segmentation process would yield less abstract segments.

A longer and more extensive familiarisation process would also benefit the older group. More time focusing on the base image, with relevant questions that force participants to engage with the whole image rather than select specific details that may not be present in a number of segments. A final stage could also be added where all the segments are shown together and the participant is asked to identify the segments they believe will be the most problematic and explain how those segments fit in with the rest of the image. Additionally, the practice phase should require participants to select every segment from amongst foils, rather than four random segments in order to identify any

problematic segments. The one problem with this approach would be the added time to the familiarisation processes, and as a consequence to the whole registration stage. Whether the added time is acceptable to the users remains to be seen.

A potential modification to improve the system is to utilise base images that contain a number of detailed objects – at least four – and then require participants to select the correct objects from amongst other objects to authenticate. In essence, the base image would not be segmented in the way it was for this study – i.e. absolute segmentation – but instead some of the objects would be extracted from that image and presented to the user in a grid along with other similar objects – similar to traditional recognition-based GAS, e.g. VIP. This approach would benefit older adults as no image segmentation would take place, and they would benefit from the context of the base image – i.e. they should be able to imagine the base image and extrapolate the correct objects.

5.5. CHAPTER SUMMARY

This chapter evaluated a new GAS that was designed to reduce the cognitive load associated with recognition-based GAS. Younger and older adults were tested with multiple Tiles codes over the course of three weeks. In comparison to study 1 (Chapter 4), it was predicted that the performance of the older group, relative to the younger group, would be relatively strong (i.e. Tiles would work to benefit older adults). However, the study found that older participants were less accurate and slower than younger participants when selecting their target segments. Only younger participants were able to take advantage of the dissimilar grid configuration while the older participants did not show a difference between the two grid configurations. Finally, the accuracy of both younger and older participants significantly declined after a week with SET 2 codes and most importantly the long-term performance with Tiles was not an improvement to that of PIN for either the younger or the older group.

6. GRAPHICAL AUTHENTICATION SYSTEMS RE-ENGINEERED: FACES AND PICTURES

6.1. RATIONALE

The previous chapter found that a GAS with partial images (Tiles) was not effective as an authentication system for older adults – i.e. it was not effective in improving the performance of older adults when compared to the earlier PIN study. This chapter will explore GAS that utilise whole images as inclusive systems for older adults based on a literature that suggests performance gains for this group. The argument in favour of the graphical systems is that humans are much better at remembering and recognising images than they are at remembering strings of text (De Angeli et al., 2002; Dhamija & Perrig, 2000). A number of studies looking at the memorability of these graphical authentication systems have found that over a time delay the graphical systems are indeed more memorable than alphanumeric passwords (e.g. Dhamija & Perrig 2000; De Angeli et al. 2002). A reason for the apparent superiority of the graphical systems is that while users rely on pure recall when authenticating with passwords, graphical authentication systems allow users to rely either cued-recall or recognition, both better than pure recall for remembering (Baddeley, 1997).

These relatively new systems have been extensively tested on younger adults, but we do not know much about their performance with a population of older adults. Previous memory research suggests that older users may not be penalised as harshly with graphical systems than with alphanumeric passwords, as recognition has been shown to be less affected by ageing than recall (Brown & Park, 2003; Craik & McDowd, 1987), most likely due to the extra effort and resources required for pure recall (Raaijmakers & Schiffrin, 1992).

Additionally, previous research also shows that visual memory appears to be less affected by ageing than verbal memory (Brown & Park, 2003), meaning that memory for pictures is likely to be superior than memory for words, and therefore remembering a graphical combination would be less work than remembering an alphanumeric password. Older adults have been shown to benefit from the Picture Superiority Effect (Park et al., 1986; Winograd et al., 1982) to the same extent as younger adults.

Moreover, older adults have been shown to be able to remember images after a one-week delay without a significant drop in recognition performance, although a longer delay appears to negatively impact their performance (Park et al., 1988).

With regards to graphical authentication systems using faces, research has shown that older adults are very adept at recognising known faces, albeit not as accurately as younger adults (Ng, Hon, & Lee, 2007; Smith & Winograd, 1978). The accuracy of recognition depends on the familiarity of the target faces (Searcy et al., 1999) which can potentially be problematic for a security system where the use of known faces may compromise the system. Hence, the use of known faces is not acceptable from a security perspective.

Younger and older adults were tested with two different GAS, a face-based system called Faces and a picture-based system called Pictures, over a three week period. These GAS were based on existing traditional recognition-based systems with a few modifications to enhance the security of the systems.

It is expected that the performance of older adults, in terms of accuracy and speed, will improve when compared with the PIN benchmark when using GAS. Based on previous work (e.g. Everitt et al., 2009) it is predicted that accuracy will drop over time but the rate of decline for the older group is expected to follow that of the younger group.

6.2. METHODOLOGY

6.2.1. EXPERIMENTAL DESIGN

The study consisted of two factorial designs as the codes were separated into 'original' accounts (SET 1) assigned to participants during the first week and tested in weeks 1, 2 and 3, and 'new' accounts (SET 2) assigned to participants during the second week and tested in weeks 2 and 3. The first set of images consisted of a 2 (participant age: young, old) x2 (system: Faces, Pictures) x3 (week of testing: week 1, 2, 3) factorial mixed design. The second set of images consisted of a 2 (participant age: young, old) x2 (system: Faces, Pictures) x2 (week of testing: week 2, 3) factorial mixed design. The factors comprised of one independent – the participants' age (young group, old group) –

and two repeated –the graphical system tested (face-based 'Faces', picture-based 'Pictures') and the time period (weeks 1, 2, 3).

Dependent factors comprised the number of successful authentication attempts (maximum of 5 per account), the average time to authenticate (in seconds) and the number of correct image selections made regardless of account (Image Interference).

6.2.2. PARTICIPANTS

36 participants were recruited in total to fit into one of the two age groups, the younger group (18-30 years old, n=18) or the older group (65-75 years old, n=18).

Younger participants (mean age: 23, SD: 2.78) were recruited from the student population in the university using an online participation pool maintained by the university. Given the diversity of the student population this sample was considered adequate.

Older participants (mean age: 69, SD: 3.56) were recruited using the lab's participant database as well as through an advert on the Elders' Council of Newcastle newsletter. All participants were given £30 to cover travel expenses to and from the university.

Participants were screened for age and for computer experience—they were required to have used a computer prior to taking part in the study. Participants were also screened for adequate vision and for previous extended contact with Caucasian people. The reason for the latter requirement was due to the use of Caucasian faces for the Faces system, and previous research has shown that people are much better at remembering faces of their own ethnic origin or of ethnicities that they have had extended contact with (e.g. Walker & Tanaka, 2003). Finally, participants were excluded if they suffered from prosopagnosia, a face recognition disorder.

6.2.3. MATERIALS

The materials were designed to reproduce a real authentication system and the experience of logging into multiple accounts. Four account names were created for the study. The names used were Bank, ATM, NHS and Email. Each account was allocated

four target images that the participants would be required to learn and remember over the course of the study.

The face-based system, Faces, was modelled on the commercial GAS Passfaces (Valentine, 1998). The foil faces on each grid were controlled for visual similarity in order to improve the security of the system with regards to observation attacks. In other aspects the Faces system resembled Passfaces with four challenge grids being presented to the participant, each containing nine faces – one of which was correct.
The picture-based system, Pictures, was loosely modelled on the original VIP system (De Angeli et al., 2002). The foil pictures on each grid were controlled in terms of similarity by belonging to the same semantic category. This approach was taken to improve the security of the system in terms of description so that the pictures cannot simply be accessed by knowing the label – i.e. cannot say "the boat" – or by observation – by observing that it is a boat an attacker cannot necessarily replicate a selection. The pictures used differed from those used in VIP as concrete everyday objects were chosen to form the categories – e.g. bollards, lamp posts, etc.

The main material used was the graphical authentication system that was created for this study using Experiment Builder v 1.5.201. Both Faces and Pictures used the same underlying program, with the difference being the stimulus that was used – faces for Faces and pictures for Pictures. The graphical authentication systems used on the same computer for all participants in order to guarantee the same experience for all participants.

Initially, the program displayed a set of simplified instructions for the participants and they were required to press the spacebar to start after the complete oral instructions were given. Four sequential challenge grids were displayed upon the participants' selection. Once four selections were made using the mouse, the system informed participants whether they had selected correctly or incorrectly, no feedback on individual selections was given. The system carried out the cycle four more iterations, meaning participants selected their codes five times in total.

6.2.3.1. GRID COMPOSITION FOR FACES

The faces were obtained from a university smartcard database and the face pool consisted of 321 males and 38 females, all numbered. The smartcard photos were taken to a constant specification, making the presentation of all the faces in the database very uniform – i.e. same background and similar poses. All faces were in full colour.

The database consisted predominantly of young white male faces, therefore in order to preserve similar levels of grid strength and avoid any complications regarding race, gender, etc. only Caucasian male faces were used. Female faces and non-Caucasian male faces were removed from the pool and resulted in 280 faces, all numbered 1-280. Faces with neutral or positive expressions were selected to make them more memorable for the older participants.

Figure 6.1: Example grids for Faces (left) and Pictures (right).

The target faces were randomly chosen from the face bank using a random number generator and were assigned to all participants. This meant all participants received the same 8 faces (Figure 6.2 and Figure 6.3 showcase the target images used for the study). The pool of remaining faces was then renumbered from 1 to 272. Each of the target faces was assigned 12 foil faces that would appear on the challenge grids. The foil faces were chosen to be similar to the target images in order to avoid the target images from standing out due to any outstanding features – i.e. distinctiveness – and ensure that all grids were similar in difficulty. This was done by grouping the faces in terms of visual similarity. The similarity of the faces was obtained by asking 16 random people across

the university to rate the bank of images on the basis of visual similarity—how likely the faces are to be confused. For every target face, the person rating was asked to select the 12 most similar faces to each target face from the bank and write the numbers down on a sheet of paper. Research by Dunphy, Nicholson, and Olivier (2008) shows that it is significantly easier to guess a target face using a description from a randomly generated grid than when the grid is built using similar faces. Based on these findings it was decided to use visually similar groupings for the grids. Once all target faces had been rated, the 12 most similar faces—classed as the 12 most recurring chosen faces—were chosen as the foils.

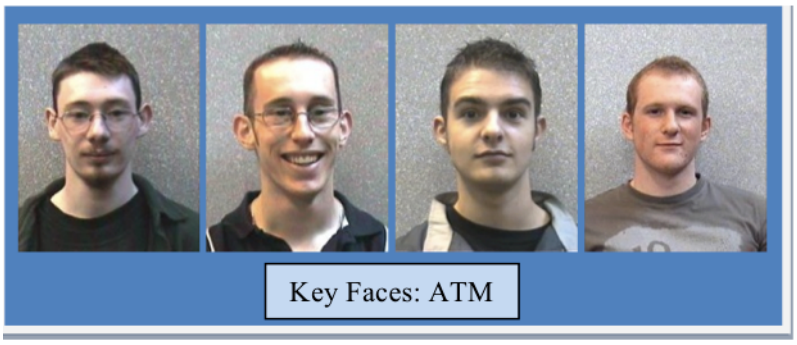


Figure 6.2: Target faces used for ATM account.

Figure 6.3: Target faces used for Bank account.

Sixteen sequences of nine numbers were generated using the random number generator. These sequences would form the challenge grids that would be displayed to participants. The position of the faces on the grid were determined by the order of the numbers in the sequences (where 1 was top left and 9 was bottom right). The sequences were inspected to make sure that an overlapping foil was present (see Image Overlap 6.2.2.3). The computer program randomised the presentation of the grids for each participant (see Appendix A for sample grids).

6.2.3.2. GRID COMPOSITION FOR PICTURES

The images used for the Pictures system were collected by the researcher from around the city of Newcastle upon Tyne and consisted of 8 categories: bollards, light posts, drains, rubbish bins, recycling bins, benches, satellite dishes, and trees. From these 8 categories, the 4 with the best quality pictures were selected (first 4 categories). Pictures of neutral objects were utilised in this study due to previous research from Charles et al. (2003) suggests that negative images are more difficult to recall for older adults. Although a number of the images were provided out of context (for example, a photograph of a drain without its surroundings), a familiarisation process was undertaken with each participant in order to help them put the pictures into context, something that should have benefitted the older group greatly (Park, Puglisi, & Smith, 1986).

The same procedure that was used for Faces was used to select the targets and foils for the Pictures system (Figure 6.4 and 6.5 showcase the target images used for the study). However, no similarity rating exercise was carried out as all pictures on the grid were from the same category (e.g. bollards) and it was thought that random selection of foils would be appropriate. To this end, 12 random foils were selected for each of the 8 target pictures from their respective categories.

The same sixteen sequences of nine numbers were that were used for the Faces were also used for Pictures. The sequences would form the challenge grids that would be displayed to participants. The position of the pictures on the grid were determined by the order of the numbers in the sequence (where 1 was top left and 9 was bottom right). The sequences were inspected to make sure that an overlapping foil was present (see Image Overlap 6.2.2.3 below). The computer program randomised the presentation of the grids for each participant (see Appendix A for sample grids).

Figure 6.4: Target pictures used for Email account.

Figure 6.5: Target pictures used for NHS account.

6.2.3.3. IMAGE OVERLAP

As mentioned previously, the grids were composed so that each grid contained one target image and eight foil images. Seven of the eight foil images were selected randomly from the foil set of 12 using the random number generator as detailed in the previous subsection (see 6.2.3.1). The final foil image was a target image from the other account. This meant that there was an overlap in the foils and targets, and each grid contained two images that had been assigned to the participant, although only one image was correct for each account (see Figure 6.6).

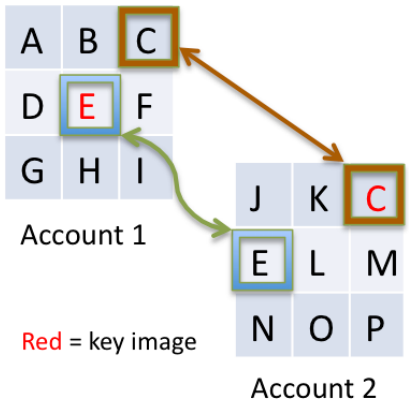


Figure 6.6: Illustration of image overlap - each grid contains one target image and one foil image used as a target for another account.

The overlap used in constructing the grids was one of the aspects that differentiated this system from existing graphical authentication systems. Previous studies investigating interference effects amongst multiple graphical authentication systems have explicitly avoided using the same images more than once to prevent the participant from becoming confused (e.g. Everitt et al., 2009), while others have do not mention whether they avoided the use of overlapping image or not.

This design was chosen to represent a realistic scenario where image databases are shared amongst providers and the possibility of one target appearing as a foil for another target becomes a reality. Previous evaluations of GAS have used optimal pictures for users – as recommended by Renaud (2009) – resulting in excellent performance by participants. However, these systems, other than Passfaces, are yet to be implemented in the real world. The aim of evaluating the systems with an image overlap – a more difficult task – is to obtain performance data of a viable system that could be implemented in the wild.

6.2.4. PROCEDURE

The procedure for this study consisted of two stages: the enrolment and the authentication stage. During enrolment, participants learned their images. During authentication, participants attempted to access their 'accounts' by identifying the four correct images that belonged to the account.

6.2.4.1. ENROLMENT STAGE & FAMILIARISATION

During the enrolment stage participants were given account names and the target images to learn and were guided through a familiarisation process consisting of seven questions aimed at creating a lasting bond between the participant and the images and 5 practice trials selecting their images (for a list of the questions see Table 6.1). The participant was given each image of an account in turn, and after having 15 seconds to study the image they were asked the respective questions by the experimenter. The participant was required to reply to the questions out loud. The questions for the Faces and for the Pictures varied slightly but both had the aim of getting the participant to study the image in as much detail as possible, and think about the different attributes of

that image to make it easier to remember at a later time. Overall, the familiarisation process for an account took approximately five minutes.

Table 6.1: List of questions used for both Faces and Pictures during the familiarisation process.

Questions for Faces
What do you think the name of this person is?
How old do you think [name] is?
Do you think [name] looks like a friendly person?
Does [name] remind you of anyone you know?
What do you think [name]'s job is?
Where do you think [name] lives?
What do you think about [name]'s family?
Questions for Pictures
What would you call this picture?
Have you seen a similar scene to this picture before?
Does this picture remind you of any particular experience or event in your life?
What stands out the most in this picture?
What is the smallest bit of detail or object that you can spot in this picture?
What do you think is happening in this picture?
What do you think the photographer was thinking when the picture was taken?

Previous research has shown that older adults heavily rely on resemblance for recognising faces (Bartlett & Fulton, 1991), therefore it was imperative to ask questions that required the participants to compare the new faces to older faces.

Once participants had been through the familiarisation process for the account code, they were asked to select their images from a set of foils to confirm that they had learned them, essentially a mock authentication attempt (see 6.2.4.2 below). The practice session also allowed participants further time to learn the images in context (e.g. discriminate from foils). If the participant failed to select their faces correctly at least three attempts in a row they were asked to repeat the mock authentication. They

were also given the option of being shown the faces again. Participants were then taken through the same process for their next account.

6.2.4.2. AUTHENTICATION STAGE

During the authentication stage participants were presented with a login screen and asked to authenticate. The login screen contained a 3x3 challenge grid where participants were required to point to their target image from the nine in four successive screens. The experimenter then clicked on the images using the mouse. The reason for not allowing participants to click for themselves was to account for the vast differences in mouse ability amongst the older participants. Following the selection of an image, participants were asked to fill in a confidence scale, requiring the participant to call out a number between 1 and 5 with 1 being not confident at all and 5 being very confident. Once they had called out a number they were then taken to the next selection screen. Once the participant selected four images, the program told them whether they selected correctly or incorrectly—if incorrect they were not told which images they got wrong. Participants were required to do this for a total of five times per set regardless of whether they selected correctly or incorrectly, a total of 20 image selections.

6.2.4.3. PROCEDURE FOR PARTICIPANTS

Participants were asked to attend the lab on three separate occasions (see Table 6.2 for overview of procedure). During the first session, participants were given their first set of images (SET 1) consisting of one Faces account and one Pictures account. The order of the accounts was randomised in order to eliminate any order effects. Participants were taken through the enrolment stage with the first account and were then taken through the enrolment stage once more with their second account. After enrolling with both systems, participants were distracted from the encoding task by taking part in a short discussion with the investigator. This discussion focused on their experience of current authentication systems such as passwords and smartcards and lasted approximately 10 minutes.

Following the discussion, participants taken through the authentication stage with the order of the accounts randomised.

Table 6.2: Overview of procedure.

Session	Activity
1	1. Enrolment with SET 1 2. Discussion (distractor) 3. Authentication with SET 1
2 (+1 week)	1. Authentication with SET 1 2. Enrolment with SET 2 3. Discussion (distractor) 4. Authentication with SET 2
3 (+1 week)	• Authentication with SET 2 • Authentication with SET 1 3. Discussion

Participants were asked to return to the lab a week after the first session. Upon their arrival they were greeted and were asked to once again authenticate using the target images they were assigned in the first session a week ago (SET 1) and once again the order of the accounts were randomised. Once participants finished authenticating, they were enrolled with two new accounts, again one from each system (SET 2). Following the enrolment with their second set, participants were engaged in another discussion, this time regarding their experience of using the two systems. This discussion lasted for approximately 10 minutes. Upon the completion of the discussion, participants were asked to authenticate with the new set of images (SET 2). They were not asked to authenticate with their first set of images.

Participants visited the lab for a final time one week after the second session. Upon their arrival they were greeted and were asked to authenticate with the images they were assigned in the previous two sessions. The order of the sets was randomised, as well as the order of the accounts in each set. Once they had authenticated with both sets of images, participants were asked questions about their experience and about their strategies for remembering the images.

6.3. RESULTS

A 3-way ANOVA with repeated measures on two factors (System and Week) and age as an independent factor was carried out on both SET 1 and SET 2. As before, accuracy (number of successful attempts) and time to authenticate were the main variables under investigation. In addition, a measure of pure image recognition was taking, comprising of correct image selected and incorrect images selected that belonged to another assigned account over total selections. This measure was important for determining the amount of interference that the image overlap caused. For a table of means see Appendix B.

The number of successful attempts measured the number of times participants selected all four images correctly in an attempt, to a maximum of five times (per account). The average time to authenticate, recorded in seconds, measured the average duration of an attempt, described as the time needed to select the four images constituting a code – from the presentation of the first image to the selection of the final image. As an additional analysis, Image Interference measured the number of images that were selected correctly along with the false-positives that were selected—images that were assigned to the participant but that did not belong to the account being tested.

6.3.1. SUCCESSFUL ATTEMPTS – ACCURACY

6.3.1.1. SET 1 – ORIGINAL CODES

Participants' scores (out of five) for each of the two codes that made up SET 1 were collated for each of the three weeks. A 2 (participant age: young, old) x2(system: Faces, Pictures) x3 (week of testing: week 1, 2, 3) mixed factorial ANOVA was carried out.

For SET 1 codes, a main effect of age was found ($F(1,34)=7.475$, $p=.01$) with younger participants (mean: 4.91) achieving more successful attempts than older participants (mean: 4.56). No main effect of system was found ($F(1,34)=1.602$, $p>.05$). A main effect of week was present ($F(2,33)=6.974$, $p<.01$) with pairwise comparisons showing that participants achieved significantly more successful attempts in the first week (mean: 5.00) compared to the third week (mean: 4.33) ($p<.010$) and also achieved significantly more attempts in the second week (4.90) when compared with the third

week (mean: 4.33) ($p<.05$). No significant difference in accuracy was found between the first week and the second week.

An interaction effect between age and system ($F(1,34)=4.450$, $p<.05$) was found (see Figure 6.7). An independent samples t-test on the *Pictures* data shows that the difference in accuracy between the younger group (mean: 4.94) and the older group (mean: 4.41) was significant ($t(106)=3.215$, $p<.05$) while the difference in accuracy between the younger group (mean: 4.90) and the older group (mean: 4.70) with *Faces* was not significant ($t(70)=1.014$, $p>.05$). This confirms that older participants could not be differentiated from the younger group when using Faces, but they were outperformed when using Pictures.

No interactions were found between age and week ($F(2, 33)=2.800$, $p>.05$) or system and week ($F(2, 33)=1.337$, $p > .05$).

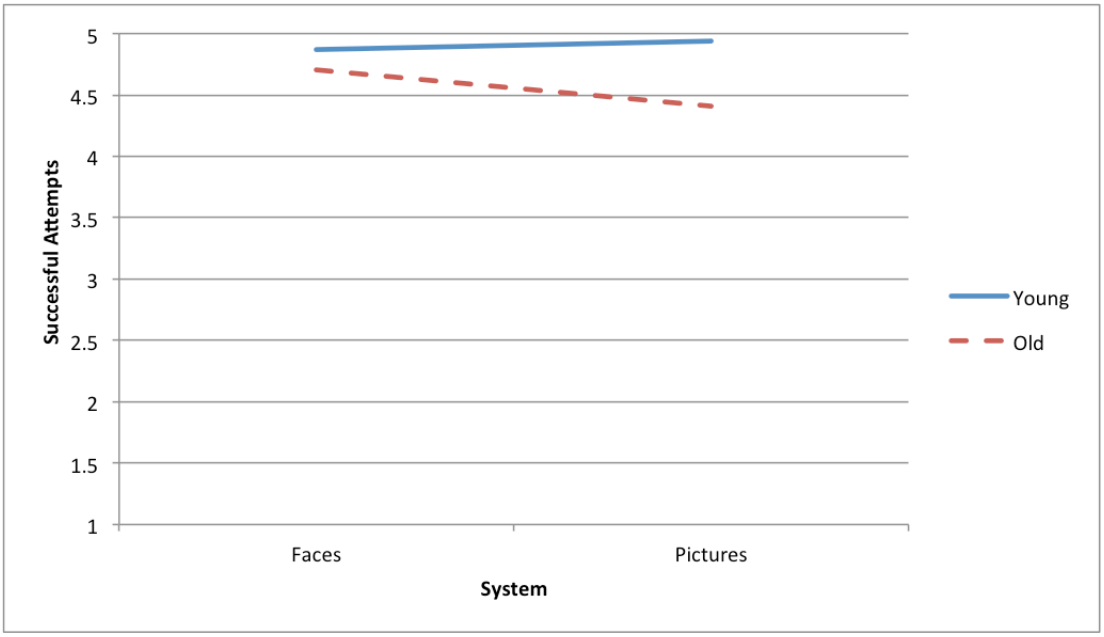


Figure 6.7: Interaction between age and system for SET 1, showing the older group being more accurate with Faces than with Pictures, while the younger group were more less accurate with Faces than with Pictures.

There was no 3-way interaction between age, system and week ($F(2,33)=1.432$, $p>.05$).

6.3.1.2. SET 2 – NEW CODES

Participants' scores (out of five) for the two codes that made up SET 2 were collated for each of the two weeks. A 2 (participant age: young, old) x2 (system: Faces, Pictures) x3 (week of testing: week 2, 3) mixed factorial ANOVA was carried out.

For SET 2 codes, a main effect of age was found ($F(1,34)=12.930$, $p=.001$) with younger participants (mean: 4.90) achieving more successful attempts than older participants (mean: 4.13). No main effect of system was found ($F(1, 34)=0.088$, $p>.05$). A main effect of week was present ($F(1,34)=18.604$, $p<.001$) with participants achieving more successful attempts in the second week (mean: 4.90) compared to the third week (mean: 4.10) ($p<.001$).

No interaction was found between age and system ($F(1,34)=1.076$, $p>.05$). An interaction was present between age and week ($F(1,34)=10.707$, $p<.01$) (see Figure 6.8). Independent samples t-tests confirm that the accuracy of younger participants did not significantly decline from week 2 (mean: 5.00) to week 3 (mean: 4.80) ($t(70)=1.324$, $p>.05$) while a performance dropoff was observed for the older participants with significant declines from week 2 (mean: 4.83) to week 3 (mean: 3.42) ($t(70)=4.056$, $p<.001$).

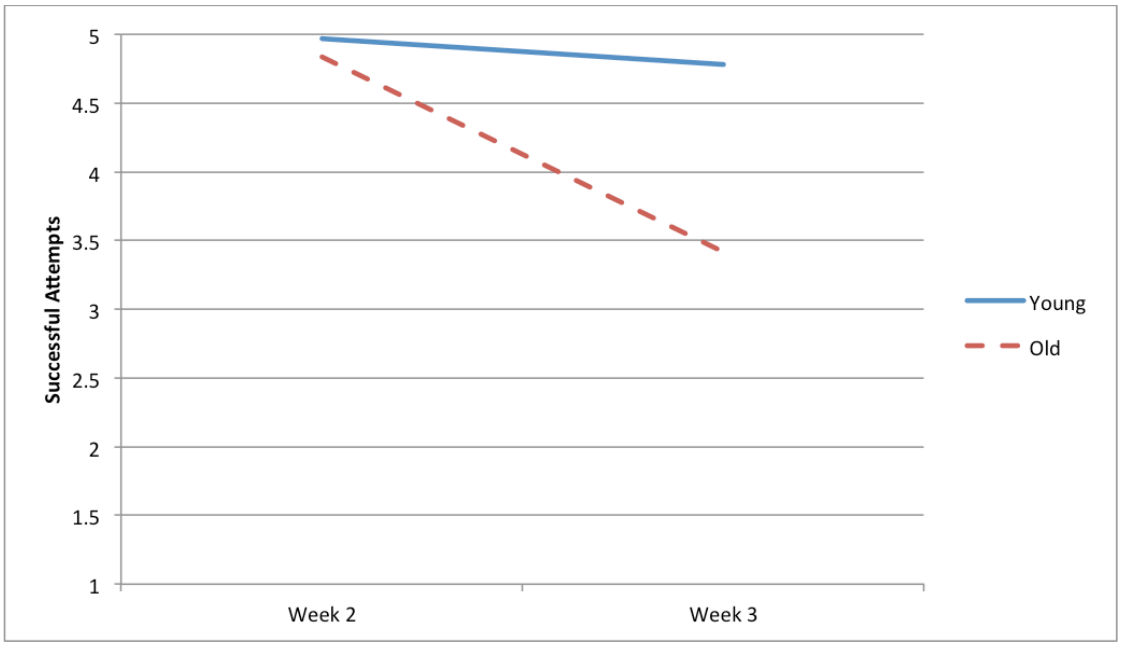


Figure 6.8: Illustration of the interaction between age and week for SET 2, showing how the accuracy of the older group dropped significantly during the second week of the study.

No interaction effect was found for system and week ($F(1,34)=0.217$, $p>.05$). There was no 3-way interaction between age, system and week ($F(1,34)=0.866$, $p>.05$).

6.3.1.3. OVERALL ACCURACY

The first trend to stand out from the results is the decline in accuracy for the older group during week three. The interaction effect between age and week with SET 2 codes confirms that older participants struggled significantly more than younger participants when recognising their images during the final week. Figure 6.9 below shows this performance drop off clearly.

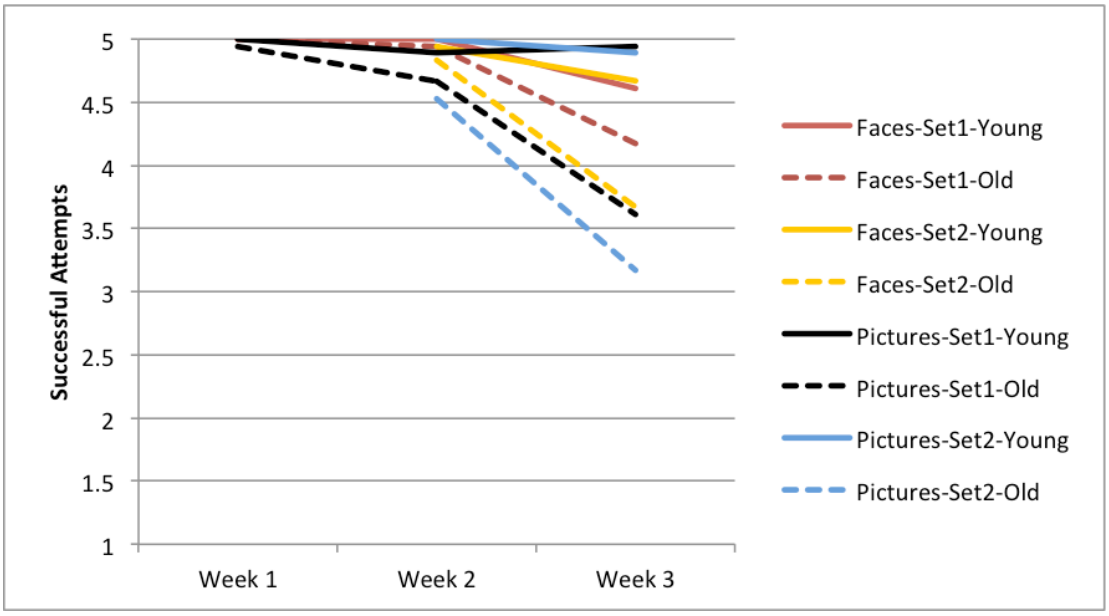


Figure 6.9: Overview of successful attempts for both sets of images.

Young participants were shown to be more accurate with Pictures than with Faces as captured by the age x system interaction with SET 1 but this effect was not present in SET 2. Older participants were shown to be more accurate with Faces than with Pictures as captured by the age by system interaction with SET 1 but this effect was masked by the addition of further codes (SET 2).

It is apparent from the study overview (Figure 6.9) that the drop off for older adults is more pronounced than that for the younger group. The gradient is particularly striking for SET 2 codes – especially for the older group. Despite the decline of accuracy over

time for both age groups, overall accuracy is an improvement on the benchmark set by PIN earlier.

6.3.2. Average Time - Speed

6.3.2.1. Set 1 – Original Codes

Participants' average time taken (in seconds) to select their images making up the codes for SET 1 were collated for each of the three weeks. A 2 (participant age: young, old) x2 (system: Faces, Pictures) x3 (week of testing: week 1, 2, 3) mixed factorial ANOVA was carried out.

For SET 1 codes, a main effect of age was found ($F(1,34)=49.252$, $p<.05$) where younger participants (mean: 12.89 seconds) were faster than older participants (mean: 18.63 seconds) when selecting their images. No main effect of system was found ($F(1,34)=0.017$, $p>.05$). However, a main effect of week was present ($F(2,33)=15.951$, $p<.05$). Pairwise comparisions show that participants took less time to select their images in the first week (mean: 13.84 seconds) over the second week (mean: 14.94 seconds) and the third week (mean: 18.49 seconds), as well as taking less time to select their images in the second week when compared with the third week.

An interaction effect was found between age and system ($F(1,34)=4.538$, $p<.05$) (see Figure 6.10). Two independent samples t-tests show no significant difference in speed between Faces and Pictures for either younger participants ($t(106)=1.730$, $p>.05$) or older participants ($t(106)=-.926$, $p>.05$). Another independent samples t-test showed no significant difference in speed between younger and older participants with either Faces or Pictures. No interaction effect was found between age and week ($F(2,33)=2.661$, $p>.05$).

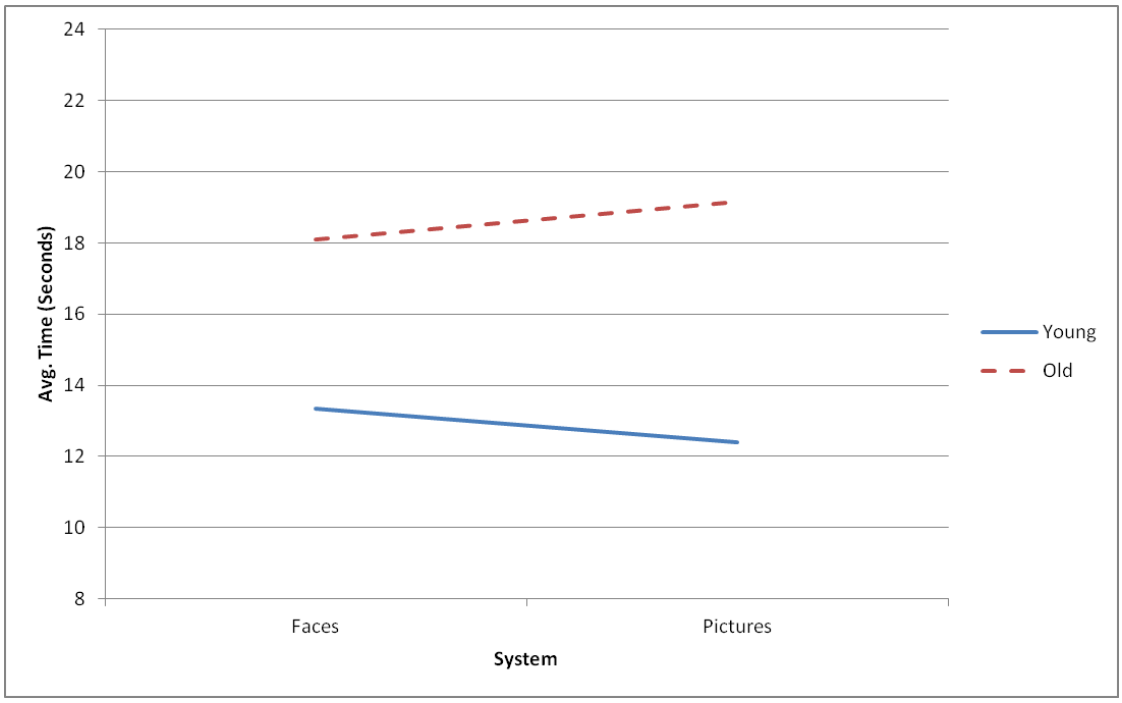


Figure 6.10: Illustration of the interaction between age and system for SET 1, showing how the older group took longer to select their Pictures, while the younger group were the other way around.

An interaction was present between system and week ($F(2,33)=11.624$, $p<.05$) (see Figure 6.11). A one-way ANOVA showed that for Faces a significant difference of speed was present where the difference between week 1 and week 2 was not significant($F(2,107)=13.771$, $p<.001$), but the difference between week 2 and week 3 was significant ($p<.001$) as was the difference between week 1 and week 3 ($p<.001$). Another one-way ANOVA showed that for Pictures a significant difference of speed with Pictures was present ($F(2, 107)=6.077$, $p<.01$) where the difference between week 1 and week 2 was not significant, the difference between week 2 and week 3 was not significant, but the difference between week 1 and week 3 was significant ($p<.01$).

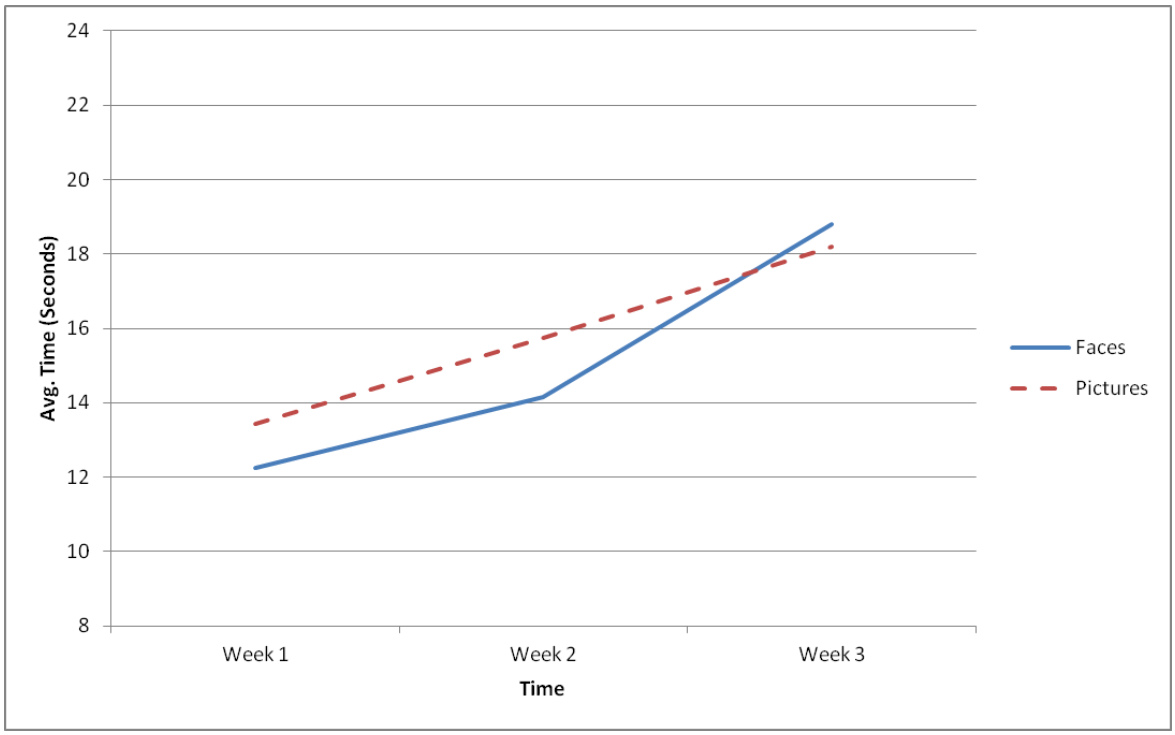


Figure 6.11: Illustration of the interaction between system and week for SET 1.

More interestingly, a 3-way interaction was found between system, time and age ($F(2,33)=5.788$, $p<.05$) (see Figure 6.12). Independent samples t-tests showed no significant difference between systems for either younger or older participants during week 1. Independent samples t-test showed no significant difference between systems for younger participants during week 2 ($t(34)=-.915$, $p>.05$) while a significant difference between systems was found for older participants during week 2 ($t(34)=-2.305$, $p<.05$) where Faces took less time to select (mean: 16 seconds) than Pictures (mean: 18.70 seconds). Independent samples t-tests found a significant difference between systems for younger participants during week 3 ($t(34)=2.427$, $p<.05$) where they were significantly faster with Pictures (mean: 12.90 seconds) than with Faces (mean: 15.91 seconds), while older participants did not show a significant difference between the systems in the third week ($t(34)=-.706$, $p>.05$).

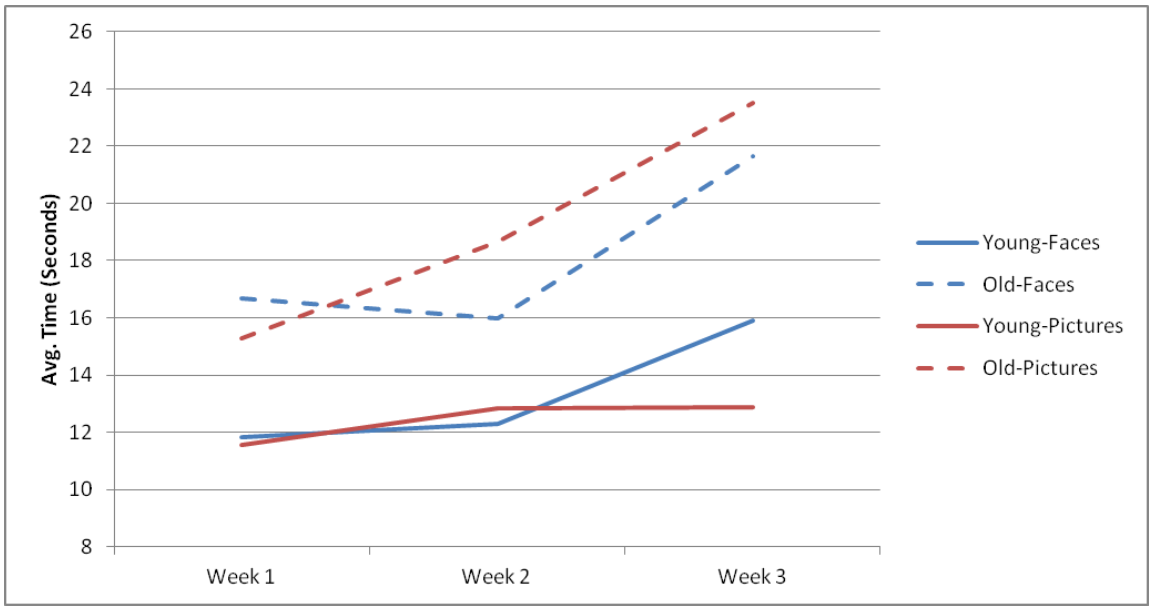


Figure 6.12: Illustration of the 3-way interaction between age, system and week for SET 1, showing how the younger group took significantly longer to select their faces during week 3 while the older group took significantly less time to select their Faces during week 2.

6.3.2.2. SET 2 – NEW CODES

Participants' average time taken (in seconds) to select their images making up the codes for SET 2 were collated for each of the three weeks. A 2 (participant age: young, old) x2 (system: Faces, Pictures) x2 (week of testing: week 2, 3) mixed factorial ANOVA was carried out.

For SET 2 codes, a main effect of age was found ($F(1,34)=52.003$, $p<.05$) where the younger group (mean: 14.51 seconds) were faster at identifying their images than the older group (mean: 20.90 seconds). No main effect of system was found ($F(1,34)=1.661$, $p > .05$). A main effect of week was also present ($F(1,34)=29.694$, $p<.05$) where participants took were quicker to select their images in the first week (mean: 15.90 seconds) over the second week (mean: 19.52 seconds).

No interaction effect was found between age and system ($F(1,34)=1.065$, $p>.05$). An interaction effect between age and week was present ($F(1,34)=7.058$, $p<.05$) (see Figure 6.13). Independent samples t-tests showed a significant difference in speed for both younger ($t(70)=-2.895$, $p<.01$) and older ($t(70)=-4.277$, $p<.001$) participants where images were selected faster during week 2 (mean young: 13.57 seconds; mean old:

18.13 seconds) than during week 3 (mean young: 15.54 seconds; mean old: 23.60 seconds).

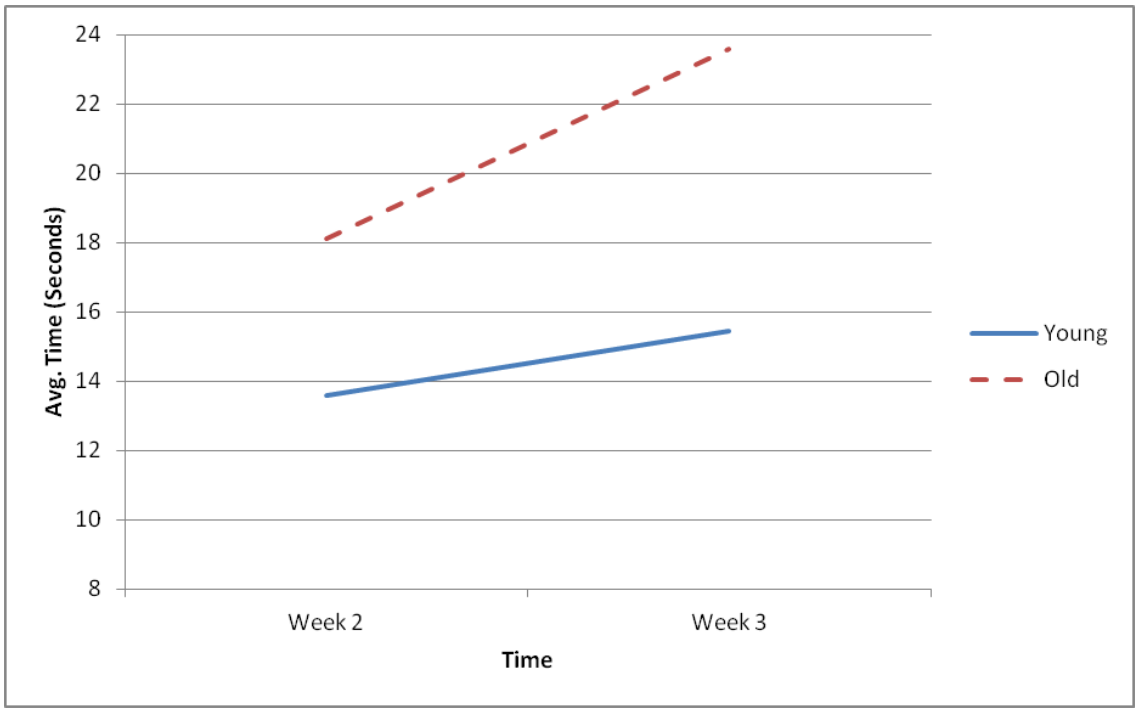


Figure 6.13: Illustration the interaction between age and week for SET 2, showing how the older group took longer to select their images in the second week of the study.

An interaction effect was also present between system and week ($F(1,34)=4.143$, $p=.050$) with Faces taking significantly longer to select in the second week (mean: 16.90 seconds) than Pictures (mean: 14.83 seconds: $t(70)=2.193$, $p<.05$) but no significant differences present in the third week (mean: 19.39 seconds; 19.70 seconds: $t(70)=-.177$, $p>.05$) (see Figure 6.14). In other words, the advantage of the Pictures system was lost by the final week.

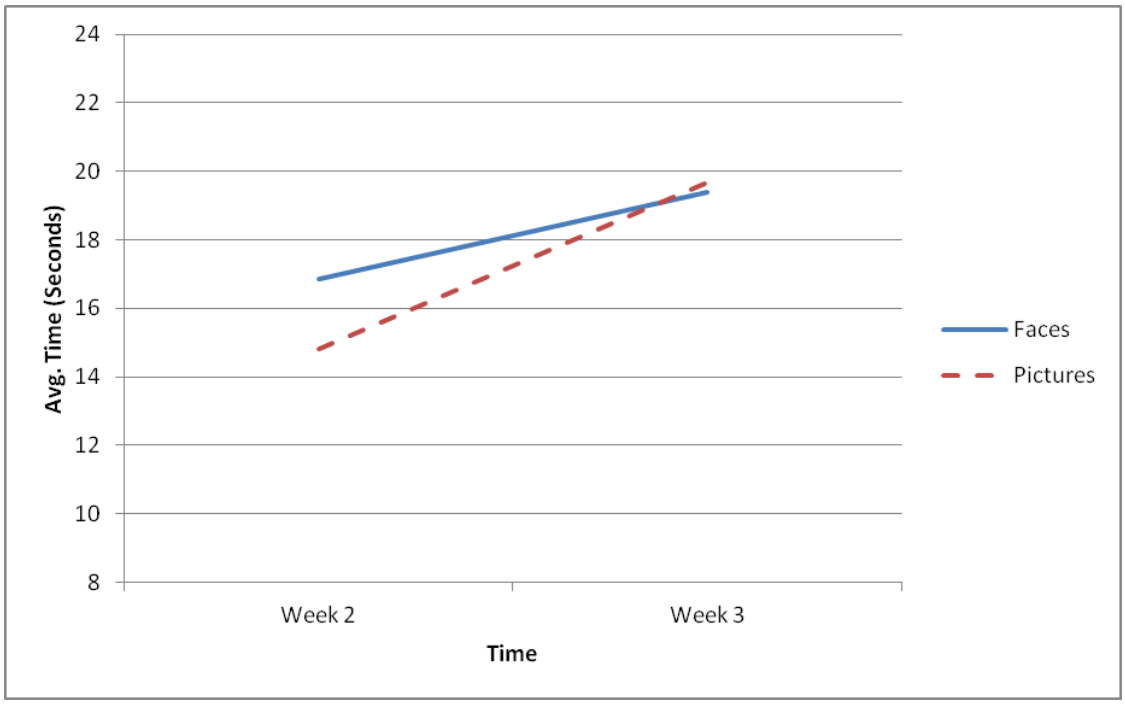


Figure 6.14: Illustration the interaction between system and week for SET 2, showing how initially Faces took longer to select but during the second week the time taken to select both images evened out.

There was no 3-way interaction between age, system and week ($F(1,34)=2.208$, $p>.05$).

6.3.1. Additional Analysis: Image Interference

6.3.1.1. Set 1 – Original Codes

Responses to the authentication task were re-scored to see if participants were able to recognise images independently of whether the images belonged to the correct codes – i.e. 'merging' of the binding between images and accounts. This analysis was used to demonstrate the extent to which the image overlap affected participants' recognition of the target images, and to obtain an idea of whether the simplification of the task – i.e. pure recognition rather than image binding and recognition – improved the performance of the older group. The maximum score for a participant was 20 per week – 5 attempts of 4 images each. A 2 (participant age: young, old) x2 (system: Faces, Pictures) x3 (week of testing: week 1, 2, 3) mixed factorial ANOVA was carried out.

When looking at pure image recognition, no main effect of age was found ($F(1,34)=2.909$, $p>.05$) which suggest that the binding of codes and accounts is problematic. There was also no main effect of system ($F(1,34)=1.280$, $p>.05$). There was a main effect of week present ($F(2,33)=4.615$, $p<.05$). Pairwise comparisons found

no significant difference between week 1 (mean: 20.00) and week 2 (mean: 19.94), between week 2 and week 3 (mean: 19.86), or between week 1 and week 3.

There were no interaction effects for age and system ($F(1,34)=0.569$, $p>.05$), age and week ($F(2,33)=1.448$, $p>.05$) or system and week ($F(2,33)=2.129$, $p>.05$). There was no 3-way interaction between age, system and week ($F(2,33)=0.558$, $p>.05$).

6.3.1.2. SET 2 – NEW CODES

Participants' scores (out of five) for each of the two codes that made up SET 2 were collated for each of the two weeks. Correct image selections and selections of other target images that did not belong to the specific account were counted as correct. A 2 (participant age: young, old) x2 (system: Faces, Pictures) x2 (week of testing: week 2, 3) mixed factorial ANOVA was carried out.

Again, when looking at the performance of partiipants without the binding require ment the age effect is gone ($F(1,34)=2.210$, $p>.05$), system ($F(1,34)=0.577$, $p>.05$), or week was present ($F(1,34)=0.063$, $p<.05$).

There were no interaction effects for age and system ($F(1,34)=0.064$, $p>.05$), age and week ($F(1,34)=0.063$, $p>.05$) or system and week ($F(1,34)=0.037$, $p>.05$). There was no 3-way interaction between age, system and week ($F(1,34)=1.807$, $p>.05$).

6.3.1.3. OVERALL IMAGE INTERFERENCE

The outcome from collating the selection of target images regardless of account is the elimination of the main effect of age. This suggests that the binding of code and account was more of an issue for the older group than for the younger group.

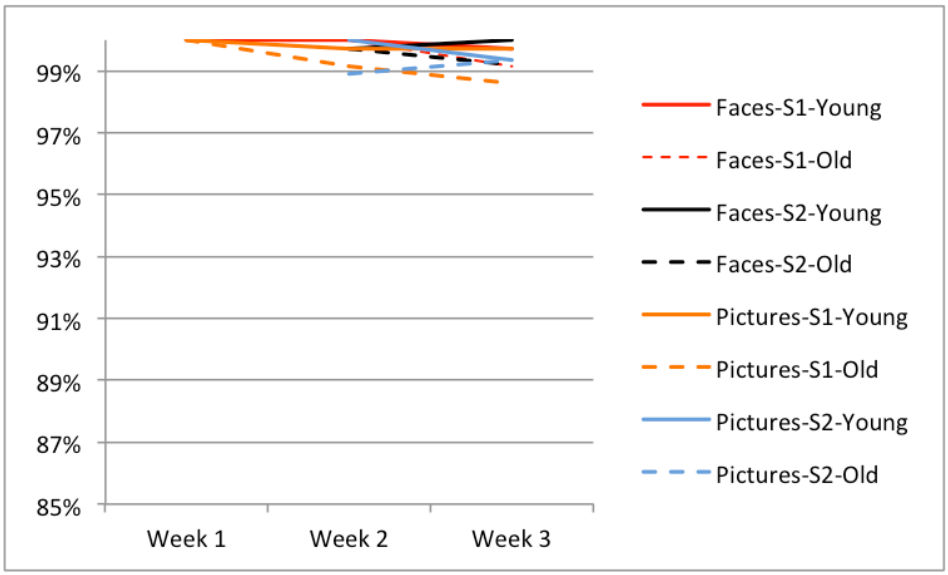


Figure 6.15: Overview of overall image interference for both sets of images (percentage correct).

In fact, upon closer inspection of the study overview (Figure 6.15) it is clear that participants are generally performing at ceiling (max score = 20) – both younger and older adults. This resulted in no effects being present – more importantly a main effect of age was not found. The fact that older adults are performing at ceiling is a good indication that graphical systems have the potential to improve the authentication experience for this age group, but that the design of the systems needs to be reviewed carefully.

6.3.2. Order of Acquisition Effects

Forgotten codes were further explored – i.e. those codes that participants were totally unable to recall after 5 attempts – as a function of order of acquisition. The underlying question was whether the order of acquisition of the code was reflected in the rate of forgetting. This was mapped out as a function of load.

A chi square test using participant age and forgotten codes as factors was carried out on the data. A significant association between participant age and forgetting of codes was found when remembering 4 codes, $\chi^2(1)=6.424$, $p<.05$. This seems to represent the fact that older participants were more likely to forget codes than younger participants (see Figure 6.16).

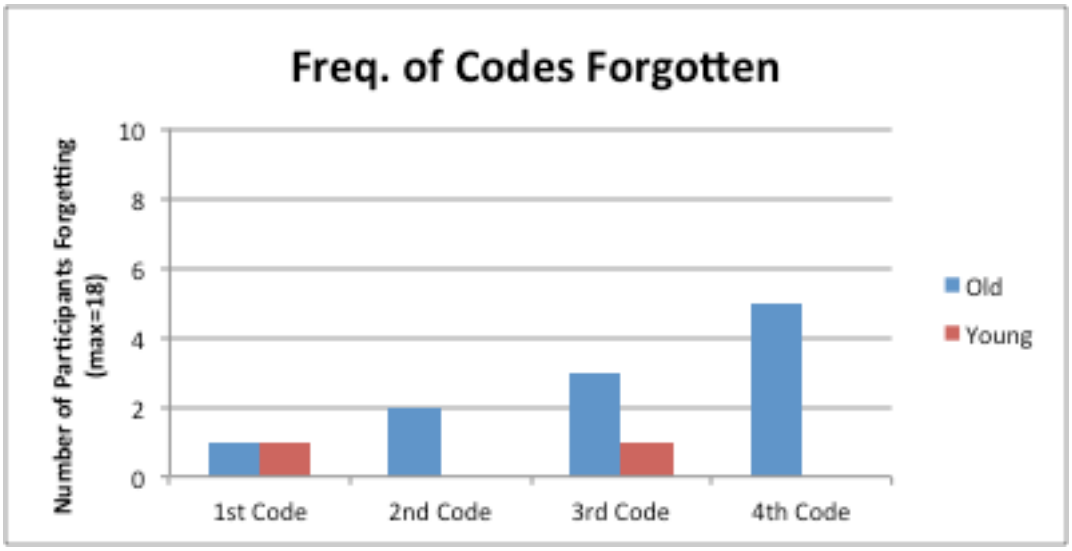


Figure 6.16: Forgetting of image codes for both younger and older participants.

Separate chi square tests were run on the younger and older participants with order of acquisition and forgetting. The test for younger participants found that order of acquisition was not significantly associated with the forgetting of codes, $\chi^2(3)=2.019$, $p>.05$. The test for older participants also revealed no association between order of acquisition and forgetting of codes, $\chi^2(3)=3.353$, $p>.05$.

6.4. DISCUSSION

Overall, a main effect of age was present for both Faces and Pictures where the younger group performed both more accurately and faster than the older group. This age effect was in accordance with the prediction that was made prior to the study. However, of more interest was the finding that older adults appeared to benefit from the Faces system more than with the Pictures system, even if no main effect of system was found. Older participants were more accurate with Faces with SET 1 codes, however, when further codes were added (SET 2) that advantage with Faces disappeared. Despite no main effect of system, there is some evidence to suggest that a face-based system could be the key to inclusive authentication after a number of improvements. It is possible that the older group were disadvantaged by the faces that were used as stimuli in this study which means the advantage of Faces could be even greater. There is evidence of an own-age effect when it comes to recognising faces, meaning that older adults tend to recognise older faces better than younger faces (e.g. Lamont et al., 2005). With this in mind, it was expected that if a difference in performance was to be found it would have been in favour of Pictures.

A possible explanation for the superiority of face-bases stimuli is that the participants processed the faces holistically (Bruce & Young, 1986), whereas Pictures were remembered by picking an odd detail from the main image. There is evidence that older adults adopt poor memorisation strategies when compared with younger adults (Naveh-Benjamin et al., 2007) and it is possible that these subpar strategies led to confusion regarding the Pictures, but not the Faces which were encoded as a whole. Additionally, it is possible that the poor strategy adopted by some older participants led them to create a very basic representation of the image that was then confused by the similar foils (Koutstaal & Schacter, 1997). The familiarisation questions aimed to neutralise these possible effect by asking participants similar questions that required them to think about past experiences, but it is unknown how much this helped.

Another possible explanation is that older users were confused by the similar images (Koustaal & Schacter, 1997). The Pictures were grouped in categories, and each grid consisted of images from the same category. Literature on categorical perception indicates that images from different categories appear more different than items from the same category (Harnad, 1987). Therefore, discriminations can be made faster between images that cross a category boundary than between two images that belong to the same category (Bornstein, 1987). In the case of Pictures, this discrimination was made more difficult due to the lack of category boundaries. Faces, on the other hand, were not grouped into categories, potentially making the choices less challenging. However, it could be argued that all faces belonged to the same category, with the images being of young males so this would indicate that older adults are more vulnerable to confusion of similar images than they are to confusion of similar faces.

A decline in performance over time – both in successful attempts and time taken to select the images – was observed for all participants with both SET 1 and SET 2 codes. The time-based performance decrement observed for both age groups is supported by previous research into multiple graphical authentication systems. In accordance with Moncur and Le Plâtre (2007) and Chiasson et al. (2009), a performance decrement was observed after a one-week delay when participants had multiple graphical systems to remember. However, an age-specific performance decrement over time was present with the new codes, meaning that older participants' accuracy declined significantly from week 2 to week 3 while that of younger participants did not. This interaction effect

makes it clear that the systems were not wholly inclusive of older adults. This finding which supports previous work once again adds credibility to the paradigm that was used. In fact, the use of the paradigm was the reason why the drop off experienced by the older group with the SET 2 codes could be observed, and the reason why the advantage older adults had with Faces – in the form of an age x system interaction with SET 1 – was discovered.

As expected, younger participants were significantly faster than older participants in selecting their images. This was the case across all conditions and presents no surprises. The speed results demonstrate that there is no accuracy-speed tradeoff and that younger participants are clearly superior than older participants.

6.4.1. Implications & Improvements

The older group were penalised more than the younger group by the image overlap that was implemented in this study. This was highlighted with the Image Interference analysis which showed that in terms of seen/not seen recognition of the images, older participants performed comparably to the younger participants. This finding is not entirely surprising due to the Associative-Deficit Hypothesis (ADH) (Naveh-Benjamin, 2000). ADH has shown that older adults are disadvantaged when creating associations between items. This problem has been shown to affect images as well (Naveh-Benjamin, 2003) and it would not be far-fetched to think that faces would be vulnerable to this as well. What the ADH demonstrates is that when remembering a single item (e.g. a picture), older adults show no difference from younger adults when asked to recognise it at a later time. This is also the case with multiple single items. However, when items are paired (e.g. two pictures that must be remembered together) older adults are seen to perform significantly worse than younger adults.

The ADH explains the main problem that older participants faced. The question remains as to whether the problem was with associating the four images to form a code, or associating a code with the right account. Based on the limited data available it seems the problem was in associating the four images and forming them into an account.

The results from this study suggest that the two GAS designed for this study were not optimal for older users in their current state, penalising them over time and making it

difficult to remember multiple sets of images. When comparing participants' performance with GAS to the PIN benchmark it becomes clear that Faces and Pictures present a marked improvement. If we consider the performance benchmark established with PIN for the older group of 2.6 average successful attempts for SET 1 and 2.7 average successful attempts for SET 2, the GAS accuracy results of 3.6/4.2 (Pictures/Faces) with SET 1 codes and an accuracy of 3.2/3.6 with SET 2 codes show an improved performance for both SET 1 and SET 2. It should be noted that the loads presented here not identical: four PINs had to be remembered while two of each Pictures and Faces had to be remembered. However, by merging the two GAS to create a load of four it can be seen that the advantage over PIN is still present, although it is debatable whether a load of four consisting of two Faces and two Pictures is equivalent to a load of four PINs.

In this study it would be expected that Caucasian participants would perform better than non-Caucasian participants (Shepherd, Deregowski, & Ellis, 1974). The participants that were recruited for this study were all Caucasian so no adverse effects regarding recognition were expected, but it does highlight the problems that individual differences incite. Females have been shown to benefit from an own-gender effect when recognising female faces (e.g. Cellerino, Borghetti, & Sartucci, 2004; Lovén, Herlitz, & Rehnman, 2011) but a similar effect has not been found for male participants. This means that the use of male faces could disadvantage female participants. As such, it is possible that the choice of faces that were used for the Faces system may have negatively impacted the performance of older participants, given the majority of the participant sample was female. Despite the possible disadvantage, the Faces system still presented the best overall performance for the older adult group. This further confirms that face-based systems have the potential to improve the authentication experience of older adults. A suggestion to improve the Faces system would be to use old faces as stimuli rather than young faces – previous research suggests that own-age effects are present for both younger and older adults (e.g. Lamont et al., 2003) meaning that older adults would be expected to recognise old faces more accurately than young faces. However, it is unknown what impact this would have on the younger group but it would be interesting to find out whether they can maintain their performance with the old faces.

Another suggestion to eliminate the interference of multiple graphical systems is to guarantee the absence of overlapping images and make the system a seen/not seen recognition task. Although this approach could have security implications, these will not be explored in this paper. However, such an approach would help both younger and older users take advantage of the Picture Superiority Effect. The feasibility of such an approach with current graphical systems is questionable, although potentially achievable if the images are distributed centrally (e.g. by an organisation). The distributor would then be able to assign specific sets for each provider (e.g. categories of boats and phone booths for NatWest Bank) and prevent other providers from using the same sets. This would eliminate the image overlap, unless users obtained multiple accounts with the same provider.

6.4.2. Conclusions

Overall, it can be seen that interference is clearly present for both graphical systems. This interference appears to affect both sets of images equally and older adults are affected more for the second set of images. This suggests that older users do not cope as well as younger users with the added memory load of more images.

Our results support previous research into multiple graphical authentication system interference showing that the memorability of the system drops off over time when more than one system has to be remembered. This effect was even more pronounced for the older group than for the younger group with SET 2 codes.

The results suggest that face-based systems may be the best solution for older adults, although improvements are still required to design a fully inclusive system.

6.5. Chapter Summary

This chapter set out to investigate how older adults performed using graphical authentication systems. It was thought that older users would match the performance of the younger users with the graphical authentications systems due to literature demonstrating that visual memory is not affected as much as other memory with age. Overall, the main effect of age was always present – i.e. older adults performed relatively poorly when compared to younger adults. However, after a more subtle

analysis taking system and week of testing into account showed that older adults' performance with the face-based system was good in the early stages (week 1, SET 1) but became progressively worse as new sets of faces were added and as the delay between encoding and testing grew. The drop off effect was particularly striking for SET 2 codes, with a particular emphasis on the Pictures system for the older group.

7. IMPROVING FACE-BASED AUTHENTICATION FOR OLDER ADULTS: YOUNGFACES AND OLDFACES

7.1. RATIONALE

In the previous chapter it was demonstrated that younger participants were significantly more accurate when selecting their graphical codes than their older counterparts. The face-based system proved to elicit better overall performance with the older group but design improvements could still be made to create an inclusive system. This chapter will look at what we know about face recognition and implement those facts and theories on a new system that will be evaluated with younger and older adults.

The advantages of facial recognition with familiar faces are well known (e.g. Pike, Kemp, & Brace, 2000). This is an effect that is present for both younger and older adults (Smith & Winograd, 1978). Research has also found that adults appear to be more adept when recognising faces from their age group than when recognising faces from other ages. This phenomenon – present for younger adults, older adults and even children (Anastasi & Rhodes, 2005) – is commonly referred to as the own-age effect. A large number of studies have reported results that support the existence of the own-age effect, but they have not always been successful in consistently finding the effects for all age groups. For example, Lamont et al. (2005) found that older adults could recognise old faces better than young faces, but younger adults did not benefit from the age of the face. Bäckman (1991) on the other hand found that younger adults recognised young faces better than old faces, and that young-older adults (>74 years old) recognised old faces better while old-older adults (75+ years old) appeared to have no preference about the age of the face. Fulton and Bartlett (1991) found that younger adults performed better with young faces, while older adults did not benefit from the age of the face. Wiese, Schweinberger, and Hansen (2008) conducted an ERP study that found that younger adults exhibited the effect, but not older adults. Harrison and Hole (2009) add further evidence to the own-age effect, but also found that exposure to the particular age range is a factor of the effect. In an experiment with primary school teachers and controls, they found that the own-age effect was present for the controls when asked to recognise faces of children but the effect was not present for the teachers, who have had extended contact with children.

Despite the discrepancies across the literature, a recent meta-analysis by Rhodes and Anastasi (2012) concluded that the own-age effect was a robust effect where people were more accurate at recognising faces of their own age. Additionally, they concluded that experience plays a role in the ability to recognise faces of other ages – similar to Harrison and Hole (2009) – but add that recency is an important factor. Simply, even though older adults have prior experience recognising young faces in the past, they still exhibit an own-age effect due to those experiences taking place a long time ago.

The purpose of this chapter is to compare the performance of younger and older adults with two face-based systems with the aim of determining whether the use of an older face-based authentication system might improve the performance for older adults without sacrificing the performance of younger adults. Two factors are explored in this experiment: a.) the effect that face age has on participants' memorability and b.) the consequence of eliminating the image interference. The literature suggests that older adults should be better at recognising old faces over young faces, while it is not certain what the impact might be on younger adults. The interference of the codes was removed based on the previous study to eliminate the need to bind the faces together, a problem that has been demonstrated amongst the older adult population (e.g. Naveh-Benjamin, 2003). The overall number of codes to be assigned to participants was set to six to determine whether participants are able to remember a relatively high number of codes under the improved conditions.

It was expected that younger adults would be more accurate than older adults when remembering multiple codes while the performance of the older group was expected to be better than the Faces benchmark when using OldFaces. This prediction was based on the own-age literature (e.g. Lamont et al., 2009). It was also expected that accuracy would decline over time based on past evaluations using multiple GAS codes (e.g. Everitt et al., 2009) but no age-specific effects were predicted.

7.2. METHOD

7.2.1. EXPERIMENTAL DESIGN

The study consisted of two factorial designs as the codes were separated into 'original' accounts (SET 1) assigned to participants during the first week and tested in weeks 1, 2 and 3, and 'new' accounts (SET 2) assigned to participants during the second week and tested in weeks 2 and 3. The first set of images were tested in a 2 (participant age: young, old) x2 (face age: young, old) x3 (week of testing: week 1, 2, 3) factorial mixed design. The second set of images were tested in a 2 (participant age: young, old) x2 (face age: young, old) x2 (week of testing: week 2, 3) factorial mixed design. There were therefore two independent factors – the participants' age (young group, old group) and the age of the faces (under 30 years old, over 50 years old) – and one repeated factor – the testing week (weeks 1, 2, 3).

Dependent factors comprised the average number of successful authentication attempts (maximum of 5 per account) and the average time (in seconds) taken to select the four faces in a code.

7.2.2. PARTICIPANTS

72 participants were recruited in total to fit into four groups: older groups learning old faces (participant age ranging 65-75, n=18), older group learning young faces (participant age range 67-75, n=18), younger group learning old faces (participant age range 18-30, n=18) and younger group learning young faces (participant age range 18-30, n=18).

Younger participants were recruited from the student population in the university with mean age of 19 (SD: 1.29) years – mean age 19 (SD: 1.61) for younger faces and mean age 20 (SD: 0.86) for older faces. Given the diversity of the student population we considered this sample adequate. Younger participants were recruited using an online participation pool maintained by the university.

Older participants, with a mean age of 70 years (SD: 3.79) – mean age 70 (SD: 3.60) for the younger faces and mean age 71 (SD: 3.68) for the older faces – were recruited using

the lab's participant database as well as through an advert on the Elders' Council of Newcastle newsletter. All participants were given £20 to cover travel expenses to and from the university.

Participants were screened for age and for computer experience—they were required to have used a computer prior to taking part in the study. Participants were also screened for adequate vision and for prosopagnosia (a face recognition deficit).

7.2.3. MATERIALS

The materials of the study were designed to reproduce a real life authentication system and the experience of logging into multiple accounts. Six account names were created for the study. The names used were Bank, Library, TV, Phone, Shop and Email. Each account was allocated eight faces – four younger and four older –that participants would be required to learn and remember over the course of the study. Participants only had to learn four faces per account, all either young or old, depending on their assigned condition.

The face-based graphical authentication system used for the study was built using Experiment Builder v 1.5.201. Initially the system displayed the simplified instructions for participants (see Figure 7.1) and once participants were ready they pressed the spacebar to start. Four sequential challenge grids were displayed upon the participants' selection. Once four selections were made, the system informed participants whether they had selected correctly or incorrectly. The system carried out the cycle four more iterations, meaning participants selected their codes five times in total. All participants were tested in the same room and same computer to guarantee the same experience.

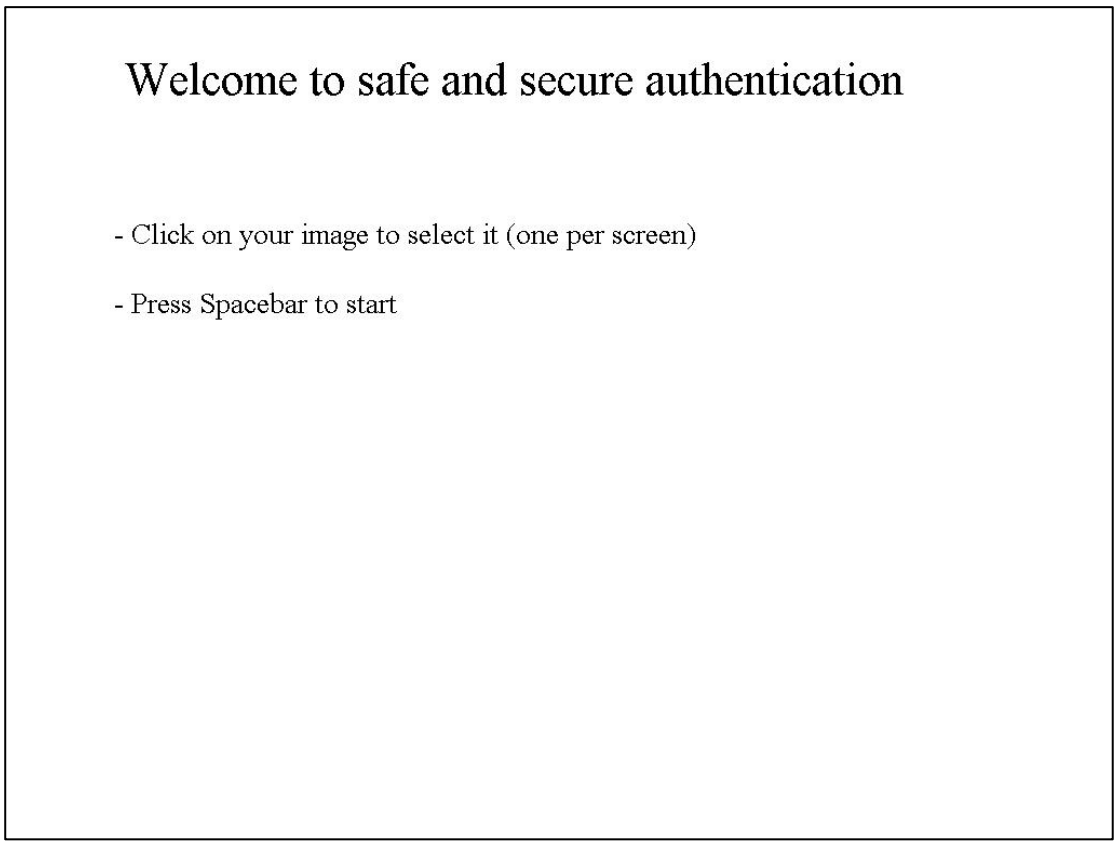


Figure 7.1: Simplified instructions to participants.

7.2.3.1. GRID COMPOSITION FOR YOUNG FACES

The young faces were obtained from a university smartcard database (same faces as Chapter 6) and the face pool consisted of 280 males, all numbered. 80 faces were randomly selected from the database using a random number generator and these would become the face bank for the study. Faces were renumbered 1 to 80. All young faces were in full colour and were converted to grey scale for consistency and for uniformity (see Older Faces, 7.2.3.2).

The 24 target young faces were selected randomly from the bank using a random number generator. The bank was then renumbered once again from 1 to 56. Each of the target faces was assigned 12 foil faces that would appear on the challenge grids. The foils were assigned partially randomly, meaning that the twelve faces were chosen using a random number generator, but were then inspected by the team to make sure there were no outstanding faces.

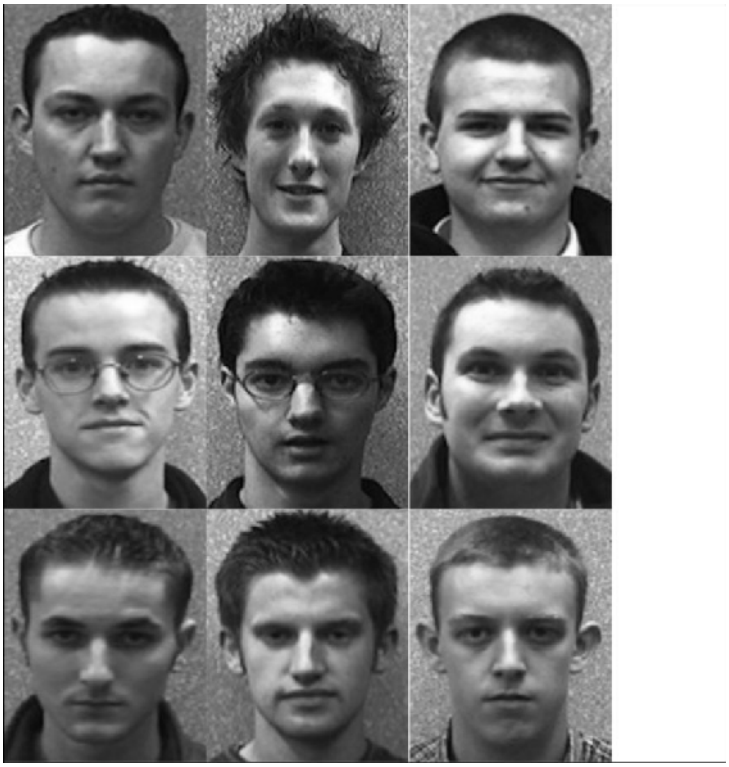


Figure 7.2: Sample grid for young faces.

Sixteen sequences of nine numbers were generated using the random number generator. These sequences would form the challenge grids that would be displayed to participants (for example see Figure 7.2). The position of the faces on the grid was determined by the order of the numbers in the sequence (where 1 is top left and 9 is bottom right). The computer program randomised the presentation of the grids for each participant (see Appendix A for sample grids).

7.2.3.2. GRID COMPOSITION FOR OLD FACES

Old faces were obtained from a number of online databases, including the Max Planck Institute (Ebner, Riediger & Lindenberger, 2010), Aberdeen, and Utrecht and consisted of 85 males. The Max Planck Institute database consisted of tagged faces of young adults, middle-aged adults, and old adults. Only the old faces were used for this study. The other databases were manually inspected for faces that appeared to be over 50 years old. The selected faces were then screened by a supervisor to confirm the perceived age of 50 or over. 80 faces formed the final old faces bank, and were all numbered from 1 to 80. The majority of the old faces were in full colour and were converted to grey scale for consistency and in order to mask any outstanding colour features (e.g. yellow shirt) and to neutralise differing backgrounds.

The 24 target old faces were randomly selected from the bank using a random number generator. The bank was then renumbered once again from 1 to 56. Each of the target faces was assigned 12 foil faces that would appear on the challenge grids. The foils were assigned partially randomly, meaning that the twelve faces were chosen using a random number generator, but were then inspected by the team to make sure no faces drew a disproportionate amount of attention from participants.

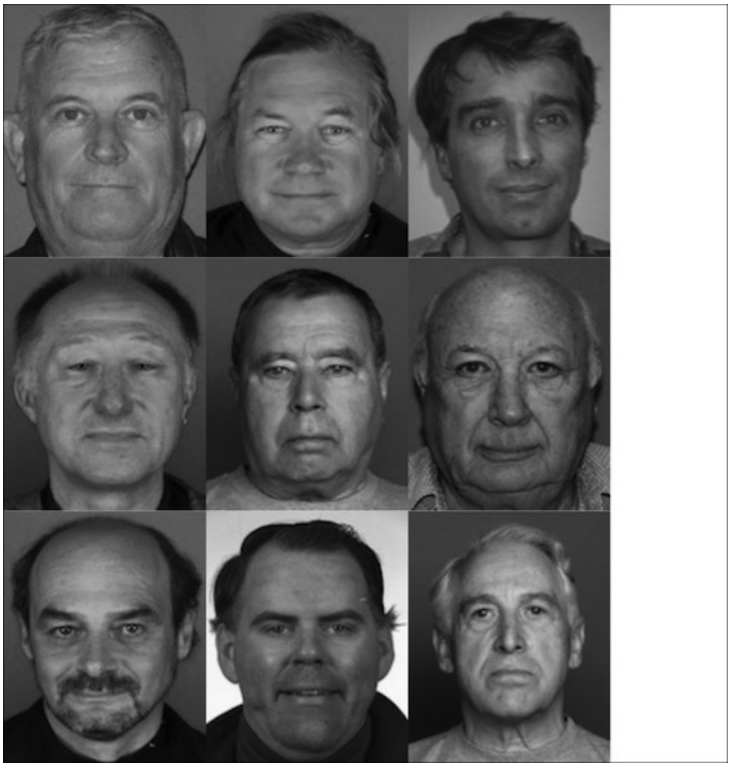


Figure 7.3: Sample grid for old faces.

The grids were constructed using the same sixteen sequences that were used for the young faces (see Figure 7.3 for example grid). Once again the grids were randomised by the computer program (see Appendix A for sample grids).

A digital voice recorder was also used to record the discussion during the first session.

7.2.3.3. IMAGE OVERLAP

It was apparent that one of the biggest issues for both the younger and older groups was the image overlap that was implemented for the Faces and Pictures systems for ecological validity (see Chapter 6). Closer analysis of the data revealed that when

participants were selecting incorrect faces, they generally belonged to one of the other assigned codes. Additionally, no age effects were found when the overlap was artificially removed. This was most likely due to age-related problems with binding items together, an effect known as the Associative-Deficit Hypothesis (Naveh-Benjamin et al., 2003). With this in mind, the image overlap was eliminated for this study. This meant that a target face for a code would not appear as a foil in any other code used in the study. The real world implementation implications are that a central organisation would be required to maintain and distribute the images in order to guarantee the absence of any overlapping images. While this may be difficult to achieve – especially as the system matures and becomes more widespread – it is a worthwhile approach if it is shown to improve the overall performance of older adults.

7.2.4. PROCEDURE

The procedure for this study consisted of two stages: the enrolment stage and the authentication stage. During enrolment, participants learned their faces. During authentication, participants attempted to access their 'accounts' by identifying the four correct faces that belonged to the account.

7.2.4.1. ENROLMENT STAGE + FAMILIARISATION

During the enrolment stage participants were given account names and target faces to learn. They were guided through a familiarisation process consisting of seven questions (see Table 7.1) to encourage the participant to associate the face with a person they knew, or to at least think of something memorable to do with that face. This was especially important for the older adults who rely on resemblance (Smith & Winograd, 1978) and pleasantness (Bartlett & Fulton, 1991) for recognition of faces. In total, the familiarisation process lasted for approximately five minutes per account (i.e. the four faces consisting of a code). Participants were given each face individually and after having 5 seconds to study the face they were asked the respective questions by the experimenter. The participant was required to reply to the questions at loud.

Table 7.1: Familiarisation questions used for both young and old faces.

Familiarisation Questions
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?

Once participants had been through the familiarisation process for the code, they were asked to select the segments that belong to their code from amongst foil faces to confirm they had learned them, essentially a mock authentication attempt (see 7.2.4.2 below). The practice session also allowed participants further time to learn the images in context (e.g. discriminate from foils). If the participant failed to select their faces correctly at least three attempts in a row they were asked to repeat the mock authentication. They were also shown the faces again if necessary. Participants were then taken through the same process for their next account.

7.2.4.2. AUTHENTICATION STAGE

During the authentication stage participants were presented with a login screen and asked to log into their accounts using the authentication codes. The login screen contained nine faces in a 3 x 3 grid and participants were required to select the face belonging to their code from the nine in four successive screens using the mouse. Once the participant selected four faces, the program told them whether they selected correctly or incorrectly—if incorrectly they were not told which faces they got wrong. Participants were required to do this for a total of five times per code regardless of whether they selected correctly or incorrectly.

7.2.4.3. PROCEDURE FOR PARTICIPANTS

Participants were asked to attend the lab on three separate occasions (see Table 7.2 for procedure overview). During the first session, participants were given their first set of

codes (SET 1) consisting of three random accounts. The order of the accounts was randomised in order to eliminate any order effects. Participants were taken through the enrolment stage with the three accounts. After enrolling with all accounts, participants were distracted from the encoding task by taking part in a short discussion with the investigator. This discussion focused on their experience of current authentication systems such as passwords and smartcards and lasted approximately 10 minutes.

Following the discussion, participants taken through the authentication stage with the order of the accounts randomised, meaning that it did not necessarily follow the same order as the enrolment.

Table 7.2: Overview of procedure for participants over the three-week period.

Week	**Activity**
1	1. Learn half the codes (SET 1) 2. Discussion (distractor) 3. Recall SET 1 codes
2 (+1 week)	1. Recall SET 1 codes 2. Learn other half of codes (SET 2) 3. Discussion (distractor) 4. Recall SET 2 codes
3 (+1 week)	• Recall SET 2 codes • Recall SET 1 codes 3. Discussion

Participants were asked to return to the lab a week after the first session. Upon their arrival they were greeted and were asked to once again log into their accounts using the authentication codes that were allocated in the previous week (SET 1). Once again the order of the accounts were randomised. Once participants finished logging in, they were enrolled with the remaining three codes (SET 2). Following the enrolment with their second set, participants were engaged in another discussion, this time regarding their thoughts about security threats. This discussion lasted for approximately 10 minutes. Upon the completion of the discussion, participants were asked to log into their

accounts using the new set of codes (SET 2). They were not asked to log into their first set of accounts.

Participants visited the lab for a final time one week after the second session. Upon their arrival they were greeted and were asked to log into the accounts they were assigned in the previous two sessions. The order of the sets was randomised, as well as the order of the accounts in each set. Once they had logged in with both sets of codes, participants were asked questions about their experience and about their strategies for remembering the faces.

The total amount of time taken to complete the study when taking into account the three sessions was approximately 120 minutes for older participants and 80 minutes for the younger participants.

7.3. RESULTS

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. Variables measured were number of successful attempts and average time taken to select the four faces making up a code. For a table of means see Appendix B.

The average number of successful attempts measured the number of times participants selected all four faces correctly in an attempt – to a maximum of five times (per account) – averaged across the total number of codes per week. The accuracy results for Set 1 and Set 2 are presented separately below.

The average time to authenticate, recorded in seconds, measured the average duration of an attempt between the presentation of the first face and the selection of the fourth face. The time results for Set 1 and Set 2 are presented separately below.

7.3.1. SUCCESSFUL ATTEMPTS- ACCURACY

7.3.1.1. SET 1 – ORIGINAL CODES

Participants' scores (max=5) for each of the three codes that made up SET 1 were collated for each of the three weeks resulting in a mean score for each week. A 2 (participant age: young, old) x2 (system: Faces, Pictures) x3 (week of testing: week 1, 2, 3) mixed factorial ANOVA was carried out.

For SET 1 codes, 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). No main effect of face age type was found ($F(1,68)=.773$, $p>.05$). A main effect of week was present ($F(2,67)=8.059$, $p=.001$). Pairwise comparisons showed that participants achieved significantly more successful attempts in the first week (mean: 4.60) compared to both the second week (mean: 4.36) ($p<.05$) and the third week (mean: 4.16) ($p<.001$). No significant difference in accuracy was present between week 2 and week 3.

No interaction was found between participant age and face age ($F(1, 68)=1.757$, $p>.05$), but there was a significant interaction effect between participant age and week ($F(2, 67)=8.783$, $p<.001$). 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) ($F(2,107)=0.031$, $p>.05$). 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) ($F(2,107)=3.597$, $p>.05$) (see Figure 7.4). No significant difference was found between week 1 (mean: 4.39) and week 2 (mean: 3.90) or between week 2 and week 3 (mean: 3.47), but a significant difference was found between week 1 and week 3 ($p>.05$).

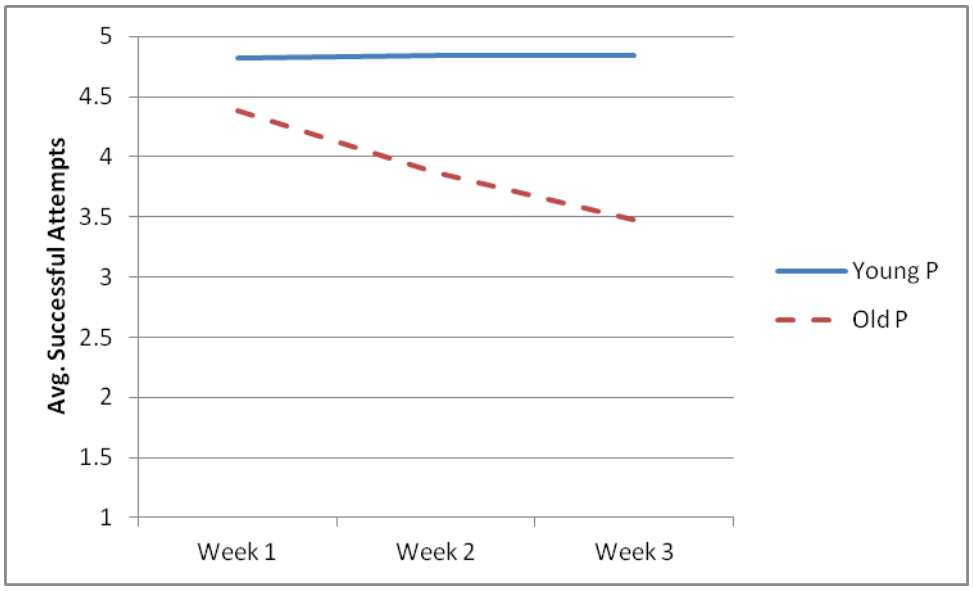


Figure 7.4: Interaction between participant age and week for SET 1.

There was no interaction effect found between face age and week ($F(2,67)=0.538$, $p>.05$). There was no significant 3-way interaction between participant age, face age and week ($F(2,33)=0.026$, $p>.05$).

7.3.1.2. SET 2 – NEW CODES

Participants' scores (max=5) for each of the three codes that made up SET 2 were collated for each of the two weeks resulting in a mean score for each week. A 2 (participant age: young, old) x2 (system: Faces, Pictures) x2 (week of testing: week 2, 3) mixed factorial ANOVA was carried out.

For SET 2 codes, a main effect of participant age was found ($F(1,68)=1092.447$, $p<.001$) with younger participants (mean: 4.71) achieving more successful attempts than older participants (mean: 3.60). No main effect of face age type was found ($F(1,68)=3.568$, $p>.05$). A main effect of week was present ($F(1,68)=14.216$, $p<.001$) with participants achieving more successful attempts in the second week (mean: 4.38) compared to the third week (mean: 3.92) ($p<.001$).

An interaction effect was found between participant age and face age ($F(1,68)=12.642$, $p<.05$) 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 7.5).

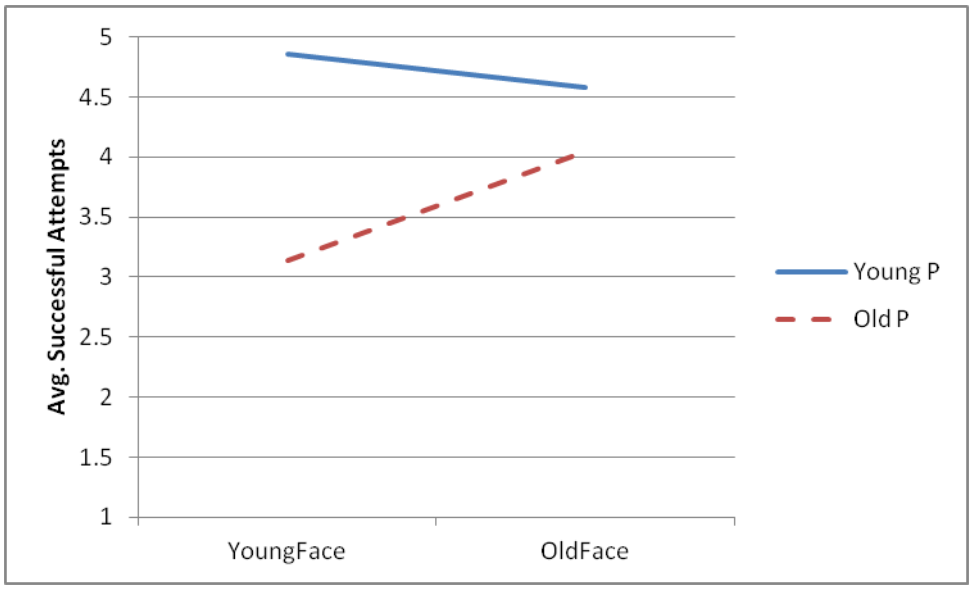


Figure 7.5: Interaction between Participant Age and Face Age for SET 2.

There were no interaction effects between participant age and week ($F(1,68)=3.844$, $p>.05$) or between face age and week ($F(1,68)=0.688$, $p>.05$).

There was a 3-way interaction between participant age, face age and week ($F(1,68)=4.146$, $p<.05$) where the accuracy with young faces was both best for the younger group but worst for the older group (see Figure 7.6).

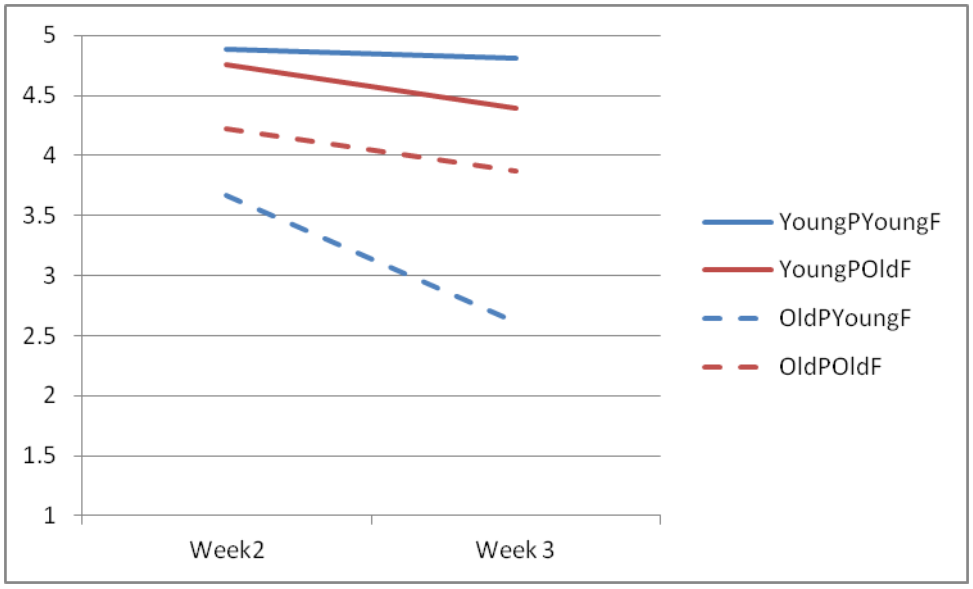


Figure 7.6: Three-way interaction between Participant Age, Face Age and Week for SET 2.

An independent samples t-test shows no significant difference in performance between the younger and older groups when using old faces in the second week ($t(34)=1.844$, $p>.05$), while another independent samples t-test shows no significant difference in

performance between young faces and old faces for the older group during week 2 (t(34)=-215, p>.05). An independent samples t-test confirms no significant age effects in accuracy with old faces in week 3 (t(106)=1.758, p>.05). Another independent samples t-test shows no significant differences between young faces and old faces for the younger group in week three (t(34)=1.484, p>.05). A final independent samples t-test shows a significant difference in accuracy between young faces and old faces for older participants (t(34)=-2.237, p<.05).

7.3.1.3. OVERVIEW OF ACCURACY

An important observation was the decline in accuracy over time, especially for the older group. An age-specific decline was present with the original codes where accuracy decreased at a faster rate for older participants. However, no age-specific declines were found for the new codes, although Figure 7.7 shows clearly that the performance of older adults by week 3 was markedly poorer than for younger adults.

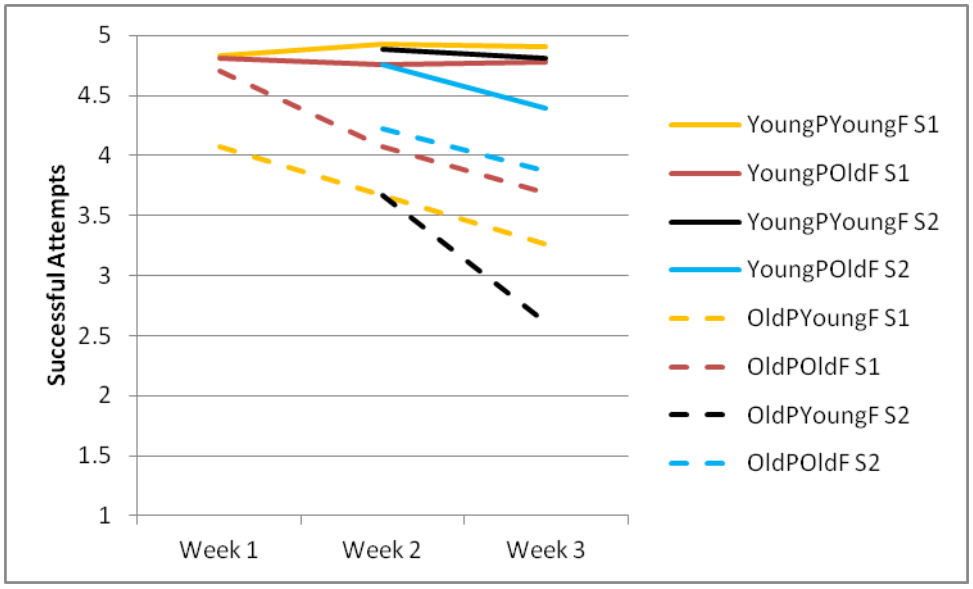


Figure 7.7: Overview of average successful attempts for both the younger and older groups using old and young faces.

Perhaps the most interesting finding was not the presence of a significant difference, but the lack of one: An independent samples t-test carried out on the accuracy of younger and older participants with OldFaces during week three revealed no significant difference. The non-significant result suggests that OldFaces may be a step in the right direction for inclusive authentication systems and will be addressed in the discussion.

7.3.2. AVERAGE TIME - SPEED

7.3.2.1. SET 1 – ORIGINAL CODES

Participants' average selection time for each of the three codes that made up SET 1 were collated for each of the three weeks resulting in a mean score for each week. A 2 (participant age: young, old) x2 (system: Faces, Pictures) x3 (week of testing: week 1, 2, 3) mixed factorial ANOVA was carried out.

For SET 1 codes, a main effect of participant age was found ($F(1,68)=56.184$, $p<.001$) with younger participants (mean: 9.53 seconds) selecting their faces faster than older participants (mean: 18.14 seconds). No main effect of face age was found ($F(1, 68)=0.320$, $p>.05$). No main effect of week was present ($F(2,67)=0.991$, $p>.05$).

No interaction was found between participant age and face age ($F(1, 68)=0.011$, $p>.05$), but there was an interaction effect present between participant age and week ($F(2,67)=5.539$, $p<.05$). A one-way ANOVA showed no significant difference in speed for the younger group between weeks 1, 2, and 3 while another one-way ANOVA showed no significant difference in speed for the older group between weeks 1, 2, and 3 (see Figure 7.8). There was no interaction between week and face age ($F(2, 67)=0.484$, $p>.05$). In other words, no speed-accuracy trade-off was observed for SET 1 codes.

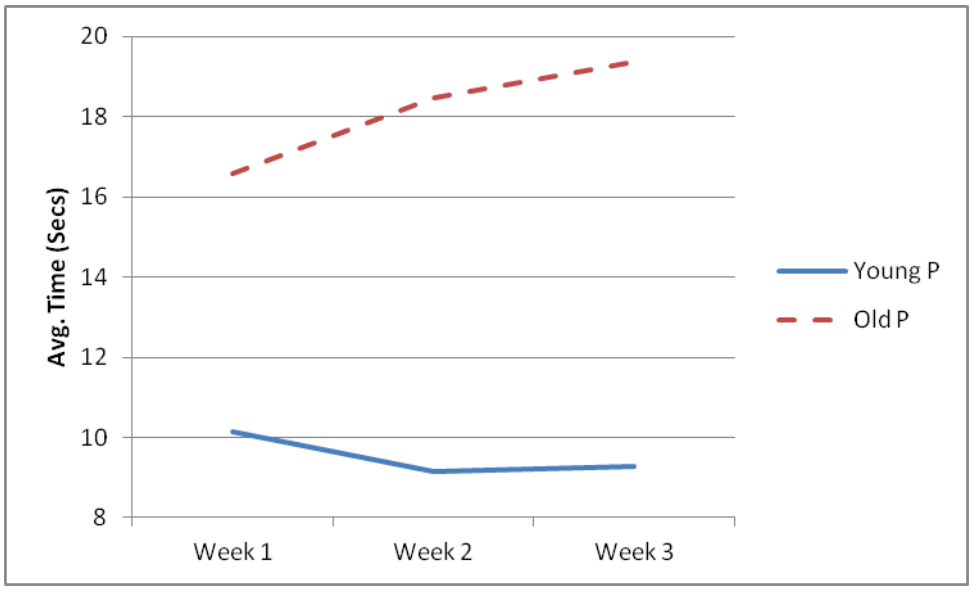


Figure 7.8: Interaction between Participant Age and Week for SET 1.

There was no 3-way interaction between participant age, face age and week ($F(2,67)=1.563$, $p>.05$).

7.3.2.2. SET 2 – NEW CODES

Participants' average selection time for each of the three codes that made up SET 2 were collated for each of the two weeks resulting in a mean score for each week. A 2 (participant age: young, old) x2 (system: Faces, Pictures) x2 (week of testing: 2, 3) mixed factorial ANOVA was carried out.

For SET 2 codes, a main effect of participant age was found ($F(1,68)=45.44$, $p<.001$) with younger participants (mean: 10.97 seconds) selecting their faces faster than older participants (mean: 19.37 seconds). No main effect of face age was found ($F(1, 68)=0.000$, $p>.05$). No main effect of week was present ($F(2,67)=3.029$, $p>.05$).

No interactions were found between participant age and face age ($F(1, 68)=0.443$, $p>.05$), week and participant age ($F(1,68)=2.796$, $p>.05$), or week and face age ($F(1, 68)=0.392$, $p>.05$). In other words, no speed-accuracy trade-off was observed for SET 2 codes.

There was no 3-way interaction between participant age, face age and week ($F(1,68)=1.217$, $p>.05$).

7.3.3. Order of Acquisition Effects

Forgotten codes were further explored – i.e. those codes that participants were totally unable to recall after 5 attempts – as a function of order of acquisition. The underlying question was whether the order of acquisition of the code was reflected in the rate of forgetting. This was mapped out as a function of load.

A chi square test using participant age and forgotten codes as factors was carried out on the data. A significant association between participant age and forgetting of codes was found when remembering both the young and old face (young: $\chi^2(1)=40.876$, $p<.001$; old: $\chi^2(1)=7.920$, $p=.005$). This seems to represent the fact that older participants were more likely to forget codes than younger participants.

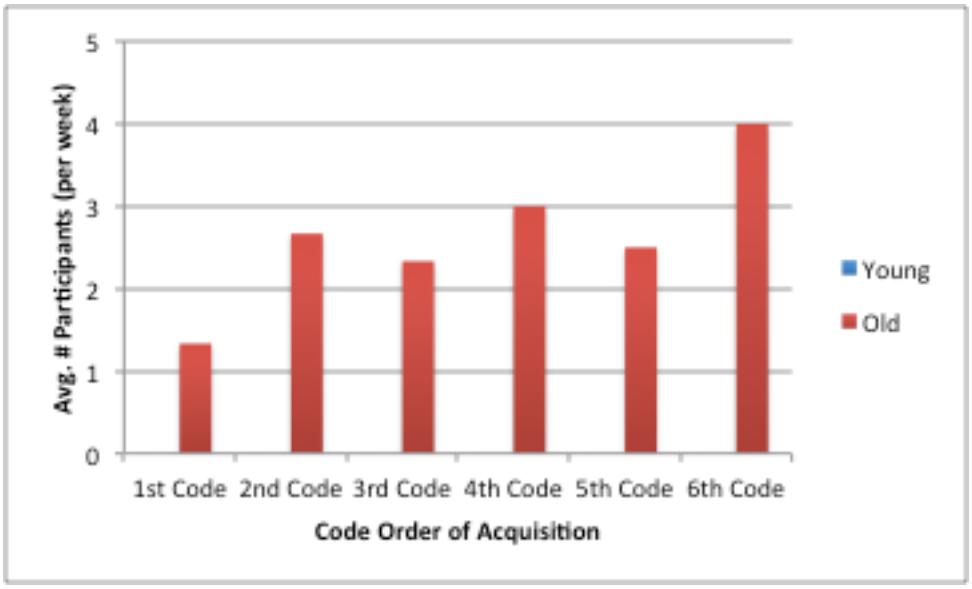


Figure 7.9: Forgetting of YoungFaces codes for both younger and older participants (NOTE: no codes forgotten by the younger group).

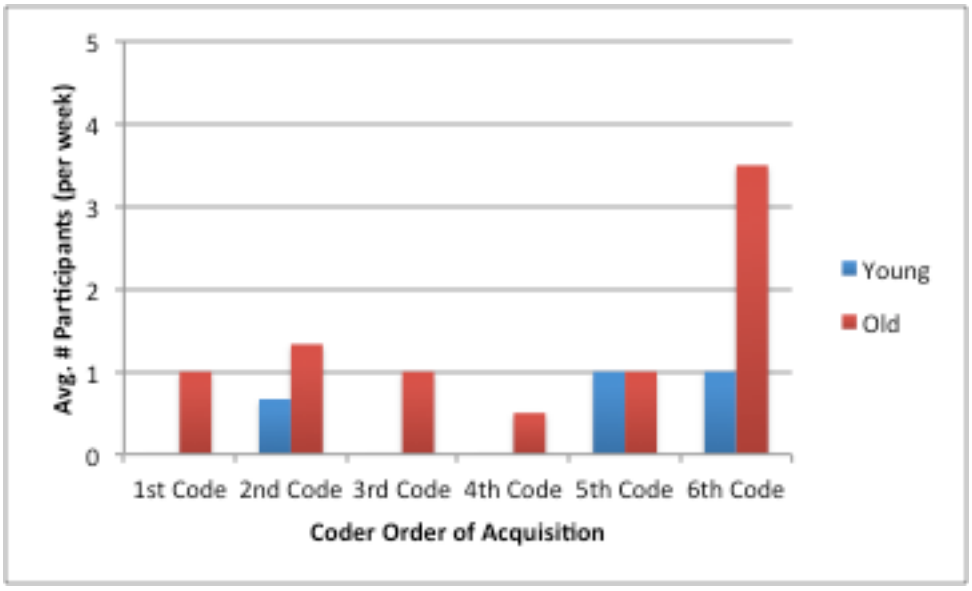


Figure 7.10: Forgetting of OldFaces codes for both younger and older participants.

Separate chi square tests were run on the younger and older participants with order of acquisition and forgetting. All tests, for both young (see Figure 7.9) and old faces (see Figure 7.10), were not statistically significant suggesting that the subsequent addition of codes did not impact on their ability to remember those codes in the long term.

7.4. DISCUSSION

Overall, a main effect of age was present for both young and old faces where the younger group performed both more accurately and faster than the older group. This finding was in accordance to the predictions made. An interaction effect was found with SET 2 codes which demonstrated the own-age effect – the older group were able to take advantage of the old faces but the younger group did not show a preference for the age of the faces. When looking specifically at the accuracy of participants with the old faces during the final week the age effects were eliminated, meaning that no significant difference was found between the accuracy of the younger and older groups. The older group was observed to be more accurate with the old faces while the younger group did not appear to benefit from the age of the face. This finding supports the hypothesis based on the literature that age-specific faces would be easier to recognise for older adults.

In the ageing literature it is common for older adults to underperform when having to learn and recognise new faces (e.g. Bartlett & Fulton, 1991). In this study older adults

were asked to learn and recognise a total of 24 faces that they had never seen before. Accuracy was good in comparison to the systems that were previously evaluated, with a low number of older adults forgetting the faces. This therefore contradicts previous research in the area and demonstrates that older adults are capable of learning new faces to a good standard. Participants were subjected to a familiarisation stage where they were encouraged to create associations with the faces and it is possible that this was the reason for the improved performance. Future research should look at the potential benefit of the familiarisation process as the older group did not excel with YoungFaces and as such the value of the process is unknown.

Based on the face recognition literature, it was predicted that an own-age effect would be present where older adults would perform better with old faces when compared with young faces. The results support a number of studies (e.g. Anastasi and Rhodes, 2005; Backman, 1991; Lamont et al., 2005) that demonstrate the own-age effect where older adults are more adept at recognising faces that match their age range, and further develops the area by confirming that the effect is present in an authentication context. An own-age effect for younger adults that was predicted by part of the literature (e.g. Fulton & Bartlett, 1991; Wiese et al., 2008) was not found in this study, further dividing the area with regards to the consistency of results across the lifespan.

Anastasi and Rhodes (2005) argue that extended exposure to own-age faces (Harrison & Hole, 2009) is unlikely to be the sole reason for the effect as if this was the case then older adults should be adept at recognising faces from any other age groups due to previous exposure. Instead they use the in-group/out-group model of face processing (IOM) developed by Sporer (1991) to justify the results. The model suggests that faces from the in-group – own age faces – are processed automatically, while those in the out-group are generally classified as out-group faces and further processing is either omitted or dependent on additional processing resources which may always be available. This is a satisfactory explanation for the results when ratings of faces are taken into consideration. Research by Ebner (2008) found that old faces are universally rated as being less attractive and less distinctive than young faces. Given that distinctiveness is thought to play a chief role in face recognition (Bruce et al., 1994; Sarno & Alley, 1997; Wickham & Morris, 2003) it would be expected that participants, regardless of age, would perform more poorly with the old faces. This was not the case, further supporting the IOM.

A decline in accuracy over time was observed for all participants with both sets of codes. The time-based performance decrement observed for both age groups is supported by previous research into multiple graphical authentication systems. In accordance with Moncur and Le Plâtre (2007) and Chiasson et al. (2009), a performance decrement was observed after a one-week delay when participants had multiple graphical systems to remember. The interaction effect found between age and week with the SET 1 codes showed a disadvantage for the older group where their accuracy declined at a significant rate from week 1 to week 3 while the accuracy of the younger group did not. However, this effect was not present with SET 2 codes, suggesting that the addition of codes affects the existing codes but does not have any negative repercussions on SET 2 codes. With an ideal system no accuracy declines would be present with the addition of new codes – for either SET 1 or SET 2 – but this result is a step in the right direction. The additional finding of a non-significant difference in accuracy between younger and older participants with old faces in week 3 further demonstrates the potential of the OldFaces system.

With regards to time taken to select the faces, it was interesting that younger participants appeared to be getting quicker at selecting their faces over time while older participants got slower – the same pattern that was observed in the accuracy data. While the overall finding – that the younger group selected their faces faster than the older group – was in accordance with the hypothesis, it was unexpected that the younger group improved their time-based performance.

7.4.1. IMPLICATIONS

The findings from this study have enormous implications for the future design of graphical authentication systems. This study has shown that an age-related performance decrement in older adults could be greatly reduced if the system is designed appropriately. Using old faces and eliminating the face binding between accounts helped in achieving this breakthrough. It is interesting that the elimination of the face binding by itself did not lead to comparative performances between both age groups as first thought based on the previous GAS experiment (see Chapter 6). The results from the older group using young faces demonstrate that an age effect was present for this

face type even with the removal of the overlap, therefore the use of old faces can be seen as a major contributing factor for the performance increase in the older group.

The results suggest that as new codes are added, the accuracy between age groups with old faces is maintained over time for the new codes – i.e. no significant differences in accuracy are found between the two age groups. It would be interesting to see whether this comparable performance between age groups is maintained over extended delays, and whether this effect is also present with the original codes over an even longer delay. Even then, these results are very encouraging, demonstrating that bridging the performance gap between age groups is possible and highlighting the fact that authentication systems that are inclusive of older users can be designed. One must be careful as it appears that the younger group were getting quicker over time while the older group were getting slower, possibly showing the effects of the added cognitive load. However, this observed effect was not specific to the old faces, with the means showing a time increase for both age groups generally.

The performance of the younger group has to be considered when thinking about this type of system as a solution to the older adults' problems. Some might argue that younger adults are being penalised in order to improve the performance of older adults, therefore not making the system inclusive but rather targeted to older adults. However, it was found that the performance difference between young faces and old faces for the younger group for SET 2 codes was not statistically significant. When looking at the means, the difference in performance between the two types of faces was minimal, and it is appropriate to conclude that the difference was not practically significant either. Hence, although the younger group performed slightly better with the young faces, they were not significantly disadvantaged by the old faces. This suggests that the use of old faces for GAS is a step in the right direction to designing an inclusive solution to authentication.

7.4.2. Future Improvements

A way to improve the system and the performance of users is to implement a larger face bank. A common concern during the debrief interviews, especially amongst older participants, was the recurrence of a large number of foils across accounts. It can be argued that regular exposure to the same foil faces – both within and across codes –

could lead to the encoding and future confusion with those faces. Participants mentioned that this was a concern when selecting the later codes during the final week as they would have been exposed continuously in a short period of time. This can perhaps explain the slowing down of the older group to an extent – older adults were having to discriminate foils while the younger group were able to select the face the stood out to them. Past work has suggested that foil faces are not learned after repeated exposure (Valentine, 1999). However, this research was carried out with younger adults, so it is possible that older adults are more vulnerable to this phenomenon. Madden (1983) demonstrated that older adults are more distracted by known stimuli when performing a visual search task after a day's delay. In this case the stimuli used were letters rather than pictures or faces, but the possibility exists that this learning of foils can be extended to the visual domain.

The inclusion of female faces should also be considered for future improvement to the system. Males were used in this study due to the unavailability of sufficient female faces, so it is possible that the incorporation of female faces could improve the performance of female users without necessarily affecting the performance of male users (Cellerino et al., 2004; Lovén et al., 2011). The inclusion of female faces would further help in increasing the size and variability of the face bank but gender-specific effects have to be considered. The literature in this area is relatively new, and as such it is possible that effects are present that we are currently not aware of.

Researchers have also found a reliable own-race effect when it comes to face recognition as confirmed by a meta-analysis by Meissner and Brigham (2001). This means that people are better able to recognise a face from their own race later on than a face from another race. Valentine and Bruce (1986) first presented evidence of an own-race effect by carrying out an experiment where participants were required to recognise inverted faces, both of their own race and other races. Recognition of the own race faces was superior as expected. Levin (2000) suggested that this effect is present due to participants encoding race-specific features at the expense of individualistic features and thus were penalised at the point of recognition. Walker and Tanaka (2003) then suggested that the effect occurs during the encoding of the faces. Michel, Rossion, Han, Chung, and Caldara (2006) support previous findings by proposing that own-race faces are encoded more holistically than other-race faces. Hence, it is important to consider

this limitation when designing face-based graphical systems as users could be easily disadvantaged.

7.5. CHAPTER SUMMARY

This chapter set out to evaluate a new face-based graphical authentication system, OldFaces, with the aim of bridging the performance gap between younger and older adults. It was predicted that the use of old faces as stimuli would improve the performance of the older group. In accordance with facial recognition literature, an own-age effect was found where younger participants performed better with young faces and older participants performed better with old faces. Younger participants outperformed older participants in both successful attempts and time with both YoungFaces and OldFaces – however, upon closer inspection, no age effects were found with SET 2 codes during the final week of the study, suggesting that OldFaces could be the key to inclusive user authentication. Finally, over the course of the three weeks, performance with SET 1 codes was subject to a performance decrement for both age groups. Meanwhile, SET 2 codes did not present any time-related performance decrements.

The following chapter will discuss the findings from all four experiments and will compare the results from this study with the PIN benchmark.

8. FINAL DISCUSSION

8.1. INTRODUCTION

Four studies were carried out using a similar methodology to explore any age differences in accuracy and time between younger and older adults with various authentication systems. Two factors remained constant across all four studies: the individual differences of participants (age) and the testing time intervals. This chapter presents an overall discussion of the work that has been undertaken. First the two research questions are revisited and answered using the results from the empirical studies. Next all authentication systems that were evaluated are directly compared. The contributions made by the thesis are then listed and put in the context of what was learned. The limitations of the thesis follow. Future work is explored before the final conclusions are made.

8.2. RESEARCH QUESTIONS

8.2.1. PERFORMANCE WITH CURRENT AUTHENTICATION SYSTEMS

The first research question that was posed was whether older adults were disadvantaged by existing authentication systems when compared with younger adults. The first study evaluated the performance of younger and older adults when learning and remembering multiple PINs over the course of three weeks. Two loads were implemented – a low load condition where participants were given four PIN codes to learn and a high load condition where participants were given six PIN codes to learn. Younger participants were shown to be significantly more accurate and quicker than older participants when entering their codes. The accuracy of the older participants was generally poor and at their best was borderline between logging in and getting locked out when using a traditional ‘three strikes’ policy.

Based on the evidence detailed above, the answer to this first research question is that older adults are disadvantaged by existing authentication systems when compared with younger adults, as demonstrated by the main effect of age. The low number of

successful attempts older participants had when recalling their codes – the PIN benchmark – further supports this notion. These findings support the survey-based results of Rasmussen and Rudmin (2010) who concluded that older adults had more difficulties in remembering their PINs when compared to younger adults. This was the first study to empirically demonstrate this problem in the context of authentication and to present benchmark data on the performance of younger and older adults with PINs. It was also the first study to directly compare the performance of older adults and younger adults when learning and recalling multiple PIN codes. The results from this study, coupled with the self-reported problems of older adults, make it clear that alternative forms of authentication need to be explored to find a solution to the current poor state of authentication.

8.2.2. Performance with Graphical Authentication Systems

The second research question that was posed was whether graphical authentication systems could improve the performance of older adults in relation to existing systems with the aim of being a more inclusive form of authentication. It should be noted that younger participants were consistently more accurate and faster than older participants, and as such only the findings in relation to the older group will be discussed in this subsection.

Three studies were carried out with the purpose of evaluating four distinct GAS with younger and older adult groups. The overall conclusion was that GAS have the potential to be used as inclusive authentication that do not overly penalise the accuracy of older adults over time. However, it is important to design the graphical systems carefully in order to avoid impacting the memorability of the systems, as witnessed in Study 2 (see Chapter 5). Tiles, the GAS designed using image segments yielded better accuracy than PIN for both younger and older participants during the first week of testing, but after a week's delay the accuracy dropped off significantly to the point where it was inferior to PIN.

The third study evaluated the performance of younger and older adults over three weeks with two GAS that were designed based on existing systems. A face-based system, Faces, was based on Passfaces (Valentine, 1998) and a picture-based system, Pictures, was based on VIP (De Angeli et al., 2002). Overall accuracy was better than with PIN,

but accuracy declined significantly between the second week and the third week suggesting a potential long-term memorability problem. Older participants showed an advantage with Faces, in terms accuracy and time taken to select the stimuli. The findings from this study demonstrated that overall performance could be improved by graphical systems that there was still room for improvement.

The fourth and final study evaluated and compared two face-based systems with younger and older adults over three weeks. One system, YoungFaces resembled the young Faces system used in the previous study, while the other, OldFaces, used faces of older adults instead of younger adults. The image overlap between codes was removed for both systems which benefited older participants. An age specific drop-off in accuracy was observed for the older group with SET 1, but that effect was mitigated by the use of old faces by SET 2. Overall, younger participants were found to be more accurate and quicker than older participants when selecting their codes overall, but it was observed that during the final week with the new codes there was no significant difference in accuracy between the two groups when using old faces. This finding suggests that it is possible to design inclusive authentication systems if done correctly.

Based on the evidence detailed above, GAS do have the potential to improve the performance of both younger and older adults when compared with existing authentication systems, but only if they are designed correctly. Study 2 demonstrated how the wrong design choice could result in the system being less memorable than PINs over time, but Study 4 also showed that with the right design choice the performance gap between age groups could be reduced. The challenge is in ascertaining what the correct design parameters are, but the final study goes some way to showing which factors play a part: the use of old faces along with a guaranteed image exclusivity between codes played a part in reducing the performance decrement observed in older adults with existing authentication systems and with other GAS. This design did not completely eradicate the decrement, but did improve upon the benchmark measure from the PIN system. Further work needs to be carried out to establish what other measures need to be taken into consideration to maximise the benefits of graphical systems for older adults.

It is important to be clear about the definition of an inclusive system. An inclusive system is one that maximises the performance of a range of groups. In this case, an

inclusive system is one that does not penalise the accuracy of older adults over time when remembering their authentication codes. With regards to that definition, the goal of inclusive authentication was achieved by the design of OldFaces. During the final week when using the new codes the accuracy of older adults was not significantly different from that of younger adults. Younger adults were not disadvantaged by OldFaces despite the apparent accuracy drop in the final week, as demonstrated by an independent samples t-test that found no significance in accuracy between YoungFaces and OldFaces during the last week. Additionally, older participants performed at a noticeably higher level than the PIN benchmark. In terms of comparing the performance between younger and older adults, it is debateable whether this main effect of age can ever be overcome by KBA – after all memory does decline with age and the accuracy of older adults is always expected to be below that of their younger counterparts. Realistically the best that can be hoped for is for a system that does not induce any age-specific declines in accuracy over time – in other words, the rate of the decline in accuracy is parallel between both age groups – and that does not produce high levels of forgetting resulting in large amounts of attempts for logging in (i.e. unlike PINs).

8.3. System Comparisons

This section directly compares the accuracy results from all four studies. The studies were carried out separately and therefore the comparison does not contain inferential statistics. However, an analysis is attempted with the two most similar designs, although the limitations are acknowledged. Comparisons are made on a system-by-system basis (i.e. PIN vs. Tiles, PIN vs. GAS, etc.). For all graphs, the red line indicates the first set of codes (original) and the black line indicates the second set of codes (new). The solid lines indicate the accuracy of the younger group while the dashed lines indicate the accuracy of the older group. Only the trends for the older adult group are described below as younger adults were consistently more accurate than their older counterparts.

8.3.1. Loads of 4 Codes

This subsection compares the authentication systems that were evaluated using four codes over three weeks. These were PIN low load condition, Tiles, and the GAS consisting of Faces and Pictures. The dependent measure for this subset of systems was successful attempts.

8.3.1.1. PIN VS. TILES

For the first set of codes, the accuracy of the older participants with Tiles during the first week was noticeably above the benchmark set by PIN. During the second week the accuracy dropped to similar standards as PIN, although still marginally higher. During the final week the accuracy of the older group was about the same as with PIN.

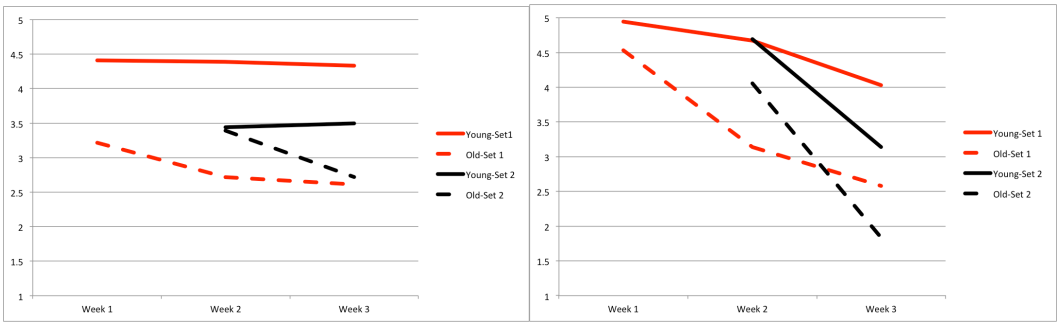


Figure 8.1: Comparison for PIN (left) and Tiles (right).

With the second set of codes, the accuracy during the second week was clearly better with the Tiles system. However, during the third week accuracy dropped alarmingly to the point where it was noticeable lower than the PIN benchmark.

This accuracy comparison between PIN and Tiles demonstrates how the graphical system improved the authentication process during a short ten-minute delay but after a long delay of one week that accuracy dropped to the same standards as PIN. These results suggest that Tiles may be a suitable authentication system for older adults if used consistently every day, as accuracy was noticeably higher after a ten-minute time delay. However, the potential dropoff after a week's delay may not be worth risking given the poor performance observed (e.g. after going on holiday for a week).

8.3.1.2. PIN VS. TRADITIONAL GAS

The accuracy of the older participants during the first week with the first set of codes appeared to be noticeably better with the traditional GAS – Faces and Pictures. During the second week the accuracy was maintained and was once again better than PIN – which also maintained the accuracy of the first week. During the final week the

accuracy of the older group fell noticeably, although still remained above the PIN benchmark.

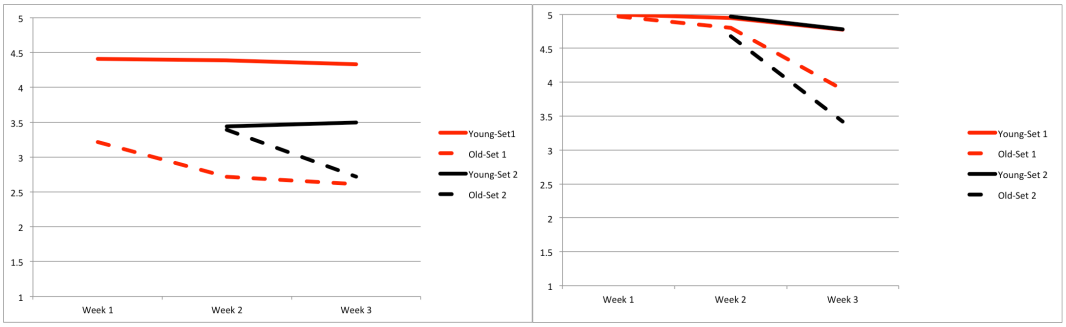


Figure 8.2: Comparison of PIN (left) and traditional GAS (right).

With the second set of codes, the accuracy during the second week was clearly better with the traditional GAS. During the third week the accuracy of the older group dropped noticeably, although it was still better than the start point in the second week with PIN.

This accuracy comparison between PIN and GAS demonstrates how the graphical system facilitated the authentication process for the older age group with both sets of codes. However, the age-specific decline in accuracy between weeks two and three with both sets of codes was very noticeable, even if accuracy remained better than with PIN. The question that arises is whether the rate of decline can be expected to continue at the same rate with more extended delays – e.g. two weeks – as if that proves to be the case then big problems can be anticipated for future code acquisitions with the graphical systems.

8.3.2. Loads of 6 Codes

This subsection compares the authentication systems that were evaluated using six codes over three weeks. These were PIN high load condition, YoungFaces, and OldFaces. The dependent measures for this subset of systems was average successful attempts.

8.3.2.1. PIN vs. YoungFaces

With the first set of codes, the accuracy of the older participants during the first week was better with the YoungFaces system than with PIN. During the second week there was a slight decline in accuracy, but once again accuracy was superior with

YoungFaces when compared with PIN for both age groups. During the final week the accuracy of the older group continued to decline steadily while remaining above the PIN benchmark.

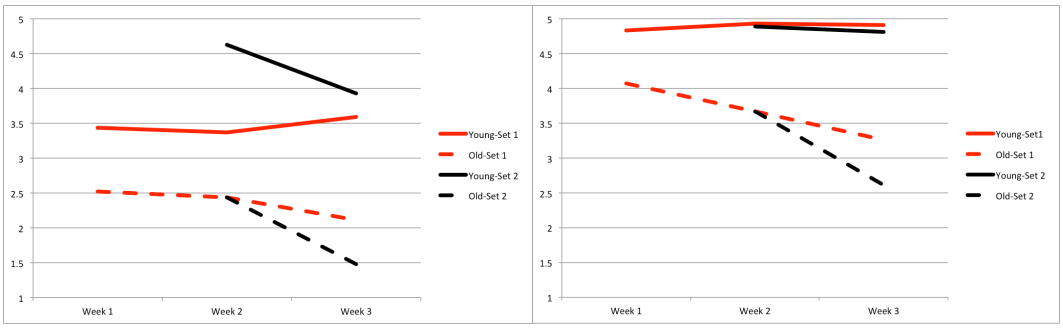


Figure 8.3: Comparison of PIN (left) and YoungFaees (right).

With the second set of codes, the accuracy during the second week was better with YoungFaces. During the third week the accuracy of the older group dropped noticeably, although remained better than the start point in the second week with PIN.

This accuracy comparison between PIN and YoungFaces demonstrates how the graphical system improved the authentication experience for the older adults with both the first and second sets of codes. However, the age-specific decline in accuracy between weeks two and three with the new codes for YoungFaces is noticeable, even if accuracy remains above the PIN benchmark.

8.3.2.2. PIN VS. OLDFACES

With the first set of codes, the accuracy of the older participants during the first week was better with the OldFaces system, to the extent where the older group were nearly as accurate as the younger group. During the second week there was a slight decline in accuracy for the older group, but once again performance remained above the PIN benchmark. During the final week the accuracy of the older group continued to decline steadily while remaining better than PIN.

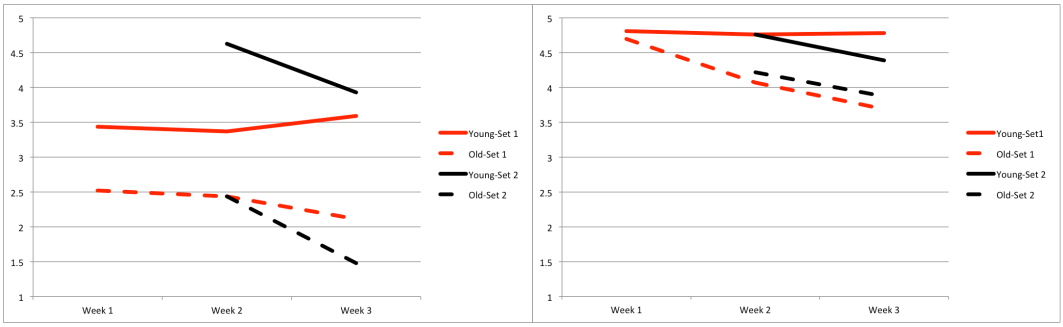


Figure 8.4: Comparison of PIN (left) and OldFaces (right).

With the second set of codes, the accuracy during the second week was better with OldFaces for the older group and once again nearly as accurate as the younger group. During the third week the accuracy of the older group dropped at a similar rate as that of the younger group while remaining noticeably higher than the PIN benchmark.

This accuracy comparison between PIN and OldFaces demonstrates how the graphical system improved the authentication experience for both age groups with both SET 1 and SET 2 codes. The similar drop in accuracy over time for both age groups is encouraging and at all stages the accuracy with OldFaces was better than with PIN. OldFaces yielded the closest accuracy performance between both age groups and as a consequence resulted in the best overall accuracy for the older group.

8.3.2.3. FURTHER ANALYSIS OF PIN VS. OLDFACES

A 3-way mixed ANOVA was carried out on the data from PIN and OldFaces (8.3.2.2). Only the scores from the high load for PIN were used in the analysis, resulting in a comparison between system that required users to learn and remember six codes throughout the three weeks. Two independent factors were used – participant age (young and old) and system (PIN, OldFaces) – while one repeated factor was used – week of testing (week 1, week 2, week 3 [SET 1 only]). It is acknowledged that analyses across separate studies can result in increased error due to the differing times of the year, slightly different recruitment strategy and the use of some participants across multiple studies. However, the two studies used the exact methodology in terms of factors and procedure, therefore the analysis was carried out with a stricter alpha level of .01.

For SET 1 codes, the analysis 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 OldFaces (mean: 4.47) than with PIN (mean: 2.91). No main effect of week was also present ($F(2,49)=3.655$, $p>.01$.

No interaction were found between age and system ($F(1,50)=.422$, $p>.01$), between system and week ($F(2,49)=1.370$, $p>.01$), or between age and week ($F(2,49)=4.195$, $p>.01$). No 3-way interaction between age, system and week was found ($F(2,49)=.828$, $p>.01$).

For the SET 2 codes, the analysis 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 ($F(1,50)=13.690$, $p=.001$) where participants were more accurate with OldFaces (mean: 4.31) than with PIN (mean: 3.12). A main effect of week was also present ($F(1,50)=15.271$, $p<.001$) with pairwise comparisons showing that participants were significantly less accurate during week 3 (mean: 3.42) than during week 2 (mean: 4.01) ($p<.001$).

An interaction was found between age and system ($F(1,50)=7.721$, $p<.01$) where the difference in accuracy between the two systems was not significant for the younger group (PIN: 4.28, OldFaces: 4.57) ($t(52)=-1.027$, $p>.01$) while the older group were significantly more accurate with OldFaces (mean: 4.05) when compared with PIN (mean: 1.96) ($t(52)=-4.987$, $p<.001$). No interaction was found between system and week ($F(1,50)=2.387$, $p>.01$) or between age and week ($F(1,50)=.155$, $p>.01$). No 3-way interaction between age, system and week was found ($F(1, 50)=.208$, $p>.01$).

These results show that while an overall age effect was present where younger participants were more accurate than older participants – as expected – a main effect of system was also present with both sets of codes where participants were more successful with OldFaces. This advantage was particularly effective for the older group with the second set of codes as demonstrated by the interaction between age and system. In essence, the analysis further suggest that a face-based GAS utilising old faces is an improvement over existing authentication systems for both age groups, but the older group benefitted to a greater degree.

8.3.3. Observed Trends

Two important trends were observed that spanned the four studies. A main effect of age was present for all systems and a main effect of week was present for all but one system. These trends are discussed in more detail below.

First, it should be noted that a main effect of age was observed throughout the testing of all systems. In every case, the younger group were shown to be superior to the older group. While it is not surprising to find that younger adults outperform older adults with memory-based authentication systems – both in terms of attempts and time – it is an important observation that demonstrates the magnitude of the challenge facing inclusive authentication for older adults. Age-related interactions have also been found in some of the studies, usually indicating a decline in accuracy over time for the older group when compared with the younger group. These interactions further disregard the inclusiveness of systems, observed for the GAS (see Chapter 6) with the second set of codes and for Faces (see Chapter 7) with the first set of codes.

Secondly, it should be noted that a main effect of week was observed throughout all studies with the exception of PIN. The drop in performance after a one-week delay does not come as a surprise, but it is important to note the consistency across systems. This main effect was not found for PIN where the overall performance was consistently poor. In summary, all graphical systems exhibited good performances in the first weeks that later decreased after a delay, while PIN exhibited poor performance from the start but it did not decrease with time. It is debateable what trend is most beneficial for an authentication system – one where performance is consistently below an acceptable threshold (i.e. PIN) or one where the initial performance is very good but after a delay it drops below an acceptable threshold (i.e. Tiles). In the context of authentication where the intervals between attempts cannot be anticipated a trend like Tiles is perhaps more desirable given the good short-term performance, while PIN consistently yields borderline – or worse – performance. However, neither trend is particularly desirable if avoidable.

Finally, it should be noted that no main effect of system manipulation was found throughout any of the standalone four studies. A number of interactions involving the

system manipulation were observed, but the main effects were usually masked by differing performances by the age groups. For example, with Tiles the younger group were more accurate with dissimilar grids while the older group was more accurate with the similar group – hence the two interactions nullified the main effect. Similarly, in Study 4 younger adults were more accurate with younger faces while older adults were more accurate with older faces, resulting in no main effect of face age. In Study 3 a slight advantage was found with Faces for the older group, but this was masked by a slight advantage for Pictures by the younger group. The one exception was PIN where the lack of a main effect of load was not masked by an interaction. This observation further demonstrates the difficulties that designers face when creating inclusive systems – one solution is not likely to benefit all parties.

8.4. CONTRIBUTIONS

This section lists the contributions that were made by this thesis and discusses what has been learned from each of the contributions.

1. This thesis presents the first performance comparison between younger and older adults with existing authentication systems – in this case PINs. PINs are one of the most common forms of knowledge-based authentication in use today, despite a vast psychology literature that predicts a long-term memory decline in older adults (e.g. Grady & Craik, 2000). Additionally, a previous self-reporting study exploring different strategies for remembering codes suggested that a problem may be present with regards to older adults remembering PINs (e.g. Rasmussen & Rudmin, 2010). Despite these clear indicators, no experimental evidence was available to validate the assumptions that PINs may disadvantage older adults. This thesis has shown that indeed older adults are disadvantaged in terms of accuracy and time when they have to remember multiple PIN codes. We have also learned that a load of six codes is not enough to induce a significant difference in accuracy rates with either younger or older adults, but that the accuracy rate with four codes is not acceptable for older adults.

2. This thesis includes the first performance comparison between younger and older adults with GAS. No other studies have used older adults as a factor when evaluating GAS. The majority of the authentication evaluation literature has utilised a restricted

population of students as participants who are not representative of older adults' requirements. Renaud and Ramsay (2007) did use older adults as participants in a field trial evaluation of Handwing but no younger adults were recruited for comparison purposes. Study 4 has shown that GAS have the potential to improve the performance of both age groups if designed correctly, but that they do not completely eliminate age effects – younger adults are always more accurate and faster than older adults. However, there is hope that the performance of older adults can be improved to a level where the age effects are no longer of a noticeable magnitude for the older group.

3. The thesis has contributed a testing methodology for evaluating the accuracy and time for authentication systems by controlling individual differences (age in this case), system differences, and time intervals (short and long). In the past various methodologies with inconsistent measurements have been used when evaluating authentication systems and as a consequence many of the findings from the studies cannot be compared. For example, some studies have compared new systems to existing systems (e.g. Dhamija & Perrig, 2000; Wiedenbeck et al., 2005), other have evaluated a single system (e.g. Brostoff & Sasse, 2000; Wiedenbeck et al., 2006), or tested variations of the same system (e.g. De Angeli et al., 2002; Chiasson et al., 2008). Other researchers have tested the short-term recall of systems (De Angeli et al., 2002; Tullis & Tedesco, 2005; Weidenbeck et al., 2006) while long-term memorability has also been evaluated (Brostoff et al., 2010; Wiedenbeck et al., 2005).

By using the proposed methodology it is possible to compare different systems, or variations of a system with various age groups – or possibly genders – and produce results that separate the codes in a way that can be examined for any specific performance declines.

8.5. LIMITATIONS

The methodology used throughout this thesis was designed to address the most important issues with other methodologies that have been used by researchers in the field – these included defined individual differences, constant time intervals and strategic comparison of systems. However, there are a number of limitations regarding the methodology used throughout the studies that will be discussed.

First of all the ecological validity of the administration of codes used throughout the studies can be debated. An effort was made in the design of the methodology to split the encoding of the codes to match the real life acquisition of codes. However, more than one code was assigned to participants in each 'sitting' which in itself is not usually representative of how users acquire codes – typically one at a time. The division of code sets makes the methodology more realistic than previous studies, but complete independence of the code assignment would have been ideal. This setup would be incredibly difficult to implement in a lab-based study due to the number of visits required by the participants. It has been shown to be possible in field studies, for example one of the conditions tested by Everitt et al. (2009), but then the lack of control that is associated with field studies has to be factored in – especially during the familiarisation process.

Secondly, in the studies that are part of this thesis participants were assigned the codes to learn and remember. In real life users would have the opportunity to change the codes – at least for PIN codes. Although it is possible to view this as a limitation, research suggests that a large number of users do not change their PINs when they are assigned (Bonneau et al., 2012; Rasmussen & Rudmin, 2010). Additionally, research has shown that codes could be more vulnerable to guessing attacks when they are selected by the users rather than when they are randomly assigned. In the context of text and numeric codes, participants have a tendency to select meaningful numbers such as birthdays (Bonneau, et al., 2012), challenge questions that are common knowledge (Just, 2005) and phrases that are well known (Keith et al., 2007), while with GAS users can also be predictable – e.g. a male user selecting female faces (Davis et al., 2004). With this in mind it is possible that the only realistic implementation method for GAS, from a security perspective, is to assign the codes to participants rather than allowing them to select their own codes, with a best case scenario seeing a user assigned two images and allowing them to select the other two. Hence, it could be argued that the results from the studies using the proposed methodology demonstrate the memorability of codes if 'best practice' recommendations are applied.

Another limitation with the design of this methodology lies in the use of absolute successful attempts (Chapter 5 and 6) as a measure of accuracy in lieu of average successful attempts as in study 1 (Chapter 4) and study 4 (Chapter 7). Absolute

successful attempts was used for the studies implementing a repeated measures design and testing participants with four codes overall – hence the week score for each set was based on a single code (single grid composition in Tiles, single Face/Picture in GAS) and susceptible to any issues associated with that specific code. On the other hand, average successful attempts relied on the data from more than one code per week, making the averaged score more reliable than the absolute score. Based on this observation, future studies implementing this methodology should be carried out evaluating different systems in an independent measures design rather than different configurations of the same system using a repeated measures design.

Other standard limitations also apply to this methodology, such as participant motivation. First of all, participants may have lacked the motivation to remember the codes over an extended period of time. Participants were asked to try and remember all codes, but were not offered further incentive to engage with the task. Although participants appeared keen on completing the tasks successfully, it is possible that other activities could have taken priority over the memory task. In real life participants have a strong incentive for remembering their codes to not be locked out of accounts so it could be argued that the results here are on the low-end of the performance spectrum. This is a standard problem that is present with the majority of usability studies and one that is very difficult to address. Performance-based financial incentives could be offered – i.e. £1 for every code correctly recalled during each session – but this would be in conflict with the ethics regulation of the university.

8.6. NEXT STEPS

The imperative question to be asked is whether GAS are the future for inclusive authentication. Part of the answer is that if they are designed correctly then they can positively influence the performance of older adults so they must be considered. However, it is also important to think about the implementation of such systems. Given hardware requirements, the most likely application for GAS appears to be online accounts: GAS cannot be set up on older hardware like alarm boxes that do not incorporate high quality screen, if any at all. GAS are unlikely to be a secure enough system for banks to implement as an alternative to PINs, and the same problem as alarm boxes would apply to ATMs and payment terminals – poor quality screens and

environmental factors such as glare from sunny days. It is possible, however, for GAS to be part of a multifactor authentication solution for online banking.

It should be noted that a large number of authentication codes are acquired from the online domain like shopping, email, social networking and forum accounts. It is not unreasonable to assume that if GAS are used for some internet accounts then the management of codes will improve. For example, if four online codes are graphical and a further four are passwords the load will have been halved. This was the initial motivation for graphical systems – to free memory for the management of other strong passwords – but the type of accounts that GAS would protect were never discussed.

An interesting question for future research is whether recognition for computer-generated faces is comparable to that of real faces. If this was found to be the case then the population of face banks would be less problematic. Currently, face banks are populated by faces of real people who have consented to be used for a commercial system. If GAS are to be widely used and image exclusivity is to be guaranteed, a very large bank will be needed. If a central organisation is allowed to generate artificial faces without impacting their memorability then exclusivity of images could become a reality.

Graphical authentication using old faces has been shown to benefit older adults while not penalising younger adults. However, it would be interesting to explore the possibility of personalised authentication with the aim of eliciting the best possible performance from each user group – the ultimate form of inclusive authentication. It has been established that own race effects are present in face recognition (e.g. Meissner and Brigham, 2001), as are own-age (e.g. Rhodes & Anastasi, 2012) and to an extent own-gender (e.g. Lovén et al., 2011). By knowing the gender, the age, and the race of the user, and tailoring the codes to that user (e.g. white Caucasian female undergraduate student receives four Caucasian female faces that are under 30 years old) an even better advantage could be achieved. This would be balanced by using foils that match the same criteria, but given the advantages described before the user should have no problems picking out the target faces. Additionally, this would make it more difficult for an attacker that does not fit the user's demographics to launch a successful attack (e.g. a young black male would be penalised by the own-race effect and would not be able to take advantage of the own-gender effect).

A further step for the personalised authentication model could be the use of a single face per account. The user would be assigned multiple poses and angles from that single face and would be required to select a subset from challenge screens when authenticating. Research has shown that identifying an unknown person in different pictures is very difficult, with the majority of participants thinking that the two pictures portrayed different people (Jenkins, White, Van Montfort, & Mike Burton, 2011). However, if the person in the picture is familiar, then the task was fairly straightforward. In this case the familiarisation process would be very important and the user would need to be shown all the possible variations of the face pictures. However, once familiarised the recognition should be much easier as the user would only be looking for one face, in effect. Additionally, this would make it easier to guarantee image exclusivity as only one face is assigned to a user per account meaning a more economical face distribution model.

8.7. Final Conclusions

The studies in this thesis have demonstrated that PINs, one of the most commonly used authentication systems, penalise older adults who have to remember multiple codes over extended periods of time more than younger adults. This performance decrement was suspected due to limited previous studies suggesting that this was the case (e.g. Rasmussen & Rudmin, 2010), but this is the first study to empirically demonstrate this decrement.

Several GAS were evaluated with the aim of improving the authentication experience for the older adult group. Results were mixed, with some systems proving to be detrimental to older adults while others proving to be beneficial. The main lesson learned was that the design of the graphical system could greatly influence the degree to which older adults managed to successfully remember their codes. Most importantly, the segmenting of images resulted in poor performance by the older group and the use of old faces improved the performance of older adults while not penalising younger adults. Work still needs to be done to eliminate the age effects – although younger adults are likely to always be more accurate and faster than older adults, but two key factors have already been identified for inclusiveness: own-age faces and the elimination of the image overlap.

Based on the studies conducted in this thesis, authentication with old faces is proposed as the best solution to inclusive authentication, with results having shown that the accuracy of both younger and older adults is comparable after a one-week delay. However, future work is encouraged on personalised authentication as it has the potential to benefit all age groups, races and genders if the literature is to be believed.

Appendices

APPENDIX A: SAMPLE GRIDS II

APPENDIX B – TABLE OF MEANS XIII

APPENDIX C – SPSS OUTPUTS XVI

APPENDIX D – PUBLISHED PAPERS II

Appendix A: Sample Grids

A selection of grids that were used for each study are presented here. Three grids from each condition are included, each from the same target for comparison. The original stimuli code is presented below each grid along with the grid number.

Study 2 (Chapter 5) – Tiles

BANK and NHS (Similar foils condition)

2.1 – Grid 1

2.1 – Grid 10

2.1 – Grid 15

EMAIL and SHOP (Dissimilar foils condition)

1.1 – Grid 1

1.1 – Grid 10

1.1 – Grid 15

Study 3 (Chapter 6) – Traditional Graphical Authentication Systems

Faces

K1-1 – Grid 1

K1-10 – Grid 10

K1-15 – Grid 15

Pictures

BolP1 – Grid 1

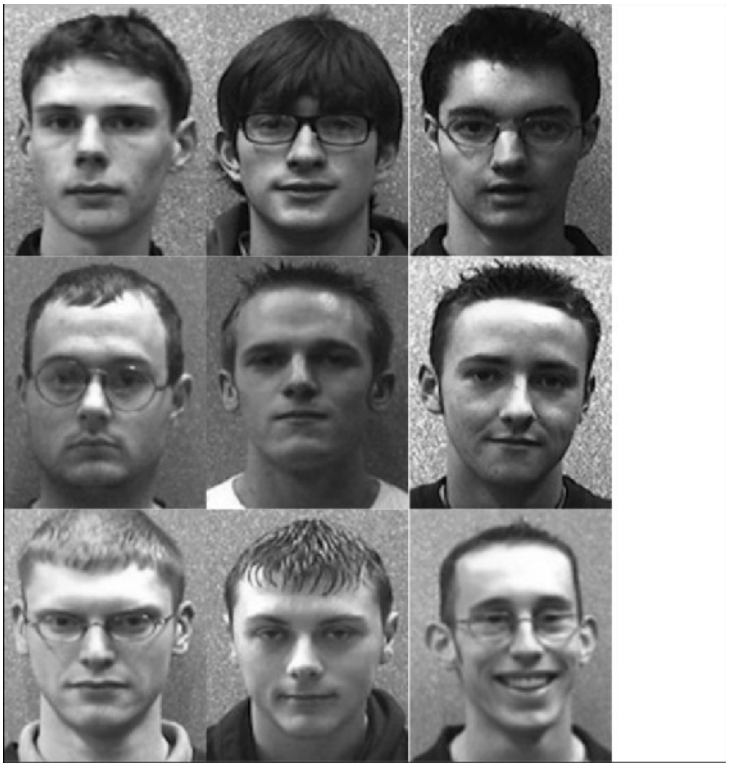
BolP10 – Grid 10

BolP15 – Grid 15

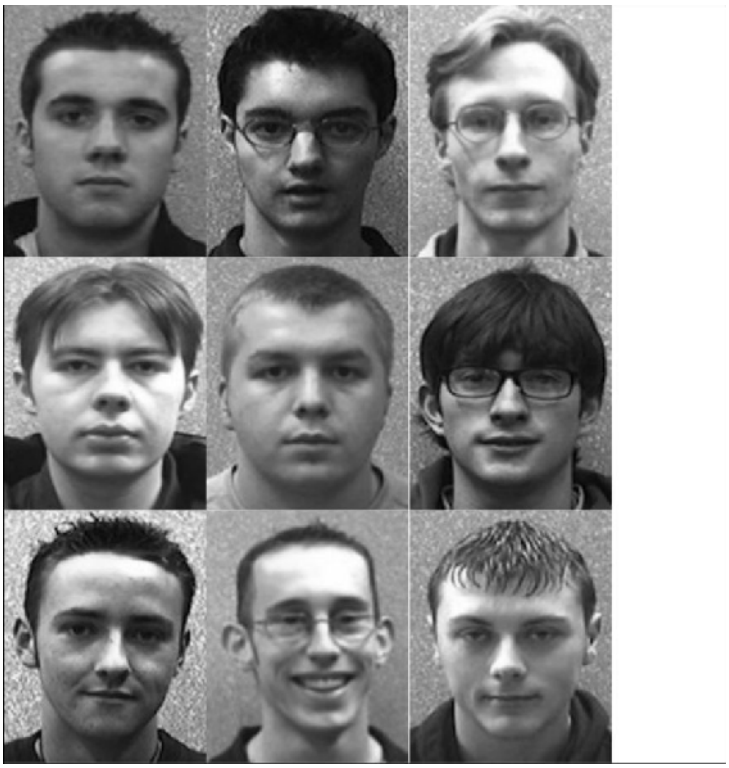
Study 4 (Chapter 7) – YoungFaces and OldFaces

YoungFaces

KMY1 – Grid 1

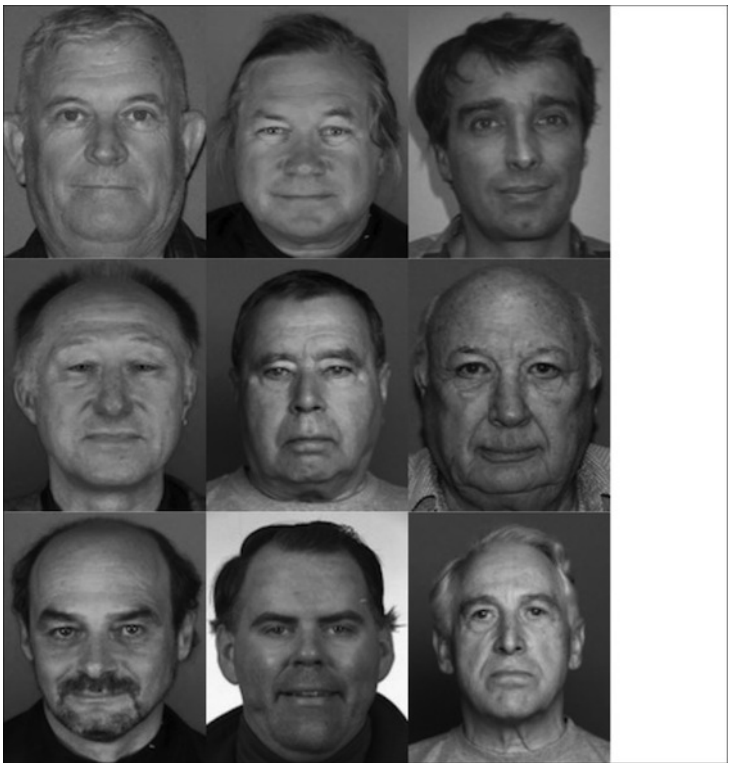


KMY1 – Grid 10

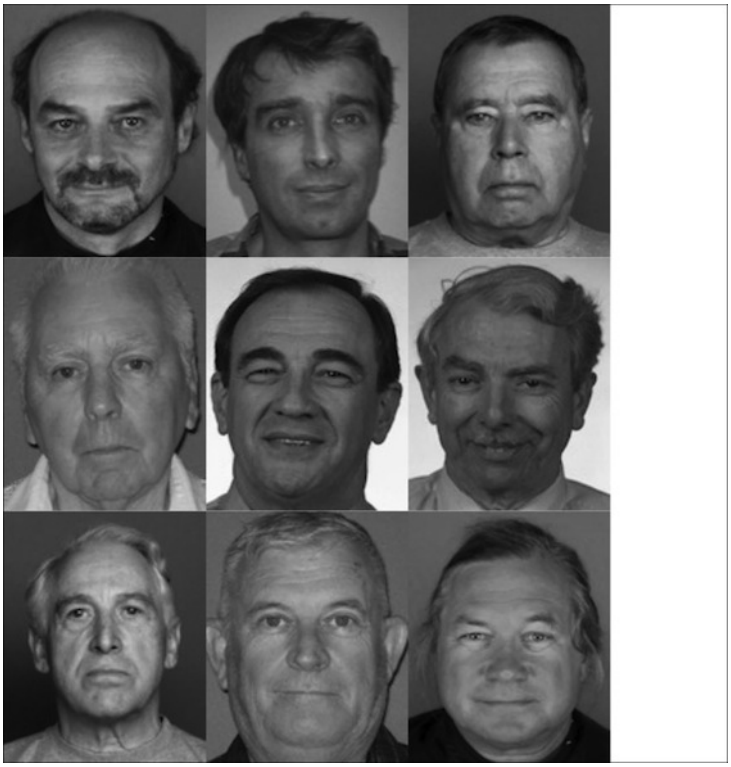


KMY1 – Grid 15

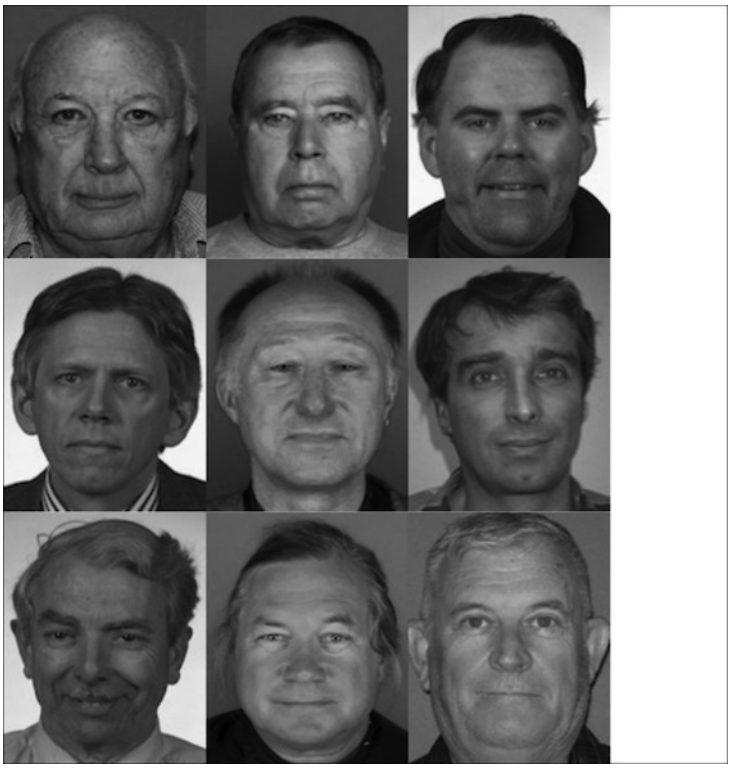
OldFaces



KMO1 – Grid 1



KMO1 – Grid 10



KMO1 – Grid 15

Appendix B – Table of Means

Study 1 (Chapter 4): PIN

SET 1

(average successful attempts)	Week 1			Week 2			Week 3		
	Young	Old	Total	Young	Old	Total	Young	Old	Total
Low Load	4.61 (0.99)	3.22 (2.43)	3.92 (1.94)	4.39 (1.65)	2.72 (2.27)	3.56 (2.11)	4.33 (1.64)	2.61 (2.16)	3.47 (2.06)
High Load	3.44 (1.97)	2.52 (2.04)	2.98 (2.00)	3.37 (2.01)	2.44 (2.15)	2.91 (2.07)	3.59 (1.86)	2.11 (1.89)	2.85 (1.97)
Total (Age)	4.02 (1.63)	2.87 (2.20)		3.88 (1.86)	2.58 (2.15)		3.96 (1.74)	2.36 (1.99)	
Total (Week)	3.45 (2.00)			3.23 (2.09)			3.16 (2.01)		

SET 2

(average successful attempts)	Week 2			Week 3		
	Young	Old	Total	Young	Old	Total
Low Load	3.44 (2.11)	3.39 (2.10)	3.42 (2.05)	3.50 (1.71)	2.72 (2.33)	3.11 (2.03)
High Load	4.63 (0.74)	2.44 (1.77)	3.54 (1.73)	3.93 (1.42)	1.48 (1.73)	2.70 (1.98)
Total (Age)	4.04 (1.65)	2.92 (1.95)		3.71 (1.54)	2.10 (2.09)	
Total (Week)	3.48 (1.87)			2.91 (1.99)		

Study 2 (Chapter 5) – Tiles

SET 1

	Similar Foils			Dissimilar Foils		
	Week 1	**Week 2**	**Week 3**	**Week 1**	**Week 2**	**Week 3**
Young Group N=18	4.89 (0.47)	4.78 (0.43)	3.78 (1.62)	5 (0)	4.56 (1.15)	4.28 (1.49)
Old Group N=18	4.39 (1.20)	3.11 (2.37)	2.44 (2.40)	4.67 (0.97)	3.17 (2.18)	2.72 (2.30)
Total	4.64 (0.93)	3.94 (1.88)	3.11 (2.11)	4.83 (0.70)	3.86 (1.85)	3.50 (2.06)

SET 2

	Similar Foils		Dissimilar Foils	
	Week 2	**Week 3**	**Week 2**	**Week 3**
Young N=18	4.83 (0.51)	2.39 (2.33)	4.56 (1.04)	3.89 (1.97)
Old N=18	3.78 (1.48)	2.28 (2.40)	4.33 (1.24)	1.39 (2.03)
Total	4.31 (1.22)	2.33 (2.33)	4.44 (1.13)	2.64 (2.34)

Study 3 (Chapter 6) – Traditional Graphical Authentication Systems

SET 1

	Faces			Pictures		
	Week 1	**Week 2**	**Week 3**	**Week 1**	**Week 2**	**Week 3**
Young Group n=18	5 (0)	5 (0)	4.61 (1.20)	5 (0)	4.89 (0.47)	4.94 (0.24)
Old Group n=18	5 (0)	4.94 (0.24)	4.17 (1.58)	4.94 (0.24)	4.67 (0.59)	3.61 (1.72)
Total	5 (0)	4.97 (0.17)	4.39 (1.40)	4.97 (0.17)	4.78 (0.54)	4.28 (1.39)

SET 2

	Faces		Pictures	
	Week 1	**Week 2**	**Week 1**	**Week 2**
Young n=18	4.94 (0.24)	4.67 (1.19)	5 (0)	4.89 (0.32)
Old n=18	4.83 (0.38)	3.67 (1.97)	4.83 (0.38)	3.17 (2.18)
Total	4.89 (0.32)	4.17 (1.68)	4.92 (0.28)	4.03 (1.76)

Study 4 (Chapter 7) – YoungFaces and OldFaces

SET 1

		Face Age		
Avg. Successful Attempts	Part. Age	Young Faces	Old Faces	Total (Week)
Week 1	Young	4.83 (0.24)	4.81 (0.35)	4.61 (0.75)
	Old	4.07 (1.15)	4.70 (0.68)	
	Total	4.45 (0.90)	4.76 (0.54)	
Week 2	Young	4.93 (0.14)	4.76 (0.55)	4.36 (1.19)
	Old	3.67 (1.61)	4.07 (1.39)	
	Total	4.30 (1.30)	4.41 (1.10)	
Week 3	Young	4.91 (0.25)	4.78 (0.50)	4.16 (1.45)
	Old	3.26 (2.03)	3.69 (1.50)	
	Total	4.08 (1.65)	4.23 (1.23)	

SET 2

		Face Age		
Avg. Successful Attempts	Part. Age	Young Faces	Old Faces	Total (Week)
Week 2	Young	4.89 (0.28)	4.76 (0.48)	4.38 (1.10)
	Old	3.67 (1.57)	4.22 (1.14)	
	Total	4.28 (1.27)	4.49 (0.90)	
Week 3	Young	4.81 (0.26)	4.39 (1.19)	3.92 (1.55)
	Old	2.61 (1.94)	3.87 (1.39)	
	Total	3.71 (1.77)	4.13 (1.30)	

Appendix C – SPSS Outputs

C1: Study 1 (Chapter 4) – PIN

C2: Study 2 (Chapter 5) – Tiles

C3: Study 3 (Chapter 6) – Traditional Graphical Authentication Systems

C4: Study 4 (Chapter 7) – YoungFaces and OldFaces

Accuracy output is presented first, followed by average time output. Analyses are split by Set, with SET 1 results being presented first, followed by SET 2.

C1: Study 1 (Chapter 4) – PIN

SET 1

[DataSetl] C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 2\Results\ St2-OVERALL.sav

Wit

hin-Subjects Factors Measure:MEASURE 1

Wee k	Dependent Variable
1	S1W1
2	S1 W2
3	S1 W3

Between-Subjects Factors

		Value Label	N
Age	1	Young	18
	2	Old	18
Load	1	Low	18
	2	High	18

Descriptive Statistics

Age		Load	Mean	Std. Deviation	N
Session 1	Young	Low	4.611111	**.9930313**	9
		High	3.444444	1.9649710	9
		Total	4.027778	1.6252199	18
	Old	Low	3.222222	2.4252720	9
		High	**2.518519**	**2.0351843**	9
		Total	2.870370	2.2018676	18
	Total	Low	3.916667	**1.9345922**	18
		Hi[9]h	2.981481	1.9982745	18
		Total	3.449074	**1.9955583**	36
Session 2	Young	Low	4.388889	1.6541194	9
		High	**3.370370**	2.0100058	9
		Total	**3.879630**	1.8610258	18
	Old	Low	2.722222	**2.2653795**	9
		High	2.444444	2.1473498	9
		Total	**2.583333**	2.1460177	18
	Total	Low	**3.555556**	2.1066344	18
		High	2.907407	2.0731888	18
		Total	**3.231481**	**2.0859613**	36
Session 3	Young	Low	**4.333333**	**1.6393596**	9

Descriptive Statistics

Age		Load	Mean	Std. Deviation	N
Session 3	Young	High	3.592593	1.8617329	9
		Total	3.962963	1.7438553	18
	Old	Low	2.611111	2.1618536	9
		High	2.111111	1.8929694	9
		Total	2.361111	1.9879128	18
	Total	Low	3.472222	2.0613547	18
		High	2.851852	1.9744188	18
		Total	3.162037	2.0140326	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Week	Pillai's Trace	.061	.999[a]	2.000	31.000	.380
	Wilks' Lambda	.939	.999[a]	2.000	31.000	.380
	Hotelling's Trace	.064	.999[a]	2.000	31.000	.380
	Roy's Largest Root	.064	.999[a]	2.000	31.000	.380
Week * Age	Pillai's Trace	.091	1.553[a]	2.000	31.000	.228
	Wilks' Lambda	.909	1.553[a]	2.000	31.000	.228
	Hotelling's Trace	.100	1.553[a]	2.000	31.000	.228
	Roy's Largest Root	.100	1.553[a]	2.000	31.000	.228
Week * Load	Pillai's Trace	.018	.285[a]	2.000	31.000	.754
	Wilks' Lambda	.982	.285[a]	2.000	31.000	.754
	Hotelling's Trace	.018	.285[a]	2.000	31.000	.754
	Roy's Largest Root	.018	.285[a]	2.000	31.000	.754
Week * Age * Load	Pillai's Trace	.057	.942[a]	2.000	31.000	.401
	Wilks' Lambda	.943	.942[a]	2.000	31.000	.401
	Hotelling's Trace	.061	.942[a]	2.000	31.000	.401
	Roy's Largest Root	.061	.942[a]	2.000	31.000	.401

a. Exact statistic

b. Design: Intercept + Age + Load + Age * Load Within Subjects Design: Week

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

[redacted]

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of

b. Design: Intercept Age * Load
Within Subjects +esign: Age + Week Load

an identity matrix.

If Within-Subjects Effects table.

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Sphericity Assumed	1.615	2	.807	1.494	.232
	Greenhouse-Geisser	1.615	1.303	1.239	1.494	.234
	Huynh-Feldt	1.615	1.463	1.104	1.494	.235
	Lower-bound	1.615	1.000	1.615	1.494	.231
Week * Age	Sphericity Assumed	.931	2	.465	.861	.428
	Greenhouse-Geisser	.931	1.303	.714	.861	.387
	Huynh-Feldt	.931	1.463	.636	.861	.398
	Lower-bound	.931	1.000	.931	.861	.360
Week * Load	Sphericity Assumed	.547	2	.273	.506	.605
	Greenhouse-Geisser	.547	1.303	.420	.506	.529
	Huynh-Feldt	.547	1.463	.374	.506	.549
	Lower-bound	.547	1.000	.547	.506	.482
Week * Age * Load	Sphericity Assumed	.282	2	.141	.261	.771
	Greenhouse-Geisser	.282	1.303	.217	.261	.674
	Huynh-Feldt	.282	1.463	.193	.261	.701
	Lower-bound	.282	1.000	.282	.261	.613
Error(Week)	Sphericity Assumed	34.588	64	.540		
	Greenhouse-Geisser	34.588	41.699	.829		
	Huynh-Feldt	34.588	46.818	.739		
	Lower-bound	34.588	32.000	1.081		

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	Week	Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Linear	1.483	1	1.483	1.934	.174
	Quadratic	.132	1	.132	.420	.522

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	Week	Type III Sum of Squares	df	Mean Square	F	Sig.
Week * Age	Linear	.889	1	.889	1.159	.290
	Quadratic	.042	1	.042	.133	.718
Week * Load	Linear	.446	1	.446	.581	.451
	Quadratic	.101	1	.101	.321	.575
Week * Age * Load	Linear	.056	1	.056	.072	.790
	Quadratic	.227	1	.227	.723	.402
Error(Week)	Linear	24.543	32	.767		
	Quadratic	10.045	32	.314		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed
Variable:Aver

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1162.520	1	1162.520	112.111	.000
Age	49.343	1	49.343	4.759	.037
Load	14.569	1	14.569	1.405	.245
Age * Load	1.565	1	1.565	.151	.700
Error	331.819	32	10.369		

Estimated Marginal Means

1. Grand Mean

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
3.281	.310	2.650	3.912

2. Age

Estimates

Measure:MEASURE 1

Age	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	3.957	.438	3.064	4.849
Old	2.605	.438	1.712	3.498

Pairwise Comparisons

Measure:MEASURE 1

(II Age	(J) Age	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Young	Old	1.352	.620	.037	.090	2.614
Old	Young	-1.352	.620	.037	-2.614	-.090

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sip.
Contrast	16.448	1	16.448	4.759	.037
Prrnr	110 ROR	32	**3 45R**		

The F tests theeffect of Age. This test is basedon thelinearly independent pairwise comparisons among the estimated marginal means.

3. Load

Estimates

Load	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Low	3.648	.438	2.756	4.541
High	2.914	.438	2.021	3.806

Pairwise Comparisons

(I) Load	(J) Load	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Low	High	.735	.620	.245	-.528	1.997
High	Low	-.735	.620	.245	-1.997	.528

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Siq.
Contrast	4.856	1	4.856	1.405	.245
Prrnr	110 ROR	32	**3 45R**		

The F tests theeffect of Load. This test is basedon thelinearly independent pairwise comparisons among the estimated marginal means.

4. Week

Estimates

Measure:MEASURE 1

Wee k	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	3.449	.321	2.794	4.104
2	3.231	.339	2.542	3.921
3	3.162	.316	2.518	3.806

Pairwise Comparisons

Measure:MEASURE 1

(I) Wee k k	(J) Wee k	Mean Difference (I-J)	Std. Error	Siq.	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	.218	.198	.842	-.283	.719
	3	.287	.206	.522	-.234	.809
2	1	-.218	.198	.842	-.719	.283
	3	.069	.090	1.000	-.159	.298
3	1	-.287	.206	.522	-.809	.234
	2	-.069	.090	1.000	-.298	.159

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Siq.
Pillars trace	.061	.999[a]	2.000	31.000	.380
Wilks' lambda	939	•999[a]	2.000	31.000	.380
Hotelling's trace	.064	.999[a]	2.000	31.000	.380
Roy's largest root	.064	.999[a]	2.000	31.000	.380

Each F tests the multivariate effect of Week. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. Age * Load

Age	Load	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	Low	4.444	.620	3.182	5.707
	High	3.469	.620	2.207	4.731
Old	Low	2.852	.620	1.590	4.114
	High	2.358	.620	1.096	3.620

6. Age *Week

Measure:MEA SURE 1

Age	Wee k	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.028	.454	3.102	4.954
	2	3.880	.479	2.904	4.855
	3	3.963	.447	3.052	4.874
Old	1	2. 870	.454	1.945	3 .796
	2	2.583	.479	1.608	3.559
	3	2.361	.447	1.450	3.272

7. Load * Week Measure

:MEASURE 1

Load	Wee k	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Low	1	3.917	.454	2.991	4.842
	2	3.556	.479	2.580	4.531
	3	3.472	.447	2.561	4.383
High	1	2.981	.454	2.056	3.907
	2	2.907	.479	1.932	3.883
	3	2.852	.447	1.941	3.763

8. Age * Load * Week Measure

:MEASURE 1

Age	Load	Wee k	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	Low	1	4.611	.643	3.302	5.920
		2	4.389	.677	3.009	5.769
		3	4.333	.633	3.045	5.622
High		1	3.444	.643	2.135	4.754
		2	3.370	.677	1.991	4.750
		3	3.593	.633	2.304	4.881

8. Age * Load * Week Measure:MEASURE 1

Age	Load	Week	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Old	Low	1	3.222	.643	1.913	**4.531**
		2	2.722	.677	1.342	4.102
		3	2.611	**.633**	1.322	3.900
High		1	2.519	.643	1.209	3.828
		2	2.444	.677	1.065	3.824
		3	2.111	**.633**	.822	3.400

SET 2

[DataSet1] C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 2\Results\ St2-OVERALL.sav

Within-Subjects Factors Measure:MEASURE 1

Week	Dependent Variable
1	S2W1
2	S2W2

Between-Subjects Factors

		Value Label	N
Age	1	Young	18
	2	Old	18
Load	1	Low	18
	2	High	18

Descriptive Statistics

Age		Load	Mean	Std. Deviation	N
S2W1	Young	Low	3.4444	2.11312	9
		High	4.6296	**.73493**	9
		Total	4.0370	1.65146	18
Old		Low	**3.3889**	2.10324	9
		High	2.4444	1.77169	9
		Total	2.9167	1.94806	18
Total		Low	3.4167	2.04544	18
		High	**3.5370**	1.73069	18
		Total	3.4769	**1.86835**	36

Descriptive Statistics

Age		Load	Mean	Std. Deviation	N
S2W2	Young	Low	3.5000	1.71391	9
		High	3.9259	1.42183	9
		Total	3.7130	1.54328	18
Old		Low	2.7222	2.33333	9
		High	1.4815	1.72491	9
		Total	2.1019	2.09039	18
Total		Low	3.1111	2.02598	18
		High	2.7037	1.98323	18
		Total	2.9074	1.98664	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Siq.
Week	Pillai's Trace	.151	5.681 a	1.000	32.000	.023
	Wilks' Lambda	.849	5.681 a	1.000	32.000	.023
	Hotelling's Trace	.178	5.681 a	1.000	32.000	.023
	Roy's Largest Root	.178	5.681 a	1.000	32.000	.023
Week * Age	Pillai's Trace	.032	1.055[a]	1.000	32.000	.312
	Wilks' Lambda	.968	1.055[a]	1.000	32.000	.312
	Hotelling's Trace	.033	1.055[a]	1.000	32.000	.312
	Roy's Largest Root	.033	1.055[a]	1.000	32.000	.312
Week * Load	Pillai's Trace	.037	1.220[a]	1.000	32.000	.278
	Wilks' Lambda	.963	1.220[a]	1.000	32.000	.278
	Hotelling's Trace	.038	1.220[a]	1.000	32.000	.278
	Roy's Largest Root	.038	1.220[a]	1.000	32.000	.278
Week * Age * Load	Pillai's Trace	.007	•235[a]	1.000	32.000	.631
	Wilks' Lambda	.993	.235[a]	1.000	32.000	.631
	Hotelling's Trace	.007	•235[a]	1.000	32.000	.631
	Roy's Largest Root	.007	•235[a]	1.000	32.000	.631

aEvaluatesbDesignIntercept+Age+Load+Age*LoadWithin Subjects Design: Week

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

[redacted]

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of

b. Design: Intercept Age * Load

Within Subjects +esign: Age + Week Load

an identity matrix.

If Within-Subjects Effects table.

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Sphericity Assumed	5.837	1	5.837	5.681	.023
	Greenhouse-Geisser	5.837	1.000	5.837	5.681	.023
	Huynh-Feldt	5.837	1.000	5.837	5.681	.023
	Lower-bound	5.837	1.000	5.837	5.681	.023
Week * Age	Sphericity Assumed	1.084	1	1.084	1.055	.312
	Greenhouse-Geisser	1.084	1.000	1.084	1.055	.312
	Huynh-Feldt	1.084	1.000	1.084	1.055	.312
	Lower-bound	1.084	1.000	1.084	1.055	.312
Week * Load	Sphericity Assumed	1.253	1	1.253	1.220	.278
	Greenhouse-Geisser	1.253	1.000	1.253	1.220	.278
	Huynh-Feldt	1.253	1.000	1.253	1.220	.278
	Lower-bound	1.253	1.000	1.253	1.220	.278
Week * Age * Load	Sphericity Assumed	.241	1	.241	.235	.631
	Greenhouse-Geisser	.241	1.000	.241	.235	.631
	Huynh-Feldt	.241	1.000	.241	.235	.631
	Lower-bound	.241	1.000	.241	.235	.631
Error(Week)	Sphericity Assumed	32.877	32	1.027		
	Greenhouse-Geisser	32.877	32.000	1.027		
	Huynh-Feldt	32.877	32.000	1.027		
	Lower-bound	32.877	32.000	1.027		

Tests of Within-Subjects Contrasts

Source	Week	Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Linear	5.837	1	5.837	5.681	.023
Week * Age	Linear	1.084	1	1.084	1.055	.312

Tests of Within-Subjects Contrasts

Source	Week	Type III Sum of Squares	df	Mean Square	F	Siq.
Week * Load	Linear	1.253	1	1.253	1.220	.278
Week * Age * Load	Linear	.241	1	.241	.235	.631
Error(Week)	Linear	32.877	32	1.027		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed
Variable:Aver

Source	Type III Sum of Squares	df	Mean Square	F	Siq.
Intercept	733.658	1	733.658	134.387	.000
Age	33.574	1	33.574	6.150	.019
Load	.371	1	.371	.068	.796
Age * Load	16.213	1	16.213	2.970	.094
Error	174.698	32	5.459		

Estimated Marginal Means

1. Grand Mean

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
3.192	.275	2.631	3.753

2. Age

Estimates

Measure:MEASURE 1

Age	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	3.875	.389	3.082	4.668
Old	2.509	.389	1.716	3.302

Pairwise Comparisons

Measure:MEASURE 1

(II Age	(J) Age	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Young	Old	1.366	.551	.019	.244	2.488
Old	Young	-1.366	.551	.019	-2.488	-.244

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sip.
Contrast	16.787	1	16.787	6.150	.019
Prrnr	R7 349	**32**	2 730		

The F tests theeffect of Age. This test is basedon thelinearly independent pairwise comparisons among the estimated marginal means.

3. Load

Estimates

Load	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Low	3.264	.389	2.471	4.057
High	3.120	.389	2.327	3.914

Pairwise Comparisons

(I) Load	(J) Load	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Low	High	.144	.551	.796	-.978	1.265
High	Low	-.144	.551	.796	-1.265	.978

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Siq.
Contrast	.185	1	.185	.068	.796
Prrnr	85 R7 349	32	2 73n		

The F tests the of Load. This test is basedon the linearly independent pairwise comparisons among the estimated marginal means.

4. Week

Estimates

Measure:MEASURE 1

Wee k	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	3.477	.295	2.875	4.079
2	2.907	.305	2.287	3.528

Pairwise Comparisons

Measure:MEASURE 1

(I) Wee k	(J) Wee k	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	.569	.239	.023	.083	1.056
2	1	-.569	.239	.023	-1.056	-.083

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Siq.
Pillars trace	.151	5.681 a	1.000	32.000	.023
Wilks'lambda	.849	5.681 a	1.000	32.000	.023
Hotelling's trace	.178	5.681 a	1.000	32.000	.023
Roy's largest root	.178	5.681 a	1.000	32.000	.023

Each F tests the multivariate effect of Week. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. Age * Load

Age	Load	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	Low	3.472	.551	2.350	4.594
	High	4.278	.551	3.156	5.400
Old	Low	3.056	.551	1.934	4.177
	High	1.963	.551	.841	3.085

6. **Age * Week** Measure:MEASURE 1

Age	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.037	.418	3.186	4.888
	2	3.713	.431	2.835	4.591
Old	1	2.917	.418	2.066	3.768
	2	2.102	.431	1.224	2.980

7. **Load * Week** Measure:MEASURE 1

Load	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Low	1	3.417	.418	2.566	4.268
	2	3.111	.431	2.233	3.989
High	1	3.537	.418	2.686	4.388
	2	2.704	.431	1.826	3.582

8. **Age * Load * Week** Measure:MEASURE 1

Age	Load	Week	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	Low	1	3.444	.591	2.241	4.648
		2	3.500	.610	2.258	4.742
	High	1	4.630	.591	3.426	5.833
		2	3.926	.610	2.684	5.168
Old	Low	1	3.389	.591	2.185	4.592
		2	2.722	.610	1.480	3.964
	High	1	2.444	.591	1.241	3.648
		2	1.481	.610	.240	2.723

Set 1

[DataSet0] C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 2\Results\ St2-OVERALLtime.sav

Within-Subjects Factors

Measure:MEASURE 1

Session	Dependent Variable
1	**SiWi**
2	S1 W2
3	S1 W3

Between-Subjects Factors

		Value Label	N
Age	1	Young	18
	2	Old	18
Load	1	Low	18
	2	High	18

Descriptive Statistics

Age		Load	Mean	Std. Deviation	N
S1 W1	Young	Low	**1.9535**	**1.08365**	9
		High	**2.4255**	1.49277	9
		Total	2.1895	1.28850	18
Old		Low	5.6619	**3.54333**	9
		High	4.9528	1.90098	9
		Total	**5.3074**	2.78245	18
Total		Low	3.8077	3.17826	18
		High	3.6891	2.10712	18
		Total	3.7484	2.65829	36
S1 W2	Young	Low	2.0987	.91004	9
		High	1.7964	.38866	9
		Total	1.9475	.69642	18
Old		Low	4.3269	1.31197	9
		High	4.8238	1.74213	9
		Total	**4.5754**	1.51777	18
Total		Low	3.2128	**1.58556**	18
		High	**3.3101**	1.98127	18
		Total	3.2614	1.76923	36
S1 W3	Young	Low	**1.9337**	.86089	9

Descriptive Statistics

Age		Load	Mean	Std. Deviation	N
51 W3	Young	High	2.3611	1.09995	9
		Total	2.1474	.98309	18
Old		Low	4.5474	.94862	9
		High	4.5764	1.49543	9
		Total	4.5619	1.21494	18
Total		Low	3.2405	1.60638	18
		High	3.4687	1.70904	18
		Total	3.3546	1.63873	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Siq.
Session	Pillai's Trace	.091	1.551 a	2.000	31.000	.228
	Wilks' Lambda	.909	1.551 a	2.000	31.000	.228
	Hotelling's Trace	.100	1.551 a	2.000	31.000	.228
	Roy's Largest Root	.100	1.551 a	2.000	31.000	.228
Session * Age	Pillai's Trace	.033	.531 a	2.000	31.000	.593
	Wilks' Lambda	.967	.531 a	2.000	31.000	.593
	Hotelling's Trace	.034	.531 a	2.000	31.000	.593
	Roy's Largest Root	.034	.531 a	2.000	31.000	.593
Session * Load	Pillai's Trace	.009	.134[a]	2.000	31.000	.875
	Wilks' Lambda	.991	.134[a]	2.000	31.000	.875
	Hotelling's Trace	.009	.134[a]	2.000	31.000	.875
	Roy's Largest Root	.009	.134[a]	2.000	31.000	.875
Session * Age * Load	Pillai's Trace	.152	2.769[a]	2.000	31.000	.078
	Wilks' Lambda	.848	2.769[a]	2.000	31.000	.078
	Hotelling's Trace	.179	2.769[a]	2.000	31.000	.078
	Roy's Largest Root	.179	2.769[a]	2.000	31.000	.078

aExactstatistic.Design:Intercept+Age+Load+Age*Load Within Subjects Design: Session

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

[redacted]

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of
b. Design: Intercept + Age + Load + Age * Load Within Subjects Design: Session

an identity matrix.

If Within-Subjects Effects table.

Tests of Within-Subjects Effects Measure:MEASURE

1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Session	Sphericity Assumed	4.811	2	2.405	1.771	.178
	Greenhouse-Geisser	4.811	1.453	3.310	1.771	.188
	Huynh-Feldt	4.811	1.647	2.921	1.771	.185
	Lower-bound	4.811	1.000	4.811	1.771	.193
Session * Age	Sphericity Assumed	2.341	2	1.171	.862	.427
	Greenhouse-Geisser	2.341	1.453	1.611	.862	.397
	Huynh-Feldt	2.341	1.647	1.421	.862	.409
	Lower-bound	2.341	1.000	2.341	.862	.360
Session * Load	Sphericity Assumed	.552	2	.276	.203	.817
	Greenhouse-Geisser	.552	1.453	.380	.203	.745
	Huynh-Feldt	.552	1.647	.335	.203	.774
	Lower-bound	.552	1.000	.552	.203	.655
Session * Age * Load	Sphericity Assumed	4.477	2	2.238	1.648	.201
	Greenhouse-Geisser	4.477	1.453	3.080	1.648	.207
	Huynh-Feldt	4.477	1.647	2.718	1.648	.205
	Lower-bound	4.477	1.000	4.477	1.648	.208
Error(Session)	Sphericity Assumed	86.926	64	1.358		
	Greenhouse-Geisser	86.926	46.506	1.869		
	Huynh-Feldt	86.926	52.713	1.649		
	Lower-bound	86.926	32.000	2.716		

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	Session	Type III Sum of Squares	df	Mean Square	F	Sig.
Session	Linear	2.791	1	2.791	1.374	.250
	Quadratic	2.020	1	2.020	2.948	.096

Tests of Within-Subjects Contrasts

Source	Session	Type III Sum of Squares	df	Mean Square	F	Sig.
Session * Age	Linear	2.226	1	2.226	1.096	.303
	Quadratic	.115	1	.115	.168	.685
Session * Load	Linear	.541	1	.541	.266	.609
	Quadratic	.011	1	.011	.016	.901
Session * Age * Load	Linear	.689	1	.689	.339	.564
	Quadratic	3.787	1	3.787	5.527	.025
Error(Session)	Linear	65.000	32	2.031		
	Quadratic	21.925	32	.685		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed
Variable:Aver

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1289.070	1	1289.070	264.037	.000
Age	199.766	1	199.766	40.918	.000
Load	.128	1	.128	.026	.872
Age * Load	.457	1	.457	.094	.762
Error	156.229	32	4.882		

Estimated Marginal Means

1. Grand Mean

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
3.455	.213	3.022	3.888

2. Age

Estimates

Measure:MEASURE 1

Age	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	2.095	.301	1.482	2.707
Old	4.815	.301	4.202	5.427

Pairwise Comparisons

Measure:MEASURE 1

(Il Age	(J) Age	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Young	Old	-2.720	.425	.000	-3.586	-1.854
Old	Young	2.720 ~	.425	.000	1.854	3.586

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sip.
Contrast	66.589	1	66.589	40.918	.000
Prrnr	52 07R	**32**	1 R27		

The F tests theeffect of Age. This test is basedon thelinearly independent pairwise comparisons among the estimated marginal means.

3. Load

Estimates

Load	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Low	3.420	.301	2.808	4.033
High	3.489	.301	2.877	4.102

Pairwise Comparisons

(I) Load	(J) Load	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Low	High	-.069	.425	.872	-.935	.797
High	Low	.069	.425	.872	-.797	.935

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Siq.
Contrast	.043	1	.043	.026	.872
Prrnr	52 07R	**32**	1 R27		

The F tests the effect of Load. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

4. Session

Estimates

Measure:MEASURE 1

Session	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	3.748	.369	2.997	4.499
2	3.261	.200	2.855	3.668
3	3.355	.188	2.972	3.737

Pairwise Comparisons

Measure:MEASURE 1

(I) Session	(J) Session	Mean Difference (I-J)	Std. Error	a Siq.	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	.487	.284	.287	-.229	1.203
	3	.394	.336	.749	-.455	1.242
2	1	-.487	.284	.287	-1.203	.229
	3	-.093	.182	1.000	-.553	.367
3	1	-.394	.336	.749	-1.242	.455
	2	.093	.182	1.000	-.367	.553

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Siq.
Pillars trace	.091	1.551 a	2.000	31.000	.228
Wilks' lambda	.909	1.551 a	2.000	31.000	.228
Hotelling's trace	.100	1.551 a	2.000	31.000	.228
Roy's largest root	.100	1.551 a	2.000	31.000	.228

Each F tests the multivariate effect of Session. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. Age * Load

Age	Load	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	Low	1.995	.425	1.129	2.861
	High	2.194	.425	1.328	3.060
Old	Low	4.845	.425	3.979	5.712
	High	4.784	.425	3.918	5.650

6. **Age * Session**

Measure:MEA SURE 1

Age	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	2.189	.521	1.127	3.251
	2	1.948	.282	1.373	2.522
	3	2.147	.266	1.606	2.689
Old	1	5.307	.521	4.245	6.369
	2	4.575	.282	4.000	5.150
	3	4.562	.266	4.020	5.103

7. **Load * Session** Measure

:MEASURE 1

Load	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Low	1	3.808	.521	2.746	4.870
	2	3.213	.282	2.638	3.788
	3	3.241	.266	2.699	3.782
High	1	3.689	.521	2.627	4.751
	2	3.310	.282	2.735	3.885
	3	3.469	.266	2.927	4.010

8. **Age * Load * Session**

Measure :MEASURE 1

Age	Load	Session	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	Low	1	1.953	.737	.452	3.455
		2	2.099	.399	1.286	2.912
		3	1.934	.376	1.168	2.699
High		1	2.426	.737	.924	3.927
		2	1.796	.399	.983	2.609
		3	2.361	.376	1.595	3.127

8. Age * Load * Session

Measure:MEASURE 1

Age	Load	Session	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Old	Low	1	5.662	.737	4.160	7.164
		2	4.327	**.399**	3.514	5.140
		3	4.547	.376	3.782	**5.313**
High		1	4.953	.737	3.451	6.455
		2	4.824	**.399**	4.011	5.637
		3	4.576	.376	3.811	5.342

>

Set 2

[DataSet0] C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 2\Results\ St2-OVERALLtime.sav

Within-Subjects Factors

Measure:MEASURE 1

Session	Dependent Variable
1	S2W1
2	S2W2

Between-Subjects Factors

		Value Label	N
Age	1	Young	18
	2	Old	18
Load	1	Low	18
	2	High	18

Descriptive Statistics

Age		Load	Mean	Std. Deviation	N
S2W1	Young	Low	2.6562	1.70859	9
		High	1.8237	.79984	9
		Total	2.2400	**1.36318**	18
	Old	Low	4.7191	3.01766	9
		High	**5.0685**	.93414	9
		Total	4.8938	2.17446	18
	Total	Low	3.6876	2.60490	18
		High	3.4461	1.87048	18

Descriptive Statistics

Age		Load	Mean	Std. Deviation	N
S2W1	Total	Total	3.5669	2.23834	36
S2W2	Young	Low	2.0920	.51447	9
		High	1.9699	.71320	9
		Total	2.0309	.60652	18
	Old	Low	4.7353	2.58967	9
		High	6.3134	2.65194	9
		Total	5.5244	2.66923	18
	Total	Low	3.4136	2.26495	18
		High	4.1417	2.92285	18
		Total	3.7776	2.60336	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Session	Pillai's Trace	.007	.217[a]	1.000	32.000	.645
	Wilks' Lambda	.993	.217[a]	1.000	32.000	.645
	Hotelling's Trace	.007	.217[a]	1.000	32.000	.645
	Roy's Largest Root	.007	.217[a]	1.000	32.000	.645
Session * Age	Pillai's Trace	.026	.860[a]	1.000	32.000	.361
	Wilks' Lambda	.974	.860[a]	1.000	32.000	.361
	Hotelling's Trace	.027	.860[a]	1.000	32.000	.361
	Roy's Largest Root	.027	.860[a]	1.000	32.000	.361
Session * Load	Pillai's Trace	.035	1.146[a]	1.000	32.000	.292
	Wilks' Lambda	.965	1.146[a]	1.000	32.000	.292
	Hotelling's Trace	.036	1.146[a]	1.000	32.000	.292
	Roy's Largest Root	.036	1.146[a]	1.000	32.000	.292
Session * Age * Load	Pillai's Trace	.003	.082[a]	1.000	32.000	.777
	Wilks' Lambda	.997	.082[a]	1.000	32.000	.777
	Hotelling's Trace	.003	.082[a]	1.000	32.000	.777
	Roy's Largest Root	.003	.082[a]	1.000	32.000	.777

a. Exact statistic b. Design: Intercept + Age +Load + Age * Load Within Subjects Design:Session

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

[redacted]

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of

b. Design: Intercept + Age + Load + Age * Load Within Subjects Design:
Session

an identity matrix.

If Within-Subjects Effects table.

Tests of Within-Subjects Effects Measure:MEASURE

1

Source		Type III Sum of Squares	df	Mean Square	F	Siq.
Session	Sphericity Assumed	.800	1	.800	.217	.645
	Greenhouse-Geisser	.800	1.000	.800	.217	.645
	Huynh-Feldt	.800	1.000	.800	.217	.645
	Lower-bound	.800	1.000	.800	.217	.645
Session * Age	Sphericity Assumed	3.172	1	3.172	.860	.361
	Greenhouse-Geisser	3.172	1.000	3.172	.860	.361
	Huynh-Feldt	3.172	1.000	3.172	.860	.361
	Lower-bound	3.172	1.000	3.172	.860	.361
Session * Load	Sphericity Assumed	4.230	1	4.230	1.146	.292
	Greenhouse-Geisser	4.230	1.000	4.230	1.146	.292
	Huynh-Feldt	4.230	1.000	4.230	1.146	.292
	Lower-bound	4.230	1.000	4.230	1.146	.292
Session * Age * Load	Sphericity Assumed	.302	1	.302	.082	.777
	Greenhouse-Geisser	.302	1.000	.302	.082	.777
	Huynh-Feldt	.302	1.000	.302	.082	.777
	Lower-bound	.302	1.000	.302	.082	.777
Error(Session)	Sphericity Assumed	118.084	32	3.690		
	Greenhouse-Geisser	118.084	32.000	3.690		
	Huynh-Feldt	118.084	32.000	3.690		
	Lower-bound	118.084	32.000	3.690		

Tests of Within-Subjects Contrasts

Source	Sessi on	Type III Sum of Squares	df	Mean Square	F	Siq.
Session	Linear	.800	1	.800	.217	.645
Session *Age	Linear	3.172	1	3.172	.860	.361

Tests of Within-Subjects Contrasts

Source	Session	Type III Sum of Squares	df	Mean Square	F	Siq.
Session * Load	Linear	4.230	1	4.230	1.146	.292
Session * Age * Load	Linear	.302	1	.302	.082	.777
Error(Session)	Linear	118.084	32	3.690		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed
Variable:Aver

Source	Type III Sum of Squares	df	Mean Square	F	Siq.
Intercept	970.961	1	970.961	292.239	.000
Age	170.049	1	170.049	51.181	.000
Load	1.065	1	1.065	.321	.575
Age * Load	9.346	1	9.346	2.813	.103
Error	106.320	32	3.322		

Estimated Marginal Means

1. Grand Mean

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
3.672	.215	3.235	4.110

2. Age

Estimates

Measure:MEASURE 1

Age	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	2.135	.304	1.517	2.754
Old	5.209	.304	4.590	5.828

Pairwise Comparisons

Measure:MEASURE 1

(Il Age	(J) Age	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Young	Old	-3.074	.430	.000	-3.949	-2.198
Old	Young	3.074 ~	.430	.000	2.198	3.949

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sip.
Contrast	85.025	1	85.025	51.181	.000
Prrnr	s 3 1Rn	32	1 R61		

The F tests theeffect of Age. This test is basedon thelinearly independent pairwise comparisons among the estimated marginal means.

3. Load

Estimates

Load	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Low	3.551	.304	2.932	4.169
High	3.794	.304	3.175	4.413

Pairwise Comparisons

(I) Load	(J) Load	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Low	High	-.243	.430	.575	-1.118	.632
High	Low	.243	.430	.575	-.632	1.118

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Siq.
Contrast	.533	1	.533	.321	.575
Prrnr	**s3** 1Rn	**32**	1 R61		

The F tests the effect of Load. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

4. Session

Estimates

Measure:MEASURE 1

Session	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	3.567	.307	2.942	4.191
2	3.778	.317	3.131	4.424

Pairwise Comparisons

Measure:MEASURE 1

(I) Session	(J) Session	Mean Difference (I-J)	Std. Error	**Siq.**[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	-.211	.453	.645	-1.133	.712
2	1	.211	.453	.645	-.712	1.133

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	**Siq.**
Pillars trace	.007	.217[a]	1.000	32.000	.645
Wilks'lambda	993	•217[a]	1.000	32.000	.645
Hotelling's trace	.007	.217[a]	1.000	32.000	.645
Roy's largest root	.007	.217[a]	1.000	32.000	.645

Each F tests the multivariate effect of Session. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. Age * Load

Measure:MEASURE 1

Age	Load	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	Low	2.374	.430	1.499	3.249
	High	1.897	.430	1.022	2.772
Old	Low	4.727	.430	3.852	5.602
	High	5.691	.430	4.816	6.566

6. Age * Session

Age	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	2.240	.434	1.357	3.123
	2	2.031	.449	1.116	2.945
Old	1	4.894	.434	4.011	5.777
	2	5.524	.449	4.610	6.439

7. **Load * Session** Measure:MEASURE 1

Load	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Low	1	3.688	.434	2.804	4.571
	2	3.414	.449	2.499	4.328
High	1	3.446	.434	2.563	4.329
	2	4.142	.449	3.227	5.056

8. **Age * Load * Session** Measure:MEASURE 1

Age	Load	Session	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	Low	1	2.656	.613	1.407	3.905
		2	2.092	.635	.799	3.385
	High	1	1.824	.613	.575	3.073
		2	1.970	.635	.677	3.263
Old	Low	1	4.719	.613	3.470	5.968
		2	4.735	.635	3.442	6.029
	High	1	5.069	.613	3.819	6.318
		2	6.313	.635	5.020	7.607

SET1

Notes

Output Created		18-Jan-2012 08:25:27
Comments		/Users/jjnicholson/Dropbox/Ph D/Study 4/Results/St4-Datasheet-Attem pts.sav
Input	**Data**	DataSetI <none>
	Active Dataset	<none> <none>
	Filter	
	Weight	
	Split File	
	N of Rows in Working Data File	36
Missing Value Handling	**Definition of Missing**	User-defined missing values are treated as missing. Statistics are based on all cases with valid data for all variables in the model.
	Cases Used	GLM S1 S1 S2S1 S3S1 S1 D1 S2D1 S3D1 BY Age /WSFACTOR=Grid 2 Polynomial Week 3 Polynomial IMETHOD=SSTYPE(3) /EMMEANS=TAB LES(OVERALL) /EMMEANS=TAB LES(Age) COMPARE ADJ(BONFERRONI) /EMMEANS=TAB LES(G rid) COMPARE ADJ(BONFERRONI) /EMMEANS=TAB LES(Week) COMPARE ADJ(BONFERRONI) /EMMEANS=TAB LES(Age*Grid) /EMMEANS=TAB LES(Age*Week) /EMMEANS=TAB LES(G rid*Week) IEMMEANS=TABLES (A a*Grid*Week) /IRINT=DESCRIPTIVE /CRITERIA=A L P H A(.05) /WSDESIGN=Grid Week Grid*Week /DESIGN=Age.
Syntax		
Resources	**Processor Time**	00:00:00.122
	Elapsed Time	00:00:01.000

Within-Subjects Factors

Measure:MEASURE 1

Grid	Week	Dependent Variable
1	1	S1S1
	2	S251
	3	S351
2		
	1 2 3	S2D1 S3D1

Between-Subjects Factors Factors

		Value Label	N
Age	1	Young	18
	2	Old	18

Descriptive ptive Statistics

	Age	Mean	Std. Deviation	N
S1-S1	Young	4.89	.471	18
	Old	4.39	1.195	18
	Total	4.64	.931	36
S2-S1	Young	4.78	.428	18
	Old	3.11	2.374	18
	Total	3.94	1.881	36
S3-S1	Young	3.78	1.629	18
	Old	2.44	2.357	18
	Total	3.11	2.108	36
S1-D1	Young	5.00	.000	18
	Old	4.67	.970	18
	Total	4.83	.697	36
S2-D1	Young	4.56	1.149	18
	Old	3.17	2.176	18
	Total	3.86	1.854	36
S3-D1	Young	4.28	1.487	18
	Old	2.72	2.296	18
	Total	3.50	2.063	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Grid	Pillars Trace	.023	.800[d]	1.000	34.000	.377
	Wilks' Lambda	.977	.800[a]	1.000	34.000	.377
	Hotelling's Trace	.024	.800[a]	1.000	34.000	.377
	Roy's Largest Root	.024	.800[a]	1.000	34.000	.377
Grid * Age	Pillars Trace	.001	.040[a]	1.000	34.000	.844
	Wilks' Lambda	.999	.040[a]	1.000	34.000	.844
	Hotelling's Trace	.001	.040[a]	1.000	34.000	.844
	Roy's Largest Root	.001	.040[a]	1.000	34.000	.844
Week	Pillars Trace	.422	12.042[d]	2.000	33.000	.000
	Wilks' Lambda	.578	12.042[a]	2.000	33.000	.000
	Hotelling's Trace	.730	12.042[a]	2.000	33.000	.000
	Roy's Largest Root	.730	12.042[a]	2.000	33.000	.000
Week *Age	Pillars Trace	.161	3.168[d]	2.000	33.000	.055
	Wilks' Lambda	.839	3.168[a]	2.000	33.000	.055
	Hotelling's Trace	.192	3.168[a]	2.000	33.000	.055
	Roy's Largest Root	.192	3.168[a]	2.000	33.000	.055
Grid * Week	Pillars Trace	.046	.787[d]	2.000	33.000	.464
	Wilks' Lambda	.954	.787[a]	2.000	33.000	.464
	Hotelling's Trace	.048	.787[a]	2.000	33.000	.464
	Roy's Largest Root	.048	.787[a]	2.000	33.000	.464
Grid * Week * Age	Pillars Trace	.015	.256[d]	2.000	33.000	.775
	Wilks' Lambda	.985	.256[a]	2.000	33.000	.775
	Hotelling's Trace	.016	.256[a]	2.000	33.000	.775
	Roy's Largest Root	.016	.256[a]	2.000	33.000	.775

a. Exact statistic

b. Design: Intercept + Age Within Subjects Design: Grid + Week + Grid * Week

Mauchly's Test of Sphericity[b]

Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	EP silon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Grid	1.000	.000	0		1.000	1.000	1.000
Week	.880	4.225	2	.121	.893	.967	.500
Grid * Week	.992	.277	2	.871	.992	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + Age Within Subjects Design: Grid + Week + Grid * Week

Tests of Within-Subjects Effects

Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Grid	Sphericity Assumed	1.500	1	1.500	.800	.377
	Greenhouse-Geisser	1.500	1.000	1.500	.800	.377
	Huynh-Feldt	1.500	1.000	1.500	.800	.377
	Lower-bound	1.500	1.000	1.500	.800	.377
Grid * Age	Sphericity Assumed	.074	1	.074	.040	.844
	Greenhouse-Geisser	.074	1.000	.074	.040	.844
	Huynh-Feldt	.074	1.000	.074	.040	.844
	Lower-bound	.074	1.000	.074	.040	.844
Error(Grid)	Sphericity Assumed	63.759	34	1.875		
	Greenhouse-Geisser	63.759	34.000	1.875		
	Huynh-Feldt	63.759	34.000	1.875		
	Lower-bound	63.759	34.000	1.875		
Week	Sphericity Assumed	74.343	2	37.171	14.678	.000
	Greenhouse-Geisser	74.343	1.785	41.638	14.678	.000
	Huynh-Feldt	74.343	1.933	38.456	14.678	.000
	Lower-bound	74.343	1.000	74.343	14.678	.001
Week * Age	Sphericity Assumed	13.787	2	6.894	2.722	.073
	Greenhouse-Geisser	13.787	1.785	7.722	2.722	.080
	Huynh-Feldt	13.787	1.933	7.132	2.722	.075
	Lower-bound	13.787	1.000	13.787	2.722	.108
Error(Week)	Sphericity Assumed	172.204	68	2.532		
	Greenhouse-Geisser	172.204	60.705	2.837		
	Huynh-Feldt	172.204	65.728	2.620		
	Lower-bound	172.204	34.000	**5.065**		
Grid * Week	Sphericity Assumed	2.028	2	1.014	.873	.423
	Greenhouse-Geisser	2.028	1.983	1.022	.873	.422
	Huynh-Feldt	2.028	2.000	1.014	.873	.423
	Lower-bound	2.028	1.000	2.028	.873	.357
Grid * Week * Age	Sphericity Assumed	.620	2	.310	.267	.767
	Greenhouse-Geisser	.620	1.983	.313	.267	.765
	Huynh-Feldt	.620	2.000	.310	.267	.767
	Lower-bound	.620	1.000	.620	.267	.609
Error(Grid*Week)	Sphericity Assumed	79.019	68	1.162		
	Greenhouse-Geisser	79.019	67.437	1.172		
	Huynh-Feldt	79.019	68.000	1.162		
	Lower-bound	79.019	34.000	2.324		

Tests of Within-Subjects Contrasts Measure:MEASURE

Source	G	Week	Type III Sum of Squares	df	Mean Square	F	Sig.
Grid	Linear		1.500	1	1.500	.800	.377
Grid * Age	Linear		.074	1	.074	.040	.844
Error(Grid)	Linear		63.759	34	1.875		
Week		Linear	73.674	1	73.674	22.452	.000
		Quadratic	.669	1	.669	.375	.544
Week * Age		Linear	9.507	1	9.507	2.897	.098
		Quadratic	4.280	1	4.280	2.400	.131
Error(Week)		Linear	111.569	34	3.281		
		Quadratic	60.634	34	1.783		
Grid * Week	Linear	Linear	.340	1	.340	.321	.575
		Quadratic	1.688	1	1.688	1.336	.256
Grid * Week * Age	Linear	Linear	.340	1	.340	.321	.575
		Quadratic	.280	1	.280	.222	.641
Error(Grid*Week)	Linear	Linear	36.069	34	1.061		
		Quadratic	42.949	34	1.263		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	3424.074	1	3424.074	588.909	.000
Age	68.907	1	68.907	11.851	.002
Error	197.685	34	5.814		

Estimated Marginal Means

1. Grand Mean

Measure:MEASURE 1

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
3.981	.164	3.648	4.315

2. Age

Estimates

Measure:MEASURE 1

Age	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	4.546	.232	4.075	5.018
Old	3.417	.232	2.945	3.888

Pairwise Comparisons

Measure:MEASURE 1

(I) Age	(J) Age	Mean Difference (I-	Std. Error	Sig.[a]	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Young	Old	1.130	.328	.002	.463	1.796
Old	Young	-1.130	.328	.002	-1.796	-.463

Based on estimated mated marginal means *. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	11.485	1	11.485	11.851	.002
Error	32.948	34	.969		

The F tests the of Age. Age. This test isbasedon the linearlyi independent pairwise comparisons among the estimated marginal means.

. Grid

Estimates

Measure:MEASURE 1

Grid	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	3.898	.188	3.516	4.280
2	4	.189	3.680	4.450

Pairwise Comparisons

asure:MEASURE 1

(I) G r i d	(J) Grid	Mean Difference (I- .It	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	-.167	.186	.377	-.545	.212
2	1	.167	.186	.377	- .212	.545

Based on estimated matedmarginal means a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.023	.800[d]	1.000	34.000	.377
Wilks' lambda	.977	.800[a]	1.000	34.000	.377
Hotelling's trace	.024	•800[a]	1.000	34.000	.377
Roy's largest root	.024	•800[a]	1.000	34.000	.377

Each F tests the multivariate effect of Grid. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

Estimates

Measure:MEASURE 1

Week	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.	.127	4.478	4.994
2	3.903	.227	3.442	4.363
3	3.306	.289	2.718	3.893

Pairwise Comparisons

Measure:MEASURE 1

(I) W e e k (J) Week		Mean Difference (I-.!t	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	.833	.221	.002	.276	1.390
	3	1.431 *	.302	.000	.670	2.191
2	1	-.833	.221	.002	-1.390	-.276
	3	.597	.266	.095	-.074	1.268
3	1	-1.431	.302	.000	-2.191	-.670
	2	-.597	.266	.095	-1.268	.074

Based on estimated matedmarginal means *. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.422	12.042[d]	2.000	33.000	.000
Wilks' lambda	.578	12.04e	2.000	33.000	.000
Hotelling's trace	.730	12.04e	2.000	33.000	.000
Roy's largest root	.730	12.04e	2.000	33.000	.000

Each F tests the multivariate effect of Week. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. Age * Grid

Measure:MEASURE 1

Age	Grid	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.481	.266	3.941	5.022
	2	4.611	.268	4.067	5.155
Old	1	3.315	.266	2.774	3.855
	2	3.519	.268	2.974	4.063

6.Age* Week

Measure:MEASURE 1

Age	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.944	.179	4.580	**5.309**
	2	4.667	.321	4.015	**5.318**
	3	4.028	.409409	3.197	4.859
Old	1	.528 4.528	.179	4.163	4.892
	2	**3.139**	.321	2.487	3.790
	3	**2.583**	.409	1.752	3.414

7. Grid * Week

Meure:MEASURE_1 ure:MEASURE 1

Grid	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	4.639	.151	4.331	4.947
	2	3.944	.284	**3.367**	4.522
	3	3.111	**.338**	2.425	3.797
2	1	**4.833**	.114	4.601	5.066
	2	3.861	.290	3.272	4.450
	3	**3.500**	.322	2.845	4.155

8. * Grid * Week

ure:MEASURE 1

Age	Grid		Mean	Std. Error	95%Confidence Interval	
					Lower **Bound**	Upper Bound
Young	1	1	4.889	.214	4.454	5.324
		2	4.778	.402	3.961	**5.595**
		3	3.778	.478	2.807	4.748
2		1	**5.000**	.162	4.671	**5.329**
		2	4.556	.410	3.722	**5.389**
		3	4.278	.456	**3.351**	5.204
Old	1	1	4.389	.214	3.954	4.824
		2	3.111	.402	2.294	**3.928**
		3	2.444	.478	1.474	3.415
2		1	4.667	.162	4.338	4.995
		2	3.167	.410	**2.333**	4.000
		3	2.722	.456	1.796	3.649

IM S2S2 S3S2 S2I2 S3D2Y Age /WSFACTOR=Grid Grid 2 PolynomialWeek 2 Polynomial.

/EMMEANS=TABLES(Grid) COMPARE ADJ(BONFERRONI) /EMMEANS=TABLES(Week) COMPARE ADJ(BONFERRONI) /EMMEANS=TABLES(Age*Grid)
/EMMEANS=TABLES(Age*Week)
/EMMEANS=TABLES(Grid*Week)
/EMMEANS=TABLES(Age*Grid*Week)
/PRINT=DESCRIPTIVE
/CRITERIA=ALPHA(.05)
/WSDESIGN=Grid Week Grid*Week
/DESIGN=Age.

Notes

[DataSetl] /Users/jjnicholson/Dropbox/PhD/Study 4/Results/St4-Datasheet-Attempts.sav

Output Created		18-Jan-2012 08:28:12
Comments		
Input	**Data**	/Users/jjnicholson/Dropbox/PhD/Study 4/Results/St4-Datasheet-Attem pts.sav DataSetl
	Active Dataset	<none>
	Filter	<none>
	Weight	<none>
	Split File	36
	N of Rows in Working Data File	User-defined missing values are treated as missing.
Missing Value Handling	**Definition of Missing**	Statistics are based on all cases with valid data for all variables in the model.
	Cases Used	GLM S2S2 S3S2 S2D2 S3D2 BY Age /WSFACTOR=Grid 2 Polynomial Week 2 Polynomial IMETHOD=SSTYPE(3) /EMMEANS=TAB LES(OVERALL) IEMMEANS=TABLES(Age) COMPARE ADJ(BONFERRONI) /EMMEANS=TAB LES(G rid) COMPARE ADJ(BONFERRONI) /EMMEANS=TAB LES(Week) COMPARE ADJ(BONFERRONI) /EMMEANS=TAB LES(Age*Grid) /EMMEANS=TAB LES(Age*Week) /EMMEANS=TAB LES(G rid*Week) IEMMEANS=TABLES (Age*Grid*Week) /PRINT=DESCRIPTIVE /CRITERIA=A L P H A(.05) /WSDESIGN=Grid Week Grid*Week /DESIGN=Age.
Syntax		
Resources	**Processor Time**	00:00:00.028
	Elapsed Time	00:00:00.000

Within-Subjects Factors

Measure:MEASURE 1

Grid	Week	Dependent Variable
1	1	S2S2
	2	S3S2
2	1	S2D2
	2	53D2

Between-Subjects Factors

		Label	N
Age	1	Young	18
	2	Old	18

Descriptive Statistics

	Age	Mean	Std. Deviation	N
S2-S2	Young	4.83	.514	18
	Old	3.78	1.478	18
	Total	4.31	1.215	36
S3-S2	Young	2.39	2.330	18
	Old	2.28	2.396	18
	Total	2.33	2.330	36
S2-D2	Young	4.56	1.042	18
	Old	4.33	1.237	18
	Total	4.44	1.132	36
S3-D2	Young	3.89	1.967	18
	Old	1.39	2.033	18
	Total	2.64	2.344	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Grid	Pillars Trace	.033	1.156[d]	1.000	34.000	.290
	Wilks' Lambda	.967	1.156[a]	1.000	34.000	.290
	Hotelling's Trace	034	1.156[a]	1.000	34.000	.290
	Roy's Largest Root	034	1.156[a]	1.000	34.000	.290
Grid * Age	Pillars Trace	.094	3.541[a]	1.000	34.000	.068
	Wilks' Lambda	.906	3.541[a]	1.000	34.000	.068
	Hotelling's Trace	104	3.541[a]	1.000	34.000	.068
	Roy's Largest Root	104	3.541[a]	1.000	34.000	.068
Week	Pillars Trace	.550	41.570[a]	1.000	34.000	.000
	Wilks' Lambda	.450	41.570[a]	1.000	34.000	.000
	Hotelling's Trace	1.223	41.570[a]	1.000	34.000	.000
	Roy's Largest Root	1.223	41.570[a]	1.000	34.000	.000
Week * Age	Pillars Trace	.037	1.295[d]	1.000	34.000	.263
	Wilks' Lambda	.963	1.295[a]	1.000	34.000	.263
	Hotelling's Trace	038	1.295[a]	1.000	34.000	.263
	Roy's Largest Root	038	1.295[a]	1.000	34.000	.263
Grid * Week	Pillars Trace	.005	.158[d]	1.000	34.000	.694
	Wilks' Lambda	.995	.158[a]	1.000	34.000	.694
	Hotelling's Trace	.005	.158[a]	1.000	34.000	.694
	Roy's Largest Root	.005	.158[a]	1.000	34.000	.694
Grid * Week * Age	Pillars Trace	.302	14.739[d]	1.000	34.000	.001
	Wilks' Lambda	.698	14.739[a]	1.000	34.000	.001
	Hotelling's Trace	434	14.739[a]	1.000	34.000	.001
	Roy's Largest Root	.434	14.739[a]	1.000	34.000	.001

a. Exact statistic

b. Design: Intercept + Age Within Subjects Design: Grid + Week + Grid * Week

Mauchly's Test of Sphericity[b]

Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Grid	1.000	.000	0		1.000	1.000	1.000
Week	1.000	.000	0		1.000	1.000	1.000
Grid * Week	1.000	.000	0		1.000	1.000	1.000

Tests thenull thatthe error covariance matrix of the orthonormalized transformeddependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + Age Within Subjects Design: Grid + Week + Grid * Week

Tests of Within-Subjects Effects

Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Grid	Sphericity Assumed	1.778	1	1.778	1.156	.290
	Greenhouse-Geisser	1.778	1.000	1.778	1.156	.290
	Huynh-Feldt	1.778	1.000	1.778	1.156	.290
	Lower-bound	1.778	1.000	1.778	1.156	.290
Grid * Age	Sphericity Assumed	5.444	1	5.444	3.541	.068
	Greenhouse-Geisser	5.444	1.000	5.444	3.541	.068
	Huynh-Feldt	5.444	1.000	5.444	3.541	.068
	Lower-bound	5.444	1.000	5.444	3.541	.068
Error(Grid)	Sphericity Assumed	52.278	34	1.538		
	Greenhouse-Geisser	52.278	34.000	1.538		
	Huynh-Feldt	52.278	34.000	1.538		
	Lower-bound	52.278	34.000	1.538		
Week	Sphericity Assumed	128.444	1	128.444	41.570	.000
	Greenhouse-Geisser	128.444	1.000	128.444	41.570	.000
	Huynh-Feldt	128.444	1.000	128.444	41.570	.000
	Lower-bound	128.444	1.000	128.444	41.570	.000
Week * Age	Sphericity Assumed	4.000	1	4.000	1.295	.263
	Greenhouse-Geisser	4.000	1.000	4.000	1.295	.263
	Huynh-Feldt	4.000	1.000	4.000	1.295	.263
	Lower-bound	4.000	1.000	4.000	1.295	.263
Error(Week)	Sphericity Assumed	105.056	34	3.090		
	Greenhouse-Geisser	105.056	34.000	3.090		
	Huynh-Feldt	105.056	34.000	3.090		
	Lower-bound	105.056	34.000	3.090		
Grid * Week	Sphericity Assumed	.250	1	.250	.158	.694
	Greenhouse-Geisser	.250	1.000	.250	.158	.694
	Huynh-Feldt	.250	1.000	.250	.158	.694
	Lower-bound	.250	1.000	.250	.158	.694
Grid * Week * Age	Sphericity Assumed	23.361	1	23.361	14.739	.001
	Greenhouse-Geisser	23.361	1.000	23.361	14.739	.001
	Huynh-Feldt	23.361	1.000	23.361	14.739	.001
	Lower-bound	23.361	1.000	23.361	14.739	.001

Tests of Within-Subjects Effects Measure:MEASURE

Source		Type III Sum of Squares	df	Mean Square
Error(Grid*Week)	Sphericity Assumed	53.889	34	1.585
	Greenhouse-Geisser	53.889	34.000	1.585
	Huynh-Feldt	53.889	34.000	1.585
	Lower-bound	53.889	34.000	1.585

Tests of Within-Subjects Contrasts Measure:MEASURE

Source	G	Week	Type III Sum of Squares	df	Mean Square	F	Sig.
Grid	Linear		1.778	1	1.778	1.156	.290
Grid * Age	Linear		5.444	1	5.444	3.541	.068
Error(Grid)	Linear		52.278	34	1.538		
Week		Linear	128.444	1	128.444	41.570	.000
Week *Age		Linear	4.000	1	4.000	1.295	.263
Error(Week)		Linear	105.056	34	3.090		
Grid * Week	Linear	Linear	.250	1	.250	.158	.694
Grid * Week * Age	Linear	Linear	23.361	1	23.361	14.739	.001
Error(Grid*Week)	Linear	Linear	53.889	34	1.585		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1694.694	1	1694.694	286.982	.000
Age	34.028	1	34.028	5.762	.022
Error	200.778	34	5.905		

Estimated Marginal Means ns

1. Grand Mean

Measure:MEASURE 1

		95% Confidence Interval	
Mean	Std. Error	Lower Bound	Upper Bound
3.431	.203	3.019	3.842

2. Age

Estimates

Measure:MEASURE 1

Age			95% Confidence Interval	
	Mean	Std. Error	Lower Bound	Upper Bound
Young	3.917	.286	3.335	4.499
Old	2.944	.286	2.362	3.526

Pairwise Comparisons

Measure:MEASURE 1

(I) Age	(J) Age	Mean Difference (I-	Std. Error	Sig.[a]	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Young	Old	.972	.405	.022	.149	1.795
Old	Young	-.972	.405	.022	-1.795	-.149

Based on estimated mated marginal means *. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	8.507	1	8.507	5.762	.022
Error	50.194	34	1.476		

The F tests theeffectof Age. This test is basedon the linearly independent ndependent pairwise comparisons among the estimatedmarginalmeans.

. Grid

Estimates

Measure:MEASURE 1

Grid	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	3.319	.245	2.821	3.817
2	3	.208	3.119	3.965

Pairwise Comparisons

asure:MEASURE 1

(I) G r i d	(J) Grid	Mean Difference (I- .It	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	-.222	.207	.290	-.642	.198
2	1	.222	.207	.290	- .198	.642

Based on estimated matedmarginal means a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.033	1.156[d]	1.000	34.000	.290
Wilks' lambda	.967	1.156[a]	1.000	34.000	.290
Hotelling's trace	.034	1.156[a]	1.000	34.000	.290
Roy's largest root	.034	1.156[a]	1.000	34.000	.290

Each F tests the multivariate effect of Grid. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

Estimates

Measure:MEASURE 1

Week	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.375	.139	4.093	4.657
2	2.486	.325	1.825	3.147

Pairwise Comparisons

Measure:MEASURE 1

(I) Week	(J) Week	Mean Difference (I-.!t	Std. Error	Sig.[a]	95% Confidence Interval for s Difference	
					Lower Bound	Upper Bound
1	2	1.889	.293	.000	1.294	2.484
2	1	-1.889	.293	.000	-2.484	-1.294

Based on estimated mated marginal means *. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.550	41.570[d]	1.000	34.000	.000
Wilks' lambda	.450	41.570[a]	1.000	34.000	.000
Hotelling's trace	1.223	41.570[a]	1.000	34.000	.000
Roy's largest root	1.223	41.570[a]	1.000	34.000	.000

Each F tests the multivariate effect of Week. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. Age * Grid

Measure:MEASURE 1

Age	Grid	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	3.611	.347	2.907	4.315
	2	4.222	. 294	3.624	4.820
Old	1	3.028	.347	2.323	3.732
	2	2.861	.294	2.263	3.459

6. Age * Week

Meure:MEASURE_1 ure:MEASURE 1

Age	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.694	.196	4.296	5.093
	2	3.139	.460	2.204	4.073
Old	1	4.056	.196	3.657	4.454
	2	1.833	.460	.899	2.768

7. Grid * Week

Measure:MEASURE 1

Grid	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	4.306	.18484	**3.931**	4.680
	2	**2.333**	.394	**1.533**	3.134
2	1	4.444	.191	4.057	4.832
	2	2.639	**.333**	1.961	**3.316**

8. * Grid *Week as

ure:MEASURE 1

Age	GridWeek	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1 1	4.833	.261	4.303	**5.363**
	2	2.389	.557	1.257	3.521
2	1	4.556	.269	4.008	5.103
	2	3.889	.472	2.931	4.847
Old	1 1	3.778	.261	3.248	4.308
	2	2.278	.557	1.146	3.410
2	1	**4.333**	.269	3.786	4.881
	2	1.389	.472	.431	2.347

C3: Study 3 (Chapter 6) – Traditional GAS

SET 1

General Linear Model

Notes

Output Created		**2011-02-10T08:46:53.704**
Comments		
Input	Data	**C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 1 \Results\Spreadsheet-Attempts. sav**
	Active Dataset	**DataSetI**
	Filter	**<none>** **<none>**
	Weight	**<none>** **36**
	Split File	**User-defined missing values are treated as missing.**
	N of Rows in Working Data File	**Statistics are based on all cases** with **valid** data for all **variables** in the **model.**
Missing Value Handling	**Definition of Missing**	
	Cases Used	

Notes

Syntax		GLM S1 F1 S2F1 S3F1 S1 P1 S2P1 S3P1 BY Age /WSFACTOR=System 2 Polynomial Session 3 Polynomial /METHOD=SSTYPE(3) /POSTHOC=Age(TUKEY) /EMM EANS=TABLES(OVERALL) /EMM EANS=TABLES(Agge) COMPARE ADJ(BONFERRONI) /EMM EANS=TABLES(System) COMPARE ADJ(BONFERRONI) /EMM EANS=TABLES(Session) COMPARE ADJ(BONFERRONI) /EMMEANS=TABLES(Age" System) /EMM EANS=TABLES (A e" Session) /EMM EANS=TABLES (System*Session) /EMM EANS=TABLES Age*System" Session /PRINT=DESCRIPTIVE /CRITERIA=A L P HA(.05) /WSDESIGN=System Session System*Session /DESIGN=Age.
Resources	Processor Time	0:00:00.140 0:00:00.169
	Elapsed Time	

[DataSetl] C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 1\Results\ Spreadsheet-Attempts.sav

Warnings

Post hoc tests are not performed for Age because there are fewer than three groups.

Within-Subjects Factors

Measure:MEASURE 1

System	Session	Dependent Variable
1	1	S1 F1
	2	52F1
	3	53F1
2	1	S1 P1
	2	52P1
	3	53P1

Between-Subjects Factors

		Value Label	N
Age	1	Young	18
	2	Old	18

Descriptive Statistics

	Age	Mean	Std. Deviation	N
S1 F1	Young	5.0000	.00000	18
	Old	5.0000	.00000	18
	Total	5.0000	.00000	36
S2F1	Young	5.0000	.00000	18
	Old	4.9444	.23570	18
	Total	4.9722	.16667	36
S3F1	Young	4.6111	1.19503	18
	Old	4.1667	1.58114	18
	Total	4.3889	1.39955	36
S1 P1	Young	5.0000	.00000	18
	Old	4.9444	.23570	18
	Total	4.9722	.16667	36
S2P1	Young	4.8889	.47140	18
	Old	4.6667	.59409	18
	Total	4.7778	.54043	36
S3P1	Young	4.9444	.23570	18
	Old	3.6111	1.71974	18
	Total	4.2778	1.38587	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
System	Pillars Trace	.045	1.602[a]	1.000	34.000	.214
	Wilks' Lambda	.955	1.602[a]	1.000	34.000	.214
	Hotelling's Trace	.047	1.602[a]	1.000	34.000	.214
	Roy's Largest Root	.047	1.602[a]	1.000	34.000	.214
System * Age	Pillars Trace	.116	4.450[a]	1.000	34.000	.042
	Wilks' Lambda	.884	4.450[a]	1.000	34.000	.042
	Hotelling's Trace	.131	4.450[a]	1.000	34.000	.042
	Roy's Largest Root	.131	4.450[a]	1.000	34.000	.042
Session	Pillars Trace	.297	6.974[a]	2.000	33.000	.003
	Wilks' Lambda	.703	6.974[a]	2.000	33.000	.003
	Hotelling's Trace	.423	6.974[a]	2.000	33.000	.003
	Roy's Largest Root	.423	6.974[a]	2.000	33.000	.003
Session * Age	Pillars Trace	.145	2.800[a]	2.000	33.000	.075
	Wilks' Lambda	.855	2.800[a]	2.000	33.000	.075
	Hotelling's Trace	.170	2.800[a]	2.000	33.000	.075

a. Exact statistic

b. Design: Intercept + Age Within Subjects Design: System + Session + System * Session

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Siq.
Session * Age	Roy's Largest Root	.170	2.800[a]	2.000	33.000	.075
System * Session	Pillai's Trace	.075	1.337[a]	2.000	33.000	.277
	Wilks' Lambda	.925	1.337[a]	2.000	33.000	.277
	Hotelling's Trace	.081	1.337[a]	2.000	33.000	.277
	Roy's Largest Root	.081	1.337[a]	2.000	33.000	.277
System * Session * Age	Pillai's Trace	.080	1.432[a]	2.000	33.000	.253
	Wilks' Lambda	.920	1.432[a]	2.000	33.000	.253
	Hotelling's Trace	.087	1.432[a]	2.000	33.000	.253
	Roy's Largest Root	.087	1.432[a]	2.000	33.000	.253

a. Exact statistic b. Design: Intercept + Age

Within Subjects Design: System + Session + System * Session

Mauchly's Test of Sphericity[b] Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Siq.	Epsilon *a* Greenhouse-Geisser	Huynh-Feldt	Lower-bound
System	1.000	.000	0		1.000	1.000	1.000
Session	.201	52.998	2	.000	.556	.578	.500
c.-+- * *c-*;-	*RCA*	***Ra 9RQ***	*9*	nnn	Ann	*RR7*	cnn

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + Age Within Subjects Design: System + Session + System * Session

Tests of Within-Subjects Effects Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Siq.
System	Sphericity Assumed	.667	1	.667	1.602	.214
	Greenhouse-Geisser	.667	1.000	.667	1.602	.214
	Huynh-Feldt	.667	1.000	.667	1.602	.214
	Lower-bound	.667	1.000	.667	1.602	.214
System * Age	Sphericity Assumed	1.852	1	1.852	4.450	.042
	Greenhouse-Geisser	1.852	1.000	1.852	4.450	.042
	Huynh-Feldt	1.852	1.000	1.852	4.450	.042
	Lower-bound	1.852	1.000	1.852	4.450	.042
Error(System)	Sphericity Assumed	14.148	34	.416		
	Greenhouse-Geisser	14.148	34.000	.416		

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Error(System)	Huynh-Feldt	14.148	34.000	.416		
	Lower-bound	14.148	34.000	.416		
Session	Sphericity Assumed	17.565	2	8.782	11.086	.000
	Greenhouse-Geisser	17.565	1.112	15.802	11.086	.001
	Huynh-Feldt	17.565	1.156	15.196	11.086	.001
	Lower-bound	17.565	1.000	17.565	11.086	.002
Session * Age	Sphericity Assumed	7.898	2	3.949	4.985	.010
	Greenhouse-Geisser	7.898	1.112	7.106	4.985	.028
	Huynh-Feldt	7.898	1.156	6.833	4.985	.027
	Lower-bound	7.898	1.000	7.898	4.985	.032
Error(Session)	Sphericity Assumed	53.870	68	.792		
	Greenhouse-Geisser	53.870	37.792	1.425		
	Huynh-Feldt	53.870	39.300	1.371		
	Lower-bound	53.870	34.000	1.584		
System * Session	Sphericity Assumed	.250	2	.125	.272	.763
	Greenhouse-Geisser	.250	1.215	.206	.272	.650
	Huynh-Feldt	.250	1.273	.196	.272	.661
	Lower-bound	.250	1.000	.250	.272	.605
System * Session * Age	Sphericity Assumed	1.843	2	.921	2.005	.142
	Greenhouse-Geisser	1.843	1.215	1.516	2.005	.162
	Huynh-Feldt	1.843	1.273	1.447	2.005	.161
	Lower-bound	1.843	1.000	1.843	2.005	.166
Error(System" Session)	Sphericity Assumed	31.241	68	.459		
	Greenhouse-Geisser	31.241	41.320	.756		
	Huynh-Feldt	31.241	43.299	.722		
	Lower-bound	31.241	34.000	.919		

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	System	Session	Type III Sum of Squares	df	Mean Square
System	Linear	Session	.667	1	.667
System * Age	Linear	Session	1.852	1	1.852
Error(System)	Linear	Session	14.148	34	.416
Session	System * Session	Linear	15.340	1	15.340
		Quadratic	2.225	1	2.225
Session * Age	System * Session	Linear	6.674	1	6.674
		Quadratic	1.225	1	1.225

Tests of Within-Subjects Contrasts

Source	System	Session	F	Siq.
System	Linear	Session	1.602	.214
System * Age	Linear	Session	4.450	.042
Error(System)	Linear	Session		
Session	System * Session	Linear	12.804	.001
		Quadratic	**5.759**	.022
Session * Age	System * Session	Linear	**5.570**	.024
		Quadratic	3.170	.084

Tests of Within-Subjects Contrasts

Source	System	Session	Type III Sum of Squares	df	Mean Square
Error(Session)	System * Session	Linear	**40.736**	34	1.198
		Quadratic	13.134	34	**.386**
System * Session	Linear	Linear	.062	1	.062
		Quadratic	.188	1	.188
System * Session * Age	Linear	Linear	**1.563**	1	**1.563**
		Quadratic	.280	1	.280
Error(System*Session)	Linear	Linear	20.125	34	**.592**
		Quadratic	11.116	34	.327

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	System	Session	F	Siq.
Error(Session)	System * Session	Linear		
		Quadratic		
System * Session	Linear	Linear	.106	.747
		Quadratic	.574	.454
System * Session * Age	Linear	Linear	2.640	.113
		Quadratic	**.857**	**.361**
Error(System*Session)	Linear	Linear		
		Quadratic		

Tests of Between-Subjects Effects Measure:MEASURE 1

Source	Type III Sum of Squares	df	Mean Square	F	Siq.
Intercept	**4835.574**	1	**4835.574**	**5406.890**	.000
Age	**6.685**	1	**6.685**	7.475	.010

Tests of Between-Subjects Effects

Measure:MEASURE _1 Transformed Variable:Aver

Sour ce	Type III Sum of Squares	df	Mean Square	F	Siq.
Error	30.407	34	.894		

Estimated Marginal Means

1. Grand Mean

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
4.731	.064	4.601	4.862

2. Age

Estimates

Measure:MEASURE 1

Age	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	4.907	.091	4.722	5.092
Old	4.556	.091	4.371	4.740

Pairwise Comparisons Measure:MEASURE 1

(Il Age	(J) Age	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Young	Old	.352	.129	.010	.090	.613
Old	Young	-.352 ~	.129	.010	-.613	-.090

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

eh

tests the effect of Age. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

Sum of Squares	df	Mean Square	F	Sic
		iiA	7 Ala	

Univariate Tests

M easu
r :MEASURE 1

Sum of Squares	df	Mean Square	F	Sip.

rT~e F tests the effect of Age. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

3. System

Estimates

Measure:MEASURE 1

System	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.787	.077	4.630	4.944
2	4.676	.078	4.517	4.835

Pairwise Comparisons

Measure:MEASURE 1

(I) System	(J) System	Mean Difference (I-J)	Std. Error	Sip.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	.111	.088	.214	-.067	.290
2	1	-.111	.088	.214	-.290	.067

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Siq.
Pillars trace	.045	1.602[a]	1.000	34.000	.214
Wilks'lambda	955	1.602[a]	1.000	34.000	.214
Hotelling's trace	.047	1.602[a]	1.000	34.000	.214
Roy's largest root	.047	1.602[a]	1.000	34.000	.214

Each F tests the multivariate effect of System. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

4. Session

Estimates

Measure:MEASURE 1

Session	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.986	.014	4.958	5.014
2	4.875	.045	4.783	4.967
3	4.333	.179	3.969	4.698

Pairwise Comparisons

Measure:MEASURE 1

(I) Session	(J) Session	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	.111	.049	.090	-.012	.235
	3	.653	.182	.003	.193	1.112
2	1	-.111	.049	.090	-.235	.012
	3	.542	.174	.011	.103	.980
3	1	-.653	.182	.003	-1.112	-.193
	2	-.542	.174	.011	-.980	-.103

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni. *. The mean difference is significant at the .05 level.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.297	6.974[a]	2.000	33.000	.003
Wilks'lambda	.703	6.974[a]	2.000	33.000	.003
Hotelling's trace	.423	6.974[a]	2.000	33.000	.003
Roy's larqest root	.423	6.974[a]	2.000	33.000	.003

Each F tests the multivariate effect of Session. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. Age * System

Measure:MEASURE 1

Age	System	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.870	.110	4.648	5.093
	2	4.944	.111	4.719	5.170
Old	1	4.704	.110	4.481	4.926
	2	4.407	.111	4.182	4.633

6. Age * Session

Age	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	5.000	.020	4.960	5.040
	2	4.944	.064	4.814	5.075
	3	4.778	.254	4.262	**5.293**
Old	1	4.972	.020	4.932	5.012
	2	4.806	.064	4.675	4.936
	3	3.889	.254	**3.373**	4.404

7. System * Session Measure:MEASURE 1

System	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	5.000	.000	5.000	5.000
	2	4.972	.028	4.916	5.029
	3	4.389	.234	3.914	4.864
2	1	4.972	.028	4.916	5.029
	2	4.778	.089	4.596	4.959
	3	4.278	.205	3.862	4.694

8. Age * System * Session Measure:MEASURE 1

Age	System	Session	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	1	1	5.000	.000	5.000	5.000
		2	5.000	.039	4.920	5.080
		3	4.611	**.330**	3.940	5.282
	2	1	5.000	.039	4.920	5.080
		2	4.889	.126	4.632	5.146
		3	4.944	.289	4.357	**5.532**
Old	1	1	5.000	.000	5.000	5.000
		2	4.944	.039	4.865	5.024
		3	4.167	**.330**	**3.495**	4.838
	2	1	4.944	.039	4.865	5.024
		2	4.667	.126	4.410	4.924

3	3.611	.289	3.023	4.199

SET 2

General Linear Model

Notes

Output Created		**2011-02-10T08:52:29.309**
Comments		
Input	Data	**C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 1 \Results\Spreadsheet-Attempts. sav**
		DataSetI
		<none> <none>
		<none> 36
	Active Dataset	User-defined missing values are treated **as** missing.
	Filter	Statistics are **based on** all **cases** with valid data for all variables in the model.
	Weight	
	Split File	
	N of Rows in Working Data File	
Missing Value Handling	**Definition of Missing**	
	Cases Used	

Notes

Syntax		GLM S2F2 S3F2 S2P2 S3P2 BY Age /WSFACTOR=System 2 Polynomial Session 2 Polynomial /METHOD=SSTYPE(3) /EMM EANS=TABLES(OVERALL) /EMM EANS=TABLES(Agge) COMPARE ADJ(BONFERRONI) /EMM EANS=TABLES(System) COMPARE ADJ(BONFERRONI) /EMM EANS=TABLES(Session) COMPARE ADJ(BONFERRONI) /EMMEANS=TABLES(Age" System) /EMM EANS=TABLES (A e" Session) /EMM EANS=TABLES (System*Session) /EMM EANS=TABLES Age*System" Session /PRINT=DESCRIPTIVE /CRITERIA=A L P HA(.05) /WSDESIGN=System Session System*Session /DESIGN=Age.
Resources	Processor Time	0:00:00.157 0:00:00.155
	Elapsed Time	

[DataSetl] C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 1\Results\ Spreadsheet-Attempts.sav

Within-Subjects Factors

Measure:MEASURE 1

Syste m	**Bess** ion	Dependent Variable
1	1	S2F2
	2	53F2
2	1	S2P2
	2	53P2

Between-Subjects Factors

		Value Label	N
Age	1	Young	18
	2	Old	18

Descriptive Statistics

	Age	Mean	Std. Deviation	N
52F2	Young	4.9444	.23570	18
	Old	4.8333	.38348	18
	Total	4.8889	.31873	36
53F2	Young	4.6667	1.18818	18

Descriptive Statistics

	Age	Mean	Std. Deviation	N
S3 F2	Old	3.6667	1.97037	18
	Total	4.1667	1.68184	36
S2P2	Young	5.0000	.00000	18
	Old	4.8333	.38348	18
	Total	4.9167	.28031	36
S3P2	Young	4.8889	.32338	18
	Old	3.1667	2.17607	18
	Total	4.0278	1.76451	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
System	Pillai's Trace	.003	.088[a]	1.000	34.000	.769
	Wilks' Lambda	.997	.088[a]	1.000	34.000	.769
	Hotelling's Trace	.003	.088[a]	1.000	34.000	.769
	Roy's Largest Root	.003	.088[a]	1.000	34.000	.769
System * Age	Pillai's Trace	.031	1.076[a]	1.000	34.000	.307
	Wilks' Lambda	.969	1.076[a]	1.000	34.000	.307
	Hotelling's Trace	.032	1.076[a]	1.000	34.000	.307
	Roy's Largest Root	.032	1.076[a]	1.000	34.000	.307
Session	Pillai's Trace	.354	18.604[a]	1.000	34.000	.000
	Wilks' Lambda	.646	18.604[a]	1.000	34.000	.000
	Hotelling's Trace	.547	18.604[a]	1.000	34.000	.000
	Roy's Largest Root	547	18.604 a	1.000	34.000	.000
Session * Age	Pillai's Trace	.239	10.707[a]	1.000	34.000	.002
	Wilks' Lambda	.761	10.707[a]	1.000	34.000	.002
	Hotelling's Trace	.315	10.707[a]	1.000	34.000	.002
	Roy's Largest Root	.315	10.707 a	1.000	34.000	.002
System * Session	Pillai's Trace	.006	.217[a]	1.000	34.000	.645
	Wilks' Lambda	.994	.217[a]	1.000	34.000	.645
	Hotelling's Trace	.006	.217[a]	1.000	34.000	.645
	Roy's Largest Root	.006	.217[a]	1.000	34.000	.645
System * Session * Age	Pillai's Trace	.025	.866[a]	1.000	34.000	.359
	Wilks' Lambda	.975	.866[a]	1.000	34.000	.359
	Hotelling's Trace	.025	.866[a]	1.000	34.000	.359
	Roy's Largest Root	.025	.866[a]	1.000	34.000	.359

a. Exact statistic

b. Design: Intercept + Age Within Subjects Design: System + Session + System * Session

Mauchly's Test of Sphericity[b] Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
System	1.000	.000	0		1.000	1.000	1.000
Session	1.000	.000	0		1.000	1.000	1.000
c.-+- * c-;-	I nnn	nnn	n		I nnn	I nnn	I nnn

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + Age Within Subjects Design: System + Session + System * Session

Tests of Within-Subjects Effects Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
System	Sphericity Assumed	.111	1	.111	.088	.769
	Greenhouse-Geisser	.111	1.000	.111	.088	.769
	Huynh-Feldt	.111	1.000	.111	.088	.769
	Lower-bound	.111	1.000	.111	.088	.769
System * Age	Sphericity Assumed	1.361	1	1.361	1.076	.307
	Greenhouse-Geisser	1.361	1.000	1.361	1.076	.307
	Huynh-Feldt	1.361	1.000	1.361	1.076	.307
	Lower-bound	1.361	1.000	1.361	1.076	.307
Error(System)	Sphericity Assumed	43.028	34	1.266		
	Greenhouse-Geisser	43.028	34.000	1.266		
	Huynh-Feldt	43.028	34.000	1.266		
	Lower-bound	43.028	34.000	1.266		
Session	Sphericity Assumed	23.361	1	23.361	18.604	.000
	Greenhouse-Geisser	23.361	1.000	23.361	18.604	.000
	Huynh-Feldt	23.361	1.000	23.361	18.604	.000
	Lower-bound	23.361	1.000	23.361	18.604	.000
Session * Age	Sphericity Assumed	13.444	1	13.444	10.707	.002
	Greenhouse-Geisser	13.444	1.000	13.444	10.707	.002
	Huynh-Feldt	13.444	1.000	13.444	10.707	.002
	Lower-bound	13.444	1.000	13.444	10.707	.002
Error(Session)	Sphericity Assumed	42.694	34	1.256		
	Greenhouse-Geisser	42.694	34.000	1.256		
	Huynh-Feldt	42.694	34.000	1.256		
	Lower-bound	42.694	34.000	1.256		
System * Session	Sphericity Assumed	.250	1	.250	.217	.645

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
System * Session	Greenhouse-Geisser	.250	1.000	.250	.217	.645
	Huynh-Feldt	.250	1.000	.250	.217	.645
	Lower-bound	.250	1.000	.250	.217	.645
System * Session * Age	Sphericity Assumed	1.000	1	1.000	.866	.359
	Greenhouse-Geisser	1.000	1.000	1.000	.866	.359
	Huynh-Feldt	1.000	1.000	1.000	.866	.359
	Lower-bound	1.000	1.000	1.000	.866	.359
Error(System*Session)	Sphericity Assumed	39.250	34	1.154		
	Greenhouse-Geisser	39.250	34.000	1.154		
	Huynh-Feldt	39.250	34.000	1.154		
	Lower-bound	39.250	34.000	1.154		

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	System	Session	Type III Sum of Squares	df	Mean Square	F	Sig.
System	Linear	Session	.111	1	.111	.088	.769
System * Age	Linear	Session	1.361	1	1.361	1.076	.307
Error(System)	Linear	Session	43.028	34	1.266		
Session	System * Session	Linear	23.361	1	23.361	18.604	.000
Session * Age	System * Session	Linear	13.444	1	13.444	10.707	.002
Error(Session)	System * Session	Linear	42.694	34	1.256		
System * Session	Linear	Linear	.250	1	.250	.217	.645
System * Session * Age	Linear	Linear	1.000	1	1.000	.866	.359
Error(System*Session)	Linear	Linear	39.250	34	1.154		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed
Variable:Aver

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	2916.000	1	2916.000	1861.859	.000
Age	20.250	1	20.250	12.930	.001
Error	53.250	34	1.566		

Estimated Marginal Means 2. Age

Estimates

Measure:MEASURE 1

Age	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	4.875	.147	4.575	5.175
Old	4.125	.147	3.825	4.425

Pairwise Comparisons Measure:MEA

SURE 1

(Il Age	(J) Age	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Young	Old	.750	.209	.001	.326	1.174
Old	Young	-.750	.209	.001	-1.174	-.326

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sip.
Contrast	5.062	1	5.062	12.930	.001
Prrnr	**13 313**	34	392		

The F tests theeffect of Age. This test is basedon thelinearly independent pairwise comparisons among the estimated marginal means.

3. System

Estimates

Measure:MEASURE 1

Syste m	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.528	.148	4.228	4.828
2	4.472	.132	4.203	4.741

Pairwise Comparisons

Measure:MEASURE 1

(I) System	(J) System	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	.056	.187	.769	-.325	.437
2	1	-.056	.187	.769	-.437	.325

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.003	.088[a]	1.000	34.000	.769
Wilks'lambda	997	•088[a]	1.000	34.000	.769
Hotelling's trace	.003	.088[a]	1.000	34.000	.769
Roy's largest root	.003	•088[a]	1.000	34.000	.769

Each F tests the multivariate effect of System. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

4. Session

Estimates

Measure:MEASURE 1

Session	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.903	.038	4.826	4.979
2	4.097	.194	3.702	4.492

Pairwise Comparisons

Measure:MEASURE 1

(I) Session	(J) Session	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	.806	.187	.000	.426	1.185
2	1	-.806	.187	.000	-1.185	-.426

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	354	18.604[a]	1.000	34.000	.000
Wilks'lambda	.646	18.604[a]	1.000	34.000	.000
Hotelling's trace	.547	18.604[a]	1.000	34.000	.000
Roy's largest root	.547	18.604[a]	1.000	34.000	.000

Each F tests the multivariate effect of Session. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. **Age * System**

Measure:MEASURE 1

Age	System	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.806	.209	4.381	5.230
	2	4.944	.187	4.564	5.325
Old	1	4.250	.209	3.826	4.674
	2	4.000	.187	3.620	4.380

6. **Age * Session**

Measure:MEASURE 1

Age	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.972	.053	4.864	5.080
	2	4.778	.275	4.219	5.336
Old	1	4.833	.053	4.725	4.942
	2	3.417	.275	2.858	3.975

7. **System * Session**

Measure:MEASURE 1

System	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	4.889	.053	4.781	4.997
	2	4.167	.271	3.616	4.718
2	1	4.917	.045	4.825	5.009
	2	4.028	.259	3.501	4.555

8. **Age * System * Session**

Measure:MEASURE 1

Age	System	Session	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	1	1	4.944	.075	4.792	5.097

8. Age * System * Session

Syste Age	m	Sess ion	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	1	2	4.667	**.383**	3.887	5.446
2		1	5.000	.064	4.870	**5.130**
		2	4.889	.367	4.144	**5.634**
Old	1	1	**4.833**	.075	4.681	4.986
		2	3.667	**.383**	2.887	4.446
2		1	**4.833**	.064	4.703	4.963
		2	3.167	.367	2.422	3.912

SET 1

General Linear Model

Notes

Output Created		**2011-02-10T08:58:41.180**
Comments		
Input	Data	**C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 1 \Results\Spreadsheet-TimeAve. sav**
	Active Dataset	**DataSetI <none>**
	Filter	**<none>**
	Weight	**<none>**
	Split File	**<none>**
	N of Rows in Working Data File	**36**
Missing Value Handling	**Definition of Missing**	User-defined missing values are treated **as** missing.
	Cases Used	Statistics are **based on** all **cases** with valid data for all variables in the model.

Notes

Syntax		GLM S1 F1 S2F1 S3F1 S1 P1 S2P1 S3P1 BY Age /WSFACTOR=System 2 Polynomial Session 3 Polynomial /METHOD=SSTYPE(3) /EMM EANS=TABLES(OVERALL) /EMM EANS=TABLES(Age) COMPARE ADJ(BONFERRONI) /EMM EANS=TABLES(System) COMPARE ADJ(BONFERRONI) /EMM EANS=TABLES(Session) COMPARE ADJ(BONFERRONI) /EMMEANS=TABLES(Age" System) /EMM EANS=TABLES (Age" Session) /EMM EANS=TABLES (System*Session) /EMM EANS=TABLES (Age" System" Session) /PRINT=DESCRIPTIVE /CRITERIA=A L P HA(.05) /WSDESIGN=System Session System*Session /DESIGN=Age.
Resources	Processor Time	0:00:00.187 0:00:00.108
	Elapsed Time	

[DataSetl] C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 1\Results\ Spreadsheet-TimeAve.sav

Within-Subjects Factors

Measure:MEASURE 1

System	Session	Dependent Variable
1	1	S1 F1
	2	52F1
	3	53F1
2	1	S1 P1
	2	52P1
	3	53P1

Between- Subjects Factors

		Value Label	N
Age	1	Young	18
	2	Old	18

Descriptive Statistics

	Age	Mean	Std. Deviation	N
S1 F1	Young	11.8225	.85141	18

Descriptive Statistics

	Age	Mean	Std. Deviation	N
S1 F1	Old	16.6664	2.57368	18
	Total	14.2445	3.09884	36
S2F1	Young	12.3014	1.31703	18
	Old	15.9811	2.66011	18
	Total	14.1412	2.78592	36
S3 F1	Young	15.9148	4.92832	18
	Old	21.6618	6.01538	18
	Total	18.7883	6.15351	36
S1 P1	Young	11.5600	1.44886	18
	Old	15.2979	2.47269	18
	Total	13.4290	2.75357	36
S2P1	Young	12.8210	2.01832	18
	Old	18.6535	4.13740	18
	Total	15.7373	4.36353	36
S3 P1	Young	12.8614	2.05233	18
	Old	23.5168	9.39208	18
	Total	18.1891	8.60736	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
System	Pillars Trace	.000	.017[a]	1.000	34.000	.898
	Wilks' Lambda	1.000	.017[a]	1.000	34.000	.898
	Hotelling's Trace	.000	.017[a]	1.000	34.000	.898
	Roy's Largest Root	.000	.017[a]	1.000	34.000	.898
System * Age	Pillars Trace	.118	4.538[a]	1.000	34.000	.040
	Wilks' Lambda	.882	4.538[a]	1.000	34.000	.040
	Hotelling's Trace	.133	4.538[a]	1.000	34.000	.040
	Roy's Largest Root	.133	4.538[a]	1.000	34.000	.040
Session	Pillars Trace	.492	15.951 a	2.000	33.000	.000
	Wilks' Lambda	.508	15.951 a	2.000	33.000	.000
	Hotelling's Trace	.967	15.951 a	2.000	33.000	.000
	Roy's Largest Root	.967	15.951 a	2.000	33.000	.000
Session * Age	Pillars Trace	.139	2.661 a	2.000	33.000	.085
	Wilks' Lambda	.861	2.661 a	2.000	33.000	.085
	Hotelling's Trace	.161	2.661 a	2.000	33.000	.085
	Roy's Largest Root	.161	2.661 a	2.000	33.000	.085

a. Exact statistic

b. Design: Intercept + Age Within Subjects Design: System + Session + System * Session

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Siq.
System * Session	Pillai's Trace	.413	11.624[a]	2.000	33.000	.000
	Wilks' Lambda	.587	11.624[a]	2.000	33.000	.000
	Hotelling's Trace	.704	11.624[a]	2.000	33.000	.000
	Roy's Largest Root	.704	11.624 a	2.000	33.000	.000
System * Session * Age	Pillai's Trace	.260	5.788[a]	2.000	33.000	.007
	Wilks' Lambda	.740	5.788[a]	2.000	33.000	.007
	Hotelling's Trace	.351	5.788[a]	2.000	33.000	.007
	Roy's Largest Root	.351	5.788[a]	2.000	33.000	.007

a. Exact statistic b. Design: Intercept + Age

Within Subjects Design: System + Session + System * Session

Mauchly's Test of Sphericity Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Siq.	Epsilon a		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
System	1.000	.000	0		1.000	1.000	1.000
Session	.442	26.928	2	.000	.642	.676	.500
c.-+- * c-;-	cn9	**99 7R9**	9	nnn	RRR	7nc	cnn

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + Age Within Subjects Design: System + Session + System * Session

Tests of Within-Subjects Effects Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Siq.
System	Sphericity Assumed	.197	1	.197	.017	.898
	Greenhouse-Geisser	.197	1.000	.197	.017	.898
	Huynh-Feldt	.197	1.000	.197	.017	.898
	Lower-bound	.197	1.000	.197	.017	.898
System * Age	Sphericity Assumed	53.194	1	53.194	4.538	.040
	Greenhouse-Geisser	53.194	1.000	53.194	4.538	.040
	Huynh-Feldt	53.194	1.000	53.194	4.538	.040
	Lower-bound	53.194	1.000	53.194	4.538	.040
Error(System)	Sphericity Assumed	398.520	34	11.721		
	Greenhouse-Geisser	398.520	34.000	11.721		
	Huynh-Feldt	398.520	34.000	11.721		

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Error(System)	Lower-bound	398.520	34.000	11.721		
Session	Sphericity Assumed	850.920	2	425.460	24.376	.000
	Greenhouse-Geisser	850.920	1.284	662.782	24.376	.000
	Huynh-Feldt	850.920	1.352	629.567	24.376	.000
	Lower-bound	850.920	1.000	850.920	24.376	.000
Session * Age	Sphericity Assumed	164.255	2	82.128	4.705	.012
	Greenhouse-Geisser	164.255	1.284	127.938	4.705	.027
	Huynh-Feldt	164.255	1.352	121.527	4.705	.025
	Lower-bound	164.255	1.000	164.255	4.705	.037
Error(Session)	Sphericity Assumed	1186.886	68	17.454		
	Greenhouse-Geisser	1186.886	43.651	27.190		
	Huynh-Feldt	1186.886	45.954	25.828		
	Lower-bound	1186.886	34.000	34.908		
System * Session	Sphericity Assumed	64.088	2	32.044	3.918	.025
	Greenhouse-Geisser	64.088	1.335	47.997	3.918	.043
	Huynh-Feldt	64.088	1.410	45.441	3.918	.040
	Lower-bound	64.088	1.000	64.088	3.918	.056
System * Session * Age	Sphericity Assumed	81.578	2	40.789	4.987	.010
	Greenhouse-Geisser	81.578	1.335	61.096	4.987	.021
	Huynh-Feldt	81.578	1.410	57.842	4.987	.020
	Lower-bound	81.578	1.000	81.578	4.987	.032
Error(System" Session)	Sphericity Assumed	556.197	68	8.179		
	Greenhouse-Geisser	556.197	45.398	12.251		
	Huynh-Feldt	556.197	47.952	11.599		
	Lower-bound	556.197	34.000	16.359		

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	System	Session	Type III Sum of Squares	df	Mean Square
System	Linear	Session	.197	1	.197
System * Age	Linear	Session	53.194	1	53.194
Error(System)	Linear	Session	398.520	34	11.721
Session	System * Session	Linear	779.072	1	779.072
		Quadratic	71.848	1	71.848
Session * Age	System * Session	Linear	137.615	1	137.615
		Quadratic	26.641	1	26.641
Error(Session)	System * Session	Linear	855.371	34	25.158

Tests of Within-Subjects Contrasts

Source	System	Session	F	Siq.
System	Linear	Session	.017	.898
System * Age	Linear	Session	4.538	.040
Error(System)	Linear	Session		
Session	System * Session	Linear	30.967	.000
		Quadratic	7.369	.010
Session * Age	System * Session	Linear	5.470	.025
		Quadratic	2.732	.108
Error(Session)	System * Session	Linear		

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	System	Session	Type III Sum of Squares	df	Mean Square
Error(Session)	System * Session	Quadratic	331.516	34	9.750
System * Session	Linear	Linear	.421	1	.421
		Quadratic	63.667	1	63.667
System * Session * Age	Linear	Linear	81.388	1	81.388
		Quadratic	.190	1	.190
Error(System*Session)	Linear	Linear	404.275	34	11.890
		Quadratic	151.922	34	4.468

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	System	Session	F	Siq.
Error(Session)	System * Session	Quadratic		
System * Session	Linear	Linear	.035	.852
		Quadratic	14.249	.001
System * Session * Age	Linear	Linear	6.845	.013
		Quadratic	.042	.838
Error(System*Session)	Linear	Linear		
		Quadratic		

Tests of Between-Subjects Effects Measure:MEASURE 1

Source	Type III Sum of Squares	df	Mean Square	F	Siq.
Intercept	53614.758	1	53614.758	1479.349	.000
Age	1785.004	1	1785.004	49.252	.000
Error	1232.233	34	36.242		

Estimated Marginal Means

1. Grand Mean

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
15.755	.410	14.922	16.587

2. Age

Estimates

Measure:MEASURE 1

Age	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	12.880	.579	11.703	14.057
Old	18.630	.579	17.452	19.807

Pairwise Comparisons Measure:MEA

SURE 1

(Il Age	(J) Age	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Young	Old	-5.749	.819	.000	-7.414	-4.085
Old	Young	5.749	.819	.000	4.085	7.414

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Siq.
Contrast	297.501	1	297.501	49.252	.000
Prrnr	**2nFi 372**	34	R n4n		

The F tests the effect of Age. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

3. System

Estimates

Measure:MEASURE 1

System	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	15.725	.462	14.786	16.663
2	15.785	.480	14.809	16.762

Pairwise Comparisons

Measure:MEASURE 1

(I) System	(J) System	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	-.060	.466	.898	-1.007	.886
2	1	.060	.466	.898	-.886	1.007

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Siq.
Pillars trace	.000	•017[a]	1.000	34.000	.898
Wilks'lambda	1.000	•017[a]	1.000	34.000	.898
Hotelling's trace	.000	•017[a]	1.000	34.000	.898
Roy's larqest root	.000	•017[a]	1.000	34.000	.898

Each F tests the multivariate effect of System. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

4. Session

Estimates

Measure:MEASURE 1

Session	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	13.837	.296	13.236	14.438
2	14.939	.392	14.143	15.735
3	18.489	.864	16.732	20.245

Pairwise Comparisons

Measure:MEASURE 1

(I) Session	(J) Session	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	-1.103	.353	.011	-1.992	-.213
	3	-4.652	.836	.000	-6.757	-2.547
2	1	1.103 ~	.353	.011	.213	1.992
	3	-3.549	.794	.000	-5.550	-1.549
3	1	4.652	.836	.000	2.547	6.757
	2	3.549	.794	.000	1.549	5.550

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.492	15.951 a	2.000	33.000	.000
Wilks' lambda	.508	15.951 a	2.000	33.000	.000
Hotelling's trace	.967	15.951 a	2.000	33.000	.000
Roy's largest root	.967	15.951 a	2.000	33.000	.000

Each F tests the multivariate effect of Session. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. **Age * System**

Measure:MEASURE 1

Age	System	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	13.346	.653	12.019	14.673
	2	12.414	.680	11.033	13.795
Old	1	18.103	.653	16.776	19.430
	2	19.156	.680	17.775	20.537

6. Age * Session

Age	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	11.691	.418	10.841	12.541
	2	12.561	.554	11.435	13.687
	3	14.388	1.223	11.904	16.873

6. Age * Session

Age	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Old	1	**15.982**	**.** 418	15.132	16. 832
	2	17.317	**.554**	16.192	18.443
	3	**22.589**	1.223	20.105	25.074

7. System * Session

System	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	14.244	.319	**13.595**	14.894
	2	14.141	**.350**	13.430	14.852
	3	18.788	.916	16.926	20.651
2	1	13.429	**.338**	12.743	14.115
	2	15.737	**.543**	**14.635**	16.840
	3	18.189	1.133	15.887	20.492

8. Age * System * Session Measure:MEASURE 1

Age	System	Session	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	1	1	11.823	.452	10.904	12.741
		2	12.301	.495	11.296	13.307
		3	**15.915**	1.296	13.281	18.549
	2	1	11.560	.478	10.589	12.531
		2	12.821	.767	11.262	14.380
		3	12.861	1.602	9.605	16.118
Old	1	1	16.666	.452	15.748	**17.585**
		2	15.981	.495	14.976	16.986
		3	21.662	1.296	19.028	24.296
	2	1	15.298	.478	14.327	16.269
		2	**18.653**	.767	17.094	20.213
		3	23.517	1.602	20.261	26.773

SET 2

General Linear Model

Notes

Output Created		**2011-02-10T09:00:08.206**
Comments		
Input	Data	**C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 1 \Results\Spreadsheet-TimeAve. sav**
	Active Dataset	**DataSetI <none>**
	Filter	**<none> <none>**
	Weight	**36**
	Split File	User-defined missing values are treated **as** missing. Statistics are **based on** all **cases** with valid data for all variables in the model.
	N of Rows in Working Data File	
Missing Value Handling	**Definition of Missing**	
	Cases Used	

Notes

Syntax		GLM S2F2 S3F2 S2P2 S3P2 BY Age /WSFACTOR=System 2 Polynomial Session 2 Polynomial /METHOD=SSTYPE(3) /EMM EANS=TABLES(OVERALL) /EMM EANS=TABLES(Agge) COMPARE ADJ(BONFERRONI) /EMM EANS=TABLES(System) COMPARE ADJ(BONFERRONI) /EMM EANS=TABLES(Session) COMPARE ADJ(BONFERRONI) /EMMEANS=TABLES(Age" System) /EMM EANS=TABLES (A e" Session) /EMM EANS=TABLES (System*Session) /EMM EANS=TABLES Age*System" Session /PRINT=DESCRIPTIVE /CRITERIA=A L P HA(.05) /WSDESIGN=System Session System*Session /DESIGN=Age.
Resources	Processor Time	0:00:00.156 0:00:00.078
	Elapsed Time	

[DataSetl] C:\Documents and Settings\u028488\My Documents\My Dropbox\PhD\Study 1\Results\ Spreadsheet-TimeAve.sav

Within-Subjects Factors

Measure:MEASURE 1

Syste m	**Bess** ion	Dependent Variable
1	1	S2F2
	2	53F2
2	1	S2P2
	2	53P2

Between-Subjects Factors

		Value Label	N
Age	1	Young	18
	2	Old	18

Descriptive Statistics

	Age	Mean	Std. Deviation	N
52F2	Young	14.5226	2.24822	18
	Old	19.2140	3.61751	18
	Total	16.8683	3.80407	36
53F2	Young	16.0921	3.29972	18

Descriptive Statistics

	Age	Mean	Std. Deviation	N
S3 F2	Old	22.6831	6.07457	18
	Total	19.3876	5.86363	36
S2P2	Young	12.6232	2.42833	18
	Old	17.0416	4.22514	18
	Total	14.8324	4.06877	36
S3P2	Young	14.8161	2.58684	18
	Old	24.5045	6.99650	18
	Total	19.6603	7.15282	36

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
System	Pillai's Trace	.047	1.661 a	1.000	34.000	.206
	Wilks' Lambda	.953	1.661 a	1.000	34.000	.206
	Hotelling's Trace	.049	1.661 a	1.000	34.000	.206
	Roy's Largest Root	.049	1.661 a	1.000	34.000	.206
System * Age	Pillai's Trace	.030	1.065[a]	1.000	34.000	.309
	Wilks' Lambda	.970	1.065[a]	1.000	34.000	.309
	Hotelling's Trace	.031	1.065[a]	1.000	34.000	.309
	Roy's Largest Root	.031	1.065[a]	1.000	34.000	.309
Session	Pillai's Trace	.466	29.694[a]	1.000	34.000	.000
	Wilks' Lambda	.534	29.694[a]	1.000	34.000	.000
	Hotelling's Trace	.873	29.694[a]	1.000	34.000	.000
	Roy's Largest Root	.873	29.694 a	1.000	34.000	.000
Session * Age	Pillai's Trace	.172	7.069[a]	1.000	34.000	.012
	Wilks' Lambda	.828	7.069[a]	1.000	34.000	.012
	Hotelling's Trace	.208	7.069[a]	1.000	34.000	.012
	Roy's Largest Root	.208	7.069[a]	1.000	34.000	.012
System * Session	Pillai's Trace	.109	4.143[a]	1.000	34.000	.050
	Wilks' Lambda	.891	4.143[a]	1.000	34.000	.050
	Hotelling's Trace	.122	4.143[a]	1.000	34.000	.050
	Roy's Largest Root	.122	4.143 a	1.000	34.000	.050
System * Session * Age	Pillai's Trace	.061	2.208[a]	1.000	34.000	.147
	Wilks' Lambda	.939	2.208[a]	1.000	34.000	.147
	Hotelling's Trace	.065	2.208[a]	1.000	34.000	.147
	Roy's Largest Root	.065	2.208[a]	1.000	34.000	.147

a. Exact statistic

b. Design: Intercept + Age Within Subjects Design: System + Session + System * Session

Mauchly's Test of Sphericity[b] Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
System	1.000	.000	0		1.000	1.000	1.000
Session	1.000	.000	0		1.000	1.000	1.000
c.-+- * c-;-	I nnn	nnn	n		I nnn	I nnn	I nnn

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + Age Within Subjects Design: System + Session + System * Session

Tests of Within-Subjects Effects Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
System	Sphericity Assumed	27.979	1	27.979	1.661	.206
	Greenhouse-Geisser	27.979	1.000	27.979	1.661	.206
	Huynh-Feldt	27.979	1.000	27.979	1.661	.206
	Lower-bound	27.979	1.000	27.979	1.661	.206
System * Age	Sphericity Assumed	17.947	1	17.947	1.065	.309
	Greenhouse-Geisser	17.947	1.000	17.947	1.065	.309
	Huynh-Feldt	17.947	1.000	17.947	1.065	.309
	Lower-bound	17.947	1.000	17.947	1.065	.309
Error(System)	Sphericity Assumed	572.794	34	16.847		
	Greenhouse-Geisser	572.794	34.000	16.847		
	Huynh-Feldt	572.794	34.000	16.847		
	Lower-bound	572.794	34.000	16.847		
Session	Sphericity Assumed	485.835	1	485.835	29.694	.000
	Greenhouse-Geisser	485.835	1.000	485.835	29.694	.000
	Huynh-Feldt	485.835	1.000	485.835	29.694	.000
	Lower-bound	485.835	1.000	485.835	29.694	.000
Session * Age	Sphericity Assumed	115.654	1	115.654	7.069	.012
	Greenhouse-Geisser	115.654	1.000	115.654	7.069	.012
	Huynh-Feldt	115.654	1.000	115.654	7.069	.012
	Lower-bound	115.654	1.000	115.654	7.069	.012
Error(Session)	Sphericity Assumed	556.278	34	16.361		
	Greenhouse-Geisser	556.278	34.000	16.361		
	Huynh-Feldt	556.278	34.000	16.361		
	Lower-bound	556.278	34.000	16.361		
System * Session	Sphericity Assumed	47.965	1	47.965	4.143	.050

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
System * Session	Greenhouse-Geisser	47.965	1.000	47.965	4.143	.050
	Huynh-Feldt	47.965	1.000	47.965	4.143	.050
	Lower-bound	47.965	1.000	47.965	4.143	.050
System * Session * Age	Sphericity Assumed	25.560	1	25.560	2.208	.147
	Greenhouse-Geisser	25.560	1.000	25.560	2.208	.147
	Huynh-Feldt	25.560	1.000	25.560	2.208	.147
	Lower-bound	25.560	1.000	25.560	2.208	.147
Error(System*Session)	Sphericity Assumed	393.662	34	11.578		
	Greenhouse-Geisser	393.662	34.000	11.578		
	Huynh-Feldt	393.662	34.000	11.578		
	Lower-bound	393.662	34.000	11.578		

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	System	Session	Type III Sum of Squares	df	Mean Square	F	Sig.
System	Linear	Session	27.979	1	27.979	1.661	.206
System * Age	Linear	Session	17.947	1	17.947	1.065	.309
Error(System)	Linear	Session	572.794	34	16.847		
Session	System * Session	Linear	485.835	1	485.835	29.694	.000
Session * Age	System * Session	Linear	115.654	1	115.654	7.069	.012
Error(Session)	System * Session	Linear	556.278	34	16.361		
System * Session	Linear	Linear	47.965	1	47.965	4.143	.050
System * Session * Age	Linear	Linear	25.560	1	25.560	2.208	.147
Error(System*Session)	Linear	Linear	393.662	34	11.578		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed
Variable:Aver

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	45048.297	1	45048.297	1616.136	.000
Age	1450.368	1	1450.368	52.033	.000
Error	947.719	34	27.874		

Estimated Marginal Means 2. Age

Estimates

Measure:MEASURE 1

Age	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	14.514	.622	13.249	15.778
Old	20.861	.622	19.596	22.125

Pairwise Comparisons Measure:MEA

SURE 1

(Il Age	(J) Age	Mean Difference (I-J)	Std. Error	Siq.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
Young	Old	-6.347	.880	.000	-8.136	-4.559
Old	Young	6.347 *	.880	.000	4.559	8.136

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sip.
Contrast	362.592	1	362.592	52.033	.000
Prrnr	***23R q30***	34	*qRq*		

The F tests theeffect of Age. This test is basedon thelinearly independent pairwise comparisons among the estimated marginal means.

3. System

Estimates

Measure:MEASURE 1

Syste m	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	18.128	.542	17.026	19.230
2	17.246	.572	16.084	18.409

Pairwise Comparisons

Measure:MEASURE 1

(I) System	(J) System	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	.882	.684	.206	-.509	2.272
2	1	-.882	.684	.206	-2.272	.509

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.047	1.661 a	1.000	34.000	.206
Wilks'lambda	.953	1.661 a	1.000	34.000	.206
Hotelling's trace	.049	1.661 a	1.000	34.000	.206
Roy's largest root	.049	1.661 a	1.000	34.000	.206

Each F tests the multivariate effect of System. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

4. Session

Estimates

Measure:MEASURE 1

Session	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	15.850	.444	14.949	16.752
2	19.524	.646	18.211	20.837

Pairwise Comparisons

Measure:MEASURE 1

(I) Session	(J) Session	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Inerval for Difference	
					Lower Bound	Upper Bound
1	2	-3.674	.674	.000	-5.044	-2.304
2	1	3.674	.674	.000	2.304	5.044

Based on estimated marginal means

*. The mean difference is significant at the .05 level. a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.466	29.694[a]	1.000	34.000	.000
Wilks'lambda	.534	29.694[a]	1.000	34.000	.000
Hotelling's trace	.873	29.694[a]	1.000	34.000	.000
Roy's largest root	.873	29.694[a]	1.000	34.000	.000

Each F tests the multivariate effect of Session. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. **Age * System**

Measure:MEASURE 1

Age	System	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	15.307	.767	13.749	16.865
	2	13.720	.809	12.076	15.364
Old	1	20.949	.767	19.391	22.507
	2	20.773	.809	19.129	22.417

6. **Age * Session**

Measure:MEASURE 1

Age	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	13.573	.627	12.298	14.848
	2	15.454	.914	13.597	17.311
Old	1	18.128	.627	16.853	19.403
	2	23.594	.914	21.736	25.451

7. **System * Session**

Measure:MEASURE 1

System	Session	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	16.868	.502	15.848	17.888
	2	19.388	.815	17.732	21.043
2	1	14.832	.574	13.665	16.000
	2	19.660	.879	17.874	21.447

8. **Age * System * Session**

Measure:MEASURE 1

Age	System	Session	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	1	1	14.523	.710	13.080	15.965

8. Age * System * Session

Syste Age	m	Sess ion	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	1	2	16.092	1.152	13.751	18.434
2		1	12.623	.812	**10.973**	14.274
		2	14.816	1.243	12.290	17.343
Old	1	1	19.214	.710	17.771	**20.657**
		2	**22.683**	1.152	20.342	**25.025**
2		1	17.042	.812	**15.391**	18.692
		2	24.504	1.243	21.978	27.031

SET1

Notes

Output Created		18-Jun-2012 01:56:16
Comments		
Input	**Data**	/Users/jjnicholson/Dropbox/PhD/Study 5/Results/St5-Attempts.say
	Active Dataset	DataSet5
	Filter	<none> <none>
	Weight	<none>
	Split File	
	N of Rows in Working Data File	72
Missing Value Handling	**Definition of Missing**	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on all cases with valid data for all variables in the model.
Syntax		GLM SETiSess1 SETISess2 SET1 Sess3 BY PartAge FaceAge /WSFACTOR=Week 3 Polynomial IMETHOD=SSTYPE(3) /EMMEANS=TABLES(OVERALL) IEMMEANS=TABLES(Page) COMPARE ADJ(BONFERRONI) /EMMEANS=TAB LES(FaceAge) COMPARE ADJ(BONFERRONI) IEMMEANS=TABLES(Week) COMPARE ADJ(BONFERRONI) IEMMEANS=TABLES (PartAge*FaceAge) IEMMEANS=TABLES (PartAge*Week) IEMMEANS=TABLES (FaceAge*Week) IEMMEANS=TABLES PartAge*FaceAge*Week /PRINT=DESCRIPTIVE /CRITERIA=A L P H A(.05) /WSDESIGN=Week /DESIGN=PartAge FaceAge PartAge*FaceAge.
Resources	**Processor Time**	00:00:00.031
	Elapsed Time	00:00:00.000

Within-Subjects Factors

Measure:MEASURE 1

Week	Dependent Variable
1	SETISessl
2	SETISess2
3	SETISess3

Between-Subjects Factors

		Value Label	N
PartAge	1.00	Young	36
	2.00	Old	3 6
FaceAge	1.00	Young	3 6
	2.00	Old	3 6

Descriptive riptive Statistics

PartAge		FaceAge	Mean	Std. Deviation	N
SETISessi	Young	Young	4.8333	.23570	18
		Old	4.8148	.34721	18
		Total	4.8241	.29262	36
Old		Young	4.0741	1.14650	1 8
		Old	4.7037	.68493	1 8
		Total	4.3889	.98400	36
Total		Young	4.4537	.90204	36
		Old	4.7593	.53814	36
		Total	4.6065	.75335	72
SETISess2	Young	Young	4.9259	.14260	18
		Old	4.7593	.54600	1 8
		Total	4.8426	.40226	36
Old		Young	3.6667	1.61286	1 8
		Old	4.0741	1.38882	1 8
		Total	3.8704	1.49768	36
Total		Young	4.2963	1.29658	36
		Old	4.4167	1.09653	36
		Total	4.3565	1.19378	72
SETISess3	Young	Young	4.9074	.25063	18
		Old	4.7778	.49836	1 8
		Total	4.8426	.39429	36
Old		Young	3.2593	2.03099	18
		Old	3.6852	1.49278	18
		Total	3.4722	1.76990	36
Total		Young	4.0833	1.65304	36
		Old	4.2315	1.22881	36
		Total	4.1574	1.44808	72

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Week	Pillai's Trace	.194	8.059[d]	2.000	67.000	.001
	Wilks' Lambda	.806	8.059[a]	2.000	67.000	.001
	Hotelling's Trace	.241	8.059[a]	2.000	67.000	.001
	Roy's Largest Root	.241	8.059[a]	2.000	67.000	.001
Week * PartAge	Pillai's Trace	.208	8.783[a]	2.000	67.000	.000
	Wilks' Lambda	.792	8.783[a]	2.000	67.000	.000
	Hotelling's Trace	.262	8.783[a]	2.000	67.000	.000
	Roy's Largest Root	.262	8.783[a]	2.000	67.000	.000
Week * FaceAge	Pillai's Trace	.016	.538[d]	2.000	67.000	.587
	Wilks' Lambda	.984	.538[a]	2.000	67.000	.587
	Hotelling's Trace	.016	.538[a]	2.000	67.000	.587
	Roy's Largest Root	.016	.538[a]	2.000	67.000	.587
Week * PartAge * FaceAge	Pillai's Trace	.001	.026[d]	2.000	67.000	.974
	Wilks' Lambda	.999	.026[a]	2.000	67.000	.974
	Hotelling's Trace	.001	.026[a]	2.000	67.000	.974
	Roy's Largest Root	.001	.026[a]	2.000	67.000	.974

a. Exact statistic

b. Design: Intercept + PartAge + FaceAge + PartAge * FaceAge Within Subjects Design: Week

Mauchly's Test of Sphericity[b]

Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Week	.888	7.930	2	.019	.900	.963	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + PartAge + FaceAge + PartAge * FaceAge Within Subjects Design: Week

Tests of Within-Subjects Effects

Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Sphericity Assumed	7.291	2	3.646	10.894	.000
	Greenhouse-Geisser	7.291	1.799	4.053	10.894	.000
	Huynh-Feldt	7.291	1.927	3.785	10.894	.000
	Lower-bound	7.291	1.000	7.291	10.894	.002
Week * PartAge	Sphericity Assumed	7.929	2	3.965	11.847	.000
	Greenhouse-Geisser	7.929	1.799	4.407	11.847	.000
	Huynh-Feldt	7.929	1.927	4.116	11.847	.000
	Lower-bound	7.929	1.000	7.929	11.847	.001
Week * FaceAge	Sphericity Assumed	.359	2	.180	.536	.586
	Greenhouse-Geisser	.359	1.799	.200	.536	.567
	Huynh-Feldt	.359	1.927	.186	.536	.579
	Lower-bound	.359	1.000	.359	.536	.466
Week * PartAge * FaceAge	Sphericity Assumed	.022	2	.011	.032	.968
	Greenhouse-Geisser	.022	1.799	.012	.032	.958
	Huynh-Feldt	.022	1.927	.011	.032	.965
	Lower-bound	.022	1.000	.022	.032	.858
Error(Week)	Sphericity Assumed	45.510	136	.335		
	Greenhouse-Geisser	45.510	122.343	.372		
	Huynh-Feldt	45.510	131.006	.347		
	Lower-bound	45.510	68.000	.669		

Tests of Within-Subjects Contrasts

Measure:MEASURE 1

Source	Week	Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Linear	7.260	1	7.260	16.267	.000
	Quadratic	.031	1	.031	.140	.710
Week * PartAge	Linear	7.871	1	7.871	17.636	.000
	Quadratic	.058	1	.058	.260	.612
Week * FaceAge	Linear	.223	1	.223	.500	.482
	Quadratic	.136	1	.136	.610	.437
Week * PartAge * FaceAge	Linear	.019	1	.019	.043	.836
	Quadratic	.002	1	.002	.010	.919
Error(Week)	Linear	30.349	68	.446		
	Quadratic	15.162	68	.223		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed
Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	4131.459	1	4131.459	1530.781	.000
PartAge	46.296	1	46.296	17.154	.000
FaceAge	1.977	1	1.977	.733	.395
PartAge * FaceAge	4.741	1	4.741	1.757	.189
Error	183.527	68	2.699		

Estimated Marginal Means

1. Grand Mean

Measure:MEASURE 1

		95% Confidence Interval	
Mean	Std. Error	Lower Bound	Upper Bound
4.373	.112	4.150	4.597

!. PartAge

Estimates

Measure:MEASURE 1

PartAge			95% Confidence Interval	
	Mean	Std. Error	Lower Bound	Upper Bound
Young	4.836	.158	4.521	5.152
Old	3.910	.158	3.595	4.226

Pairwise Comparisons

Measure:MEASURE 1

(I) PartAge	(J) PartAge	Mean Difference (I-.!t	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Young	Old	.926	.224	.000	.480	1.372
Old	Young	-.926	.224	.000	-1.372	-.480

Based on estimated mated marginal means *. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	15.432	1	15.432	17.154	.000
Error	61.176	68	.900		

The F tests the effect of PartAge. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

!. FaceAge

Estimates

Measure:MEASURE 1

FaceAge			95% Confidence Interval	
	Mean	Std. Error	Lower Bound	Upper Bound
Youngng	4.278	.158	3.962	4.593
Old	4.469	.158	4.154	4.785

U Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	.659	1	.659	.733	.395
Error	61.176	68	.900		

F tests the effect of FaceAge. This testis based on the linear! y independent pairwise comparisons among the estimated marginal means.

Estimates

Measure:MEASURE 1

Week	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.606	.082	4.442	4.771
2	4	.130	4.098	4.615
3	4.157	.152	3.854	4.461

Pairwise Comparisons

Measure:MEASURE 1

(I) W e e k (J) Week		Mean Difference (I-.!t	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	.250	.089	.019	.032	.468
	3	.449	.111	.000	.176	.722
2	1	-.250	.089	.019	-.468	-.032
	3	.199	.087	.077	-.015	.413
3	1	-.449	.111	.000	-.722	-.176
	2	-.199	.087	.077	-.413	.015

Based on estimated matedmarginal means *. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.194	8.059[d]	2.000	67.000	.001
Wilks' lambda	.806	8.05e	2.000	67.000	.001
Hotelling's trace	.241	8.05e	2.000	67.000	.001
Roy's largest root	.241	8.05e	2.000	67.000	.001

Each F tests the multivariate effect of Week. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. PartAge * FaceAge

Measure:MEASURE 1

PartAge	FaceAge	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	Young	4.889	.224	4.443	5.335
	Old	4.784	.224	4.338	5.230
Old	Young	3.667	.224	3.221	4.113
	Old	4.154	.224	3.708	4.600

6.PartAge* Week

Measure:MEASURE 1

PartAge	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.824	.117	4.591	5.057
	2	4.843	.18383	4.476	5.209
	3	4.843	.215	4.413	5.272
Old	1	4.389	.117	4.156	4.622
	2	3.870	.183	3.504	4.237
	3	3.472	.215	3.043	3.902

7. * Week

Meure:MEASURE_1 ure:MEASURE 1

FaceAge	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
ng Young	1	4.454	.117	4.221	4.686
	2	4.296	.183	3.930	4.662
	3	4.083	.215	3.654	4.513
Old	1	4.759	.117	4.526	4.992
	2	4.417	.183	4.051	4.783
	3	4.231	.215	3.802	4.661

8.PartAge* FaceAge * Week

Meure:MEASURE_1 ure:MEASURE 1

FaceAge		Week	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	Young	1	4.833	.165	4.504	5.163
		2	4.926	.260	4.408	5.444
		3	4.907	.304	4.300	5.515
Old		1	4.815	.165	4.486	5.144
		2	4.759	.260	4.241	5.277
		3	4.778	.304	4.171	5.385
Old	Young	1	4.074	.165	3.745	4.403
		2	3.667	.260	3.149	4.184
		3	3.259	.304	2.652	3.866
Old		1	4.704	.165	4.374	5.033
		2	4.074	.260	3.556	4.592
		3	3.685	.304	3.078	4.292

Notes

[DataSet5] /Users/jjnicholson/Dropbox/PhD/Study 5/Results/St5-Attempts.sav

Output Created		18-Jun-2012 01:57:02
Comments		
Input	**Data**	/Users/jjnicholson/Dropbox/PhD/Study 5/Results/St5-Attempts.say
	Active Dataset	DataSet5
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	72
Missing Value Handling	**Definition of Missing**	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on all cases with valid data for all variables in the model.
Syntax		GLM SET2Sess2 SET2Sess3 BY PartAge FaceAge /WSFACTOR Week 2 Polynomial IMETHOD=SSTYPE(3) /EMMEANS=TAB LES(OVERALL) /EMMEANS=TABLES(PartA e) COMPARE ADJ(BONFERRONIJ /EMMEANS=TAB LES(FaceAge) COMPARE ADJ(BONFERRONI) IEMMEANS=TABLES(Week) COMPARE ADJ(BONFERRONI) IEMMEANS=TABLES (PartAge*FaceAge) IEMMEANS=TABLES (PartAge*Week) IEMMEANS=TABLES (FaceAge*Week) IEMMEANS=TABLES (PartAge*FaceAge*Week) /PRINT=DESCRIPTIVE /CRITERIA=A L P H A(.05) /WSDESIGN=Week /DESIGN=PartAge FaceAge PartAge*FaceAge.
Resources	**Processor Time**	00:00:00.029
	Elapsed Time	00:00:00.000

Within-Subjects Factors

Measure:MEASURE 1

Week	Dependent Variable
1	SET2Sess2
2	SET2Sess3

Between-Subjects Factors

		Value Label	N
PartAge	1.00	Young	36
	2.00	Old	36
FaceAge	1.00	Young	36
	2.00	Old	36

Descriptive Statistics

PartAge		FaceAge	Mean	Std. Deviation	N
SET2Sess2	Young	Young	4.8889	.28006	1 8
		Old	4.7593	.48244	1 8
		Total	4.8241	.39429	36
	Old	Young	3.6667	1.57181	1 8
		Old	4.2222	1.13759	1 8
		Total	3.9444	1.38128	36
	Total	Young	4.2778	1.27366	36
		Old	4.4907	.90321	36
		Total	4.3843	1.10151	72
SET2Sess3	Young	Young	4.8148	.26127	18
		Old	4.3889	1.18955	18
		Total	4.6019	.87585	36
	Old	Young	2.6111	1.94449	18
		Old	3.8704	1.38686	18
		Total	3.2407	1.78283	36
	Total	Young	3.7130	1.76591	36
		Old	4.1296	1.30025	36
		Total	3.9213	1.55393	72

Multivariate ate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Week	Pillai's Trace	.173	14.216[d]	1.000	68.000	.000
	Wilks' Lambda	.827	14.216[a]	1.000	68.000	.000
	Hotelling's Trace	209	14.216[a]	1.000	68.000	.000
	Roy's Largest Root	209	14.216[a]	1.000	68.000	.000
Week * PartAge	Pillai's Trace	.054	3.844[a]	1.000	68.000	.054
	Wilks' Lambda	.946	3.844[a]	1.000	68.000	.054
	Hotelling's Trace	057	3.844[a]	1.000	68.000	.054
	Roy's Largest Root	057	3.844[a]	1.000	68.000	.054
Week * FaceAge	Pillai's Trace	.010	.688[d]	1.000	68.000	.410
	Wilks' Lambda	.990	.688[a]	1.000	68.000	.410
	Hotelling's Trace	010	.688[a]	1.000	68.000	.410
	Roy's Largest Root	010	.688[a]	1.000	68.000	.410
Week * PartAge * FaceAge	Pillai's Trace	.057	4.146[d]	1.000	68.000	.046
	Wilks' Lambda	.943	4.146[a]	1.000	68.000	.046
	Hotelling's Trace	061	4.146[a]	1.000	68.000	.046
	Roy's Largest Root	061	4.146[a]	1.000	68.000	.046

a. Exactstatistic b. Design: Intercept + PartAge + FaceAge + PartAge * FaceAge Within Subjects Design: Week

Mauchly's Test of Sphericity[b]

Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Week	1.000	.000	0		1.000	1.000	1.000

Tests thenull thatthe error covariance matrix of the orthonormalized transformeddependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + PartAge + FaceAge + PartAge * FaceAge Within Subjects Design: Week

Tests of Within-Subjects Effects

Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Sphericity Assumed	7.716	1	7.716	14.216	.000
	Greenhouse-Geisser	7.716	1.000	7.716	14.216	.000
	Huynh-Feldt	7.716	1.000	7.716	14.216	.000
	Lower-bound	7.716	1.000	7.716	14.216	.000
Week * PartAge	Sphericity Assumed	2.086	1	2.086	3.844	.054
	Greenhouse-Geisser	2.086	1.000	2.086	3.844	.054
	Huynh-Feldt	2.086	1.000	2.086	3.844	.054
	Lower-bound	2.086	1.000	2.086	3.844	.054
Week * FaceAge	Sphericity Assumed	.373	1	.373	.688	.410
	Greenhouse-Geisser	.373	1.000	.373	.688	.410
	Huynh-Feldt	.373	1.000	.373	.688	.410
	Lower-bound	.373	1.000	.373	.688	.410
Week * PartAge * FaceAge	Sphericity Assumed	2.250	1	2.250	4.146	.046
	Greenhouse-Geisser	2.250	1.000	2.250	4.146	.046
	Huynh-Feldt	2.250	1.000	2.250	4.146	.046
	Lower-bound	2.250	1.000	2.250	4.146	.046
Error(Week)	Sphericity Assumed	36.907	68	.543		
	Greenhouse-Geisser	36.907	68.000	.543		
	Huynh-Feldt	36.907	68.000	.543		
	Lower-bound	36.907	68.000	.543		

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	Week	Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Linear	7.716	1	7.716	14.216	.000
Week * PartAge	Linear	2.086	1	2.086	3.844	.054
Week * FaceAge	Linear	.373	1	.373	.688	.410
Week * PartAge * FaceAge	Linear	2.250	1	2.250	4.146	.046
Error(Week)	Linear	36.907	68	.543		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	2483.361	1	2483.361	1092.477	.000
PartAge	45.188	1	45.188	19.879	.000
FaceAge	3.568	1	3.568	1.570	.215
PartAge * FaceAge	12.642	1	12.642	5.561	.021
Error	154.574	68	2.273		

stimated Marginal Means

1. Grand Mean

Measure:MEASURE 1

		95% Confidence Interval	
Mean	Std. Error	Lower Bound	Upper Bound
4.153	.126	3.902	4.403

. PartAge

Estimates

Measure:MEASURE 1

PartAge			95% Confidence Interval	
	Mean	Std. Error	Lower Bound	Upper Bound
Young	4.713	.178	4.358	5.068
Old	3.593	.178	3.238	3.947

Pairwise Comparisons

Measure:MEASURE 1

(I) PartAge	(J) PartAge	Mean Difference (I-.!t	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Young	Old	1.120	.251	.000	.619	1.622
Old	Young	-1.120	.251	.000	-1.622	-.619

Based on estimated marginal means

*. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	22.594	1	22.594	19.879	.000
Error	77.287	68	1.137		

The F tests the effect of PartAge. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

Estimates

Measure:MEASURE 1

FaceAge	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	3.995	.178	3.641	4.350
Old	4.310	.178	3.956	4.665

Pairwise Comparisons

Measure:MEASURE 1

(I) FaceAge	FaceAge	Mean Difference (I-	Std. Error	Sig.[a]	95% Confidence Interval for s Difference	
					Lower Bound	Upper Bound
Young	Old	-.315	.251	.215	-.816	.187
Old	Young	.315	.251	.215	-.187	.816

Based on estimated marginal means a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	1.784	1	1.784	1.570	.215
Error	77.287	68	1.137		

The F tests the of FaceAge. FaceAge. This test isbased on the linearly independent pairwise comparisons among the estimated marginal means.

ł. Week

Estimates

Measure:MEASURE 1

Week	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.384	.119	4.147	4.622
2	3.921	.158	3.606	4.237

Pairwise Comparisons

Measure:MEASURE 1

(I) Week	(J) Week	Mean Difference (I- .!t	Std. Error	Sig.[a]	95% Confidence In$_s$terval for Difference	
					Lower Bound	Upper Bound
1	2	.463	.123	.000	.218	.708
2	1	- .463	.123	.000	- .708	- .218

Based on estimated marginal means

*. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillai's trace	.173	14.216[d]	1.000	68.000	.000
Wilks' lambda	.827	14.216[a]	1.000	68.000	.000
Hotelling's trace	.209	14.216[a]	1.000	68.000	.000
Roy's largest root	.209	14.216[a]	1.000	68.000	.000

Each F tests the multivariate effect of Week. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. PartAge * FaceAge

Measure:MEASURE 1

PartAge	FaceAge	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	Young	4.852	.251	4.350	5.353
	Old	4.574	.251	4.073	5.076
Old	Young	3.139	.251	2.637	3.640
	Old	4.046	.251	3.545	4.548

6. PartAge * Week

Measure:MEASURE 1

PartAge	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.824	.168	4.488	5.160
	2	4.602	.223	4.156	5.048
Old	1	3.944	.168	3.609	4.280
	2	3.241	.223	2.795	3.687

7. FaceAge * Week

Measure:MEASURE 1

FaceAge	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	4.278	.168	3.942	4.613
	2	3.713	.223	3.267	4.159
Old	1	4.491	.168	4.155	4.826
	2	4.130	.223	3.684	4.575

8. PartAge * FaceAge * Week

Measure:MEASURE 1

PartAge	FaceAge	Week	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	Young	1	4.889	.238	4.414	5.364
		2	4.815	.316	4.184	5.445
Old		1	4.759	.238	4.284	5.234
		2	4.389	.316	3.758	5.019
Old	Young	1	3.667	.238	3.192	4.141
		2	2.611	.316	1.981	3.242
Old		1	4.222	.238	3.747	4.697

	2	3.870	.316	3.240	4.501

SET1

Notes

Output Created		21-Jun-2012 09:54:17
Comments		/Users/j jnicholson/Documents/St 5-Time.sav DataSetI
Input	**Data**	<none> <none>
		<none>
	Active Dataset	72
	Filter	User-defined missing values are treated as missing.
	Weight	Statistics are based on all cases with valid data for all variables in the model.
	Split File	GLM SETiSess1 SETISess2 SET1 Sess3 BY PartAge FaceAge /WSFACTOR=Week 3 Polynomial IMETHOD=SSTYPE(3) /EMMEANS=TAB LES(OVERALL) /EMMEANS=TABLES(PartA e) COMPARE ADJ(BONFERRONIJ /EMMEANS=TAB LES(FaceAge) COMPARE ADJ(BONFERRONI) IEMMEANS=TABLES(Week) COMPARE ADJ(BONFERRONI) IEMMEANS=TABLES (PartAge*FaceAge) IEMMEANS=TABLES (PartAge*Week) IEMMEANS=TABLES (FaceAge*Week) IEMMEANS=TABLES (PartAge*FaceAge*Week) /PRINT=DESCRIPTIVE /CRITERIA=A L P H A(.05) /WSDESIGN=Week /DESIGN=PartAge FaceAge PartAge*FaceAge.
	N of Rows in Working Data File	
Missing Value Handling	**Definition of Missing**	
	Cases Used	
Syntax		
Resources	**Processor Time**	00:00:00.046 00:00:00.000
	Elapsed Time	

Within-Subjects Factors

Measure:MEASURE 1

Week	Dependent Variable
1	SETi Sessi
2	SETi Sess2
3	SETi Sess3

Between-Subjects Factors

		Value Label	N
PartAge	1.00	Young	36
	2.00	Old	36
FaceAge	1.00	Young	36
	2.00	Old	36

D Descriptive Statistics

PartAge		FaceAge	Mean	Std. Deviation	N
SETISessi	Young	Young	9.4808	1.88997	18
		Old	10.8162	4.78964	18
		Total	10.1485	3.65187	36
Old		Young	16.8604	5.90808	18
		Old	16.3272	5.74912	18
		Total	16.5938	5.75162	36
Total		Young	13.1706	5.71775	36
		Old	13.5717	5.91662	36
		Total	13.3711	5.78045	72
SETISess2	Young	Young	8.9023	1.34064	18
		Old	9.4178	1.65904	18
		Total	9.1600	1.50936	36
Old		Young	17.5713	7.34871	18
		Old	19.3449	7.23524	18
		Total	18.4581	7.24332	36
Total		Young	13.2368	6.81381	36
		Old	14.3813	7.21832	36
		Total	13.8091	6.99316	72
SETISess3	Young	Young	9.0465	2.01211	18
		Old	9.5133	1.90492	18
		Total	9.2799	1.94551	36
Old		Young	19.2032	9.44568	18
		Old	19.5483	8.87877	18
		Total	19.3757	9.03640	36
Total		Young	14.1249	8.47520	36
		Old	14.5308	8.12078	36
		Total	14.3278	8.24376	72

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Week	Pillai's Trace	.029	.991[d]	2.000	67.000	.376
	Wilks' Lambda	.971	.991[a]	2.000	67.000	.376
	Hotelling's Trace	.030	.991[a]	2.000	67.000	.376
	Roy's Largest Root	.030	.991[a]	2.000	67.000	.376
Week * PartAge	Pillai's Trace	.142	5.539[a]	2.000	67.000	.006
	Wilks' Lambda	.858	5.539[a]	2.000	67.000	.006
	Hotelling's Trace	.165	5.539[a]	2.000	67.000	.006
	Roy's Largest Root	.165	5.539[a]	2.000	67.000	.006
Week * FaceAge	Pillai's Trace	.014	.484[d]	2.000	67.000	.619
	Wilks' Lambda	.986	.484[a]	2.000	67.000	.619
	Hotelling's Trace	.014	.484[a]	2.000	67.000	.619
	Roy's Largest Root	.014	.484[a]	2.000	67.000	.619
Week * PartAge * FaceAge	Pillai's Trace	.045	1.563[d]	2.000	67.000	.217
	Wilks' Lambda	.955	1.563[a]	2.000	67.000	.217
	Hotelling's Trace	.047	1.563[a]	2.000	67.000	.217
	Roy's Largest Root	.047	1.563[a]	2.000	67.000	.217

a. Exact statistic

b. Design: Intercept + PartAge + FaceAge + PartAge * FaceAge Within Subjects Design: Week

Mauchly's Test of Sphericity[b]

Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Week	.788	15.964	2	.000	.825	.880	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + PartAge + FaceAge + PartAge * FaceAge Within Subjects Design: Week

Tests of Within-Subjects Effects

Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Sphericity Assumed	33.028	2	16.514	1.378	.256
	Greenhouse-Geisser	33.028	1.650	20.015	1.378	.255
	Huynh-Feldt	33.028	1.761	18.760	1.378	.255
	Lower-bound	33.028	1.000	33.028	1.378	.245
Week * PartAge	Sphericity Assumed	132.606	2	66.303	5.531	.005
	Greenhouse-Geisser	132.606	1.650	80.360	5.531	.008
	Huynh-Feldt	132.606	1.761	75.322	5.531	.007
	Lower-bound	132.606	1.000	132.606	5.531	.022
Week * FaceAge	Sphericity Assumed	6.589	2	3.294	.275	.760
	Greenhouse-Geisser	6.589	1.650	3.993	.275	.717
	Huynh-Feldt	6.589	1.761	3.742	.275	.732
	Lower-bound	6.589	1.000	6.589	.275	.602
Week * PartAge * FaceAge	Sphericity Assumed	22.097	2	11.049	.922	.400
	Greenhouse-Geisser	22.097	1.650	13.391	.922	.385
	Huynh-Feldt	22.097	1.761	12.552	.922	.390
	Lower-bound	22.097	1.000	22.097	.922	.340
Error(Week)	Sphericity Assumed	1630.279	136	11.987		
	Greenhouse-Geisser	1630.279	112.210	14.529		
	Huynh-Feldt	1630.279	119.716	13.618		
	Lower-bound	1630.279	68.000	23.975		

Tests of Within-Subjects Contrasts

Measure:MEASURE 1

Source	Week	Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Linear	32.950	1	32.950	1.956	.166
	Quadratic	.078	1	.078	.011	.917
Week * PartAge	Linear	119.937	1	119.937	7.120	.010
	Quadratic	12.669	1	12.669	1.777	.187
Week * FaceAge	Linear	.000	1	.000	.000	.997
	Quadratic	6.588	1	6.588	.924	.340
Week * PartAge * FaceAge	Linear	6.866	1	6.866	.408	.525
	Quadratic	15.232	1	15.232	2.136	.148
Error(Week)	Linear	1145.397	68	16.844		
	Quadratic	484.882	68	7.131		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed
Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	41350.005	1	41350.005	579.935	.000
PartAge	4005.995	1	4005.995	56.184	.000
FaceAge	22.852	1	22.852	.320	.573
PartAge * FaceAge	.804	1	.804	.011	.916
Error	4848.474	68	71.301		

Estimated Marginal Means

1. Grand Mean

Measure:MEASURE 1

		95% Confidence Interval	
Mean	Std. Error	Lower Bound	Upper Bound
13.836	.575	12.690	14.982

!. PartAge

Estimates

Measure:MEASURE 1

PartAge			95% Confidence Interval	
	Mean	Std. Error	Lower Bound	Upper Bound
Young	9.529	.813	7.908	11.151
Old	18.143	.813	16.521	19.764

Pairwise rwise Comparisons

Measure:MEASURE 1

(I) PartAge	(J) PartAge	Mean Difference (I-.!t	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Young	Old	-8.613	1.149	.000	-10.906	-6.320
Old	Young	8.613	1.149	.000	6.320	10.906

Based onesti estimated marginal means *. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	1335.332	1	1335.332	56.184	.000
Error	1616.158	68	23.767		

The F tests the effect of PartAge. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

!. FaceAge

Estimates

Measure:MEASURE 1

FaceAge			95% Confidence Interval	
	Mean	Std. Error	Lower Bound	Upper Bound
Young	13.511	.813	11.889	15.132
Old	14.161	.813	12.540	15.783

Pairwise Comparisons

Measure:MEASURE 1

(I) FaceAge	(J) FaceAge	Mean Difference (I-	Std. Error	Sig.[a]	95% Confidence Interval for s Difference	
					Lower Bound	Upper Bound

Young	Old	-.651	1.149	.573	-2.943	1.642
Old	Young	.651	1.149	.573	-1.642	2.943

Based on estimated mated marginal means a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	7.617	1	7.617	.320	.573
Error	1616.158	68	23.767		

The F tests theeffectof FaceAge. This test isbased on the linearly nearly independent pairwise comparisons among the estimated marginal means.

1. Week

Estimates

Measure:MEASURE 1

Week	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	13.371	.573	12.228	14.514
2	13.809	.621	12.571	15.047
3	14.328	.781	12.769	15.887

Pairwise Comparisons

Measure:MEASURE 1

(I) Week	(J) Week	Mean Difference (I-.!t	Std. Error	a Sig.	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	-.438	.450	1.000	-1.542	.666
	3	-.957	.684	.499	-2.636	.722
2	1	.438	.450	1.000	-.666	1.542
	3	-.519	.573	1.000	-1.926	.888
3	1	.957	.684	.499	-.722	2.636
	2	.519	.573	1.000	-.888	1.926

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillars trace	.029	.991[d]	2.000	67.000	.376
Wilks' lambda	.971	.991[a]	2.000	67.000	.376
Hotelling's trace	.030	.991[a]	2.000	67.000	.376
Roy's largest root	.030	.991[a]	2.000	67.000	.376

Each F tests the multivariate effect of Week. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. PartAge * FaceAge

Measure:MEASURE 1

PartAge	FaceAge	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	Young	9.143	1.149	6.850	11.436
	Old	9.916	1.149	7.623	12.209
Old	Young	17.878	1.149	15.585	20.171
	Old	18.407	1.149	16.114	20.700

6. PartAge * Week

Measure:MEASURE 1

PartAge	Week		Std. Error	95% ConfidenceInterval$_{Mean}$	
				LowerBound	U p p e r n d
Young	1	10.148	.810	8.532	11.765
	2	9.160	.878	7.409	10.911
	3	9.280	1.105	7.076	11.484
Old	1	16.594	.810	14.978	18.210
	2	18.458	.878	16.707	20.209
	3	19.376	1.105	17.171	21.580

7.FaceAge* Week

Meure:MEASURE_1 ure:MEASURE 1

FaceAge	Week		Std. Error	95% ConfidenceInterval$_{Mean}$	
				Lower Bound	Upper Bound
Young	1	13.171	.810	11.554	14.787
	2	13.237	.878	11.486	14.988
	3	14.125	1.105	11.920	16.329
Old	1	13.572	.810	11.955	15.188
	2	14.381	.878	12.630	16.133
	3	14.531	1.105	12.326	16.735

8. PartAge * FaceAge * Week

Measure:MEASURE 1

PartAge	FaceAge	Week	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	Young	1	9.481	1.145	7.195	11.766
		2	8.902	1.241	6.426	11.379
		3	9.046	1.562	5.929	12.164
Old		1	10.816	1.145	8.530	13.102
		2	9.418	1.241	6.941	11.894
		3	9.513	1.562	6.396	12.631
Old	Young	1	16.860	1.145	14.575	19.146
		2	17.571	1.241	15.095	20.048
		3	19.203	1.562	16.086	22.321
Old		1	16.327	1.145	14.041	18.613
		2	19.345	1.241	16.868	21.821
		3	19.548	1.562	16.431	22.666

/METHOD= SSTYPE(3) /EMMEANS=TABLES(OVERALL)

```
  /EMMEANS=TABLES(PartAge) COMPARE ADJ(BONFERRONI) /EMMEANS=TABLES(FaceAge) COMPARE
  ADJ(BONFERRONI) /EMMEANS=TABLES(Week) COMPARE ADJ(BONFERRONI)
  /EMMEANS=TABLES(PartAge*FaceAge) /EMMEANS=TABLES(PartAge*Week)
  /EMMEANS=TABLES(FaceAge*Week) /EMMEANS=TABLES(PartAge*FaceAge*Week)
  /PRINT=DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /WSDESIGN=Week
  /DESIGN=PartAge FaceAge PartAge*FaceAge.
```

ET 2

Notes

Output Created		**21-Jun-2012 09:56:03**
Comments		
Input	**Data**	**/Users/jjnicholson/Documents/St 5-Time.sav DataSetI**
	Active Dataset	**<none>**
	Filter	**<none>**
	Weight	**<none>**
	Split File	
	N of Rows in Working Data File	**72**
Missing Value Handling	**Definition of Missing**	**User-defined missing values are treated as missing.**
	Cases Used	**Statistics are based on all cases with valid data for all variables in the model.**
Syntax		**GLM SET2Sess2 SET2Sess3 BY PartAge FaceAge /WSFACTOR=Week 2 Polynomial** /METHOD=SSTYPE(3) **/EMMEANS=TABLES(OVERALL) /EMMEANS=TABLES(PartA e) COMPARE ADJ(BONFERRONIl) /EMMEANS=TABLES(FaceAge) COMPARE ADJ(BONFERRONI) /EMMEANS=TABLES(Week) COMPARE ADJ(BONFERRONI) /EMMEANS=TABLES (PartAge*FaceAge) /EMMEANS=TABLES (PartAge*Week) /EMMEANS=TABLES (FaceAge*Week) /EMMEANS=TABLES (PartAge" FaceAge*Week) /PRINT=DESCRIPTIVE /CRITERIA=ALPHA(.05) /WSDESIGN=Week /DESIGN=PartAge FaceAge PartAge*FaceAge.**
Resources	**Processor Time**	**00:00:00.029 00:00:00.000**
	Elapsed Time	

[DataSeti] /Users/jjnicholson/Documents/St5-Time.sav

Within-Subjects Factors

Measure:MEASURE 1

Week	Dependent Variable
1	SET2Sess2
2	SET2Sess3

Between-Subjects Factors

		Value Label	N
PartAge	1.00	Young	36
	2.00	Old	36
FaceAge	1.00	Young	36
	2.00	Old	36

Descriptive Statistics

PartAge		FaceAge	Mean	Std. Deviation	N
SET2Sess2	Young	Young	10.6758	4.24554	1 8
		Old	11.2123	2.58247	1 8
		Total	10.9440	3.47392	36
	Old	Young	18.1900	6.68781	1 8
		Old	18.4322	7.24076	1 8
		Total	18.3111	6.87058	36
	Total	Young	14.4329	6.70807	36
		Old	14.8222	6.48913	36
		Total	14.6276	6.55580	72
SET2Sess3	Young	Young	10.4230	2.73499	1 8
		Old	11.5495	4.19588	1 8
		Total	10.9863	3.53705	36
	Old	Young	21.3692	8.60239	1 8
		Old	19.4727	7.60187	1 8
		Total	20.4210	8.05834	36
	Total	Young	15.8961	8.38970	36
		Old	15.5111	7.26378	36
		Total	15.7036	7.79391	72

Multivariate ate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Week	Pillai's Trace	.043	3.029[d]	1.000	68.000	.086
	Wilks' Lambda	.957	3.029[a]	1.000	68.000	.086
	Hotelling's Trace	045	3.029[a]	1.000	68.000	.086
	Roy's Largest Root	045	3.029[a]	1.000	68.000	.086
Week * PartAge	Pillai's Trace	.039	2.796[a]	1.000	68.000	.099
	Wilks' Lambda	.961	2.796[a]	1.000	68.000	.099
	Hotelling's Trace	041	2.796[a]	1.000	68.000	.099
	Roy's Largest Root	041	2.796[a]	1.000	68.000	.099
Week * FaceAge	Pillai's Trace	.006	.392[d]	1.000	68.000	.533
	Wilks' Lambda	.994	.392[a]	1.000	68.000	.533
	Hotelling's Trace	006	.392[a]	1.000	68.000	.533
	Roy's Largest Root	006	.392[a]	1.000	68.000	.533
Week * PartAge * FaceAge	Pillai's Trace	.018	1.217[d]	1.000	68.000	.274
	Wilks' Lambda	.982	1.217[a]	1.000	68.000	.274
	Hotelling's Trace	.018	1.217[a]	1.000	68.000	.274
	Roy's Largest Root	018	1.217[a]	1.000	68.000	.274

a. Exactstatistic b. Design: Intercept + PartAge + FaceAge + PartAge * FaceAge Within Subjects Design: Week

Mauchly's Test of Sphericity[b]

Measure:MEASURE 1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Week	1.000	.000	0		1.000	1.000	1.000

Tests thenull thatthe error covariance matrix of the orthonormalized transformeddependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + PartAge + FaceAge + PartAge * FaceAge Within Subjects Design: Week

Tests of Within-Subjects Effects

Measure:MEASURE 1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Sphericity Assumed	41.684	1	41.684	3.029	.086
	Greenhouse-Geisser	41.684	1.000	41.684	3.029	.086
	Huynh-Feldt	41.684	1.000	41.684	3.029	.086
	Lower-bound	41.684	1.000	41.684	3.029	.086
Week * PartAge	Sphericity Assumed	38.475	1	38.475	2.796	.099
	Greenhouse-Geisser	38.475	1.000	38.475	2.796	.099
	Huynh-Feldt	38.475	1.000	38.475	2.796	.099
	Lower-bound	38.475	1.000	38.475	2.796	.099
Week * FaceAge	Sphericity Assumed	5.396	1	5.396	.392	.533
	Greenhouse-Geisser	5.396	1.000	5.396	.392	.533
	Huynh-Feldt	5.396	1.000	5.396	.392	.533
	Lower-bound	5.396	1.000	5.396	.392	.533
Week * PartAge * FaceAge	Sphericity Assumed	16.753	1	16.753	1.217	.274
	Greenhouse-Geisser	16.753	1.000	16.753	1.217	.274
	Huynh-Feldt	16.753	1.000	16.753	1.217	.274
	Lower-bound	16.753	1.000	16.753	1.217	.274
Error(Week)	Sphericity Assumed	935.795	68	13.762		
	Greenhouse-Geisser	935.795	68.000	13.762		
	Huynh-Feldt	935.795	68.000	13.762		
	Lower-bound	935.795	68.000	13.762		

Tests of Within-Subjects Contrasts Measure:MEASURE 1

Source	Week	Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Linear	41.684	1	41.684	3.029	.086
Week * PartAge	Linear	38.475	1	38.475	2.796	.099
Week * FaceAge	Linear	5.396	1	5.396	.392	.533
Week * PartAge * FaceAge	Linear	16.753	1	16.753	1.217	.274
Error(Week)	Linear	935.795	68	13.762		

Tests of Between-Subjects Effects

Measure:MEASURE 1 Transformed
Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	33119.305	1	33119.305	592.269	.000
PartAge	2540.685	1	2540.685	45.435	.000
FaceAge	.000	1	.000	.000	.999
PartAge * FaceAge	24.758	1	24.758	.443	.508
Error	3802.517	68	55.919		

stimated Marginal Means

1. Grand Mean

Measure:MEASURE 1

		95% Confidence Interval	
Mean	Std. Error	Lower Bound	Upper Bound
15.166	.623	13.922	16.409

. PartAge

Estimates

Measure:MEASURE 1

PartAge			95% Confidence Interval	
	Mean	Std. Error	Lower Bound	Upper Bound
Young	10.965	.881	9.207	12.724
Old	19.366	.881	17.607	21.125

Pairwise Comparisons

Measure:MEASURE 1

(I) PartAge	(J) PartAge	Mean Difference (I-.!t	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Young	Old	-8.401	1.246	.000	-10.888	-5.914
Old	Young	8.401	1.246	.000	5.914	10.888

Based on estimated marginal means

*. The mean difference is significant at the

a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

Measure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	1270.342	1	1270.342	45.435	.000
Error	1901.258	68	27.960		

The F tests the effect of PartAge. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

Estimates

Measure:MEASURE 1

FaceAge	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Young	15.164	.881	13.406	16.923
Old	15.167	.881	13.408	16.925

Pairwise Comparisons

Measure:MEASURE 1

(I) FaceAge	FaceAge	Mean Difference (I-	Std. Error	Sig.[a]	95% Confidence Interval for s Difference	
					Lower Bound	Upper Bound
Young	Old	-.002	1.246	.999	-2.489	2.485
Old	Young	.002	1.246	.999	-2.485	2.489

Based on estimated marginal means a. Adjustment for multiple comparisons: Bonferroni.

Univariate Tests

ure:MEASURE 1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	8.682E-5	1	8.682E-5	.000	.999
Error	1901.258	68	27.960		

The F tests theeffectof FaceAge. This testisbased on the linearly nearly independent pairwise comparisons among the estimated marginal means.

ł. Week

Estimates

Measure:MEASURE 1

Week	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	14.628	.650	13.330	15.926
2	15.704	.738	14.231	17.176

Pairwise rwise Comparisons

Measure:MEASURE 1

(I) Week	(J) Week	Mean Difference (I- .!t	Std. Error	Sig.[a]	95% Confidence Interval for s Difference	
					Lower Bound	Upper Bound
1	2	-1.076	.618	.086	-2.310	.158
2	1	1.076	.618	.086	- .158	2.310

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillai's trace	.043	3.029[d]	1.000	68.000	.086
Wilks' lambda	.957	3.029[a]	1.000	68.000	.086
Hotelling's trace	.045	3.029[a]	1.000	68.000	.086
Roy's largest root	.045	3.029[a]	1.000	68.000	.086

Each F tests the multivariate effect of Week. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

5. PartAge * FaceAge

Measure:MEASURE 1

PartAge	FaceAge	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	Young	10.549	1.246	8.062	13.036
	Old	11.381	1.246	8.894	13.868
Old	Young	19.780	1.246	17.293	22.267
	Old	18.952	1.246	16.465	21.439

6. PartAge * Week

Measure:MEASURE 1

PartAge	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	10.944	.920	9.108	12.780
	2	10.986	1.044	8.904	13.069
Old	1	18.311	.920	16.476	20.147
	2	20.421	1.044	18.338	22.504

7. FaceAge * Week

Measure:MEASURE 1

FaceAge	Week	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Young	1	14.433	.920	12.597	16.268
	2	15.896	1.044	13.813	17.979
Old	1	14.822	. 920	12.987	16. 658
	2	15.511	1.044	13.428	17.594

8. PartAge * FaceAge * Week Measure:MEASURE

1

PartAge	FaceAge	Week	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Young	Young	1	10.676	1.301	8.080	13.272
		2	10.423	1.476	7.478	13.368
Old		1	11.212	1.301	8.616	13.808
		2	11.550	1.476	8.604	14.495
Old	Young	1	18.190	1.301	15.594	20.786
		2	21.369	1.476	18.424	24.315
Old		1	18.432	1.301	15.836	21.028
		2	19.473	1.476	16.527	22.418

Appendix D – Published Papers

D1:
Nicholson, J., Dunphy, P., Coventry, L., Briggs, P., & Olivier, P. L. (2012). A security assessment of Tiles: a new portfolio-based graphical authentication system. *CHI 2012 Works-in-Progress*. Texas, USA.

REFERENCES

Akatsu, H., & Miki, H. (2004). Usability Research for the elderly people. *Oki Technical Review*, *71*(3), 54–57.

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*(6), 1043–1047.

Arbor, A. (2001). Memory Starts To Decline In Our Mid-20s. *Science Daily*. Retrieved November 25, 2010, from http://www.sciencedaily.com/releases/2001/08/010814063231.htm

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

Bartlett, J., & Fulton, A. (1991). Familiarity and recognition of faces in old age. *Memory & Cognition*, *19*(3), 229–238.

Bastin, C., & Van der Linden, M. (2006). The effects of aging on the recognition of different types of associations. *Experimental Aging Research*, *32*(1), 61–77.

Bonneau, J., Preibusch, S., & Anderson, R. (2012). A birthday present every eleven wallets? The security of customer-chosen banking PINs. *In Proceedings of Financial Cryptography 2012* (pp. 1–15).

Borgida, E., & Nisbett, R. E. (1977). The Differential Impact of Abstract vs. Concrete Information on Decisions. *Journal of Applied Social Psychology*, *7*(3), 258–271.

Bornstein, M. H. (1987). Perceptual categories in vision and audition. *S. Harnad (Ed.), Categorical perception: The groundwork of cognition* (pp. 287–300).

Brigham, J. C., & Williamson, N. L. (1979). Cross-racial Recognition and Age: When You're Over 60, Do They Still "All Look Alike?" *Personality and Social Psychology Bulletin*, *5*(2), 218–222.

Brostoff, S, & Sasse, M. (2000). Are Passfaces more usable than passwords? A field trial investigation. *In Proceedings of Human Computer Interaction* (pp. 405–424).

Brostoff, S., Inglesant, P., & Sasse, M. A. (2010). Evaluating the usability and security of a graphical one-time PIN system. *In Proceedings of HCI 2010*.

Brostoff, S., & Sasse, M. A. (2003). "Ten strikes and you're out": Increasing the number of login attempts can improve password usability. *In Proceedings of CHI 2003* (pp. 1–4).

Brown, S., & Park, D. (2003). Theoretical Models of Cognitive Aging and Implications for Translational Research in Medicine. *The Gerontologist*, *43*(1), 57–67.

Bruce, V. (1982). Changing faces: visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*(Pt 1), 105–116.

Bruce, V, Burton, M., & Dench, N. (1994). What's distinctive about a distinctive face? *The Quarterly Journal of Experimental Psychology Section A*, *47*(1), 119–141.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207–218.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327.

Burton, A., & Wilson, S. (1999). Face Recognition in Poor-Quality Video: Evidence From Security Surveillance. *Psychological Science*, *10*(3), 243–248.

Bäckman, L. (1991). Recognition memory across the adult life span: the role of prior knowledge. *Memory & Cognition*, *19*(1), 63–71.

Cellerino, A., Borghetti, D., & Sartucci, F. (2004). Sex differences in face gender recognition in humans. *Brain research bulletin*, *63Lnaveh*(6), 443–449.

Chalfonte, B. L., & Johnson, M. K. (1996). Feature memory and binding in young and older adults. *Memory & cognition*, *24*(4), 403–16.

Chiasson, S., Forget, A., Biddle, R., & Van Oorschot, P. C. (2008). Influencing Users Towards Better Passwords: Persuasive Cued Click-Points. *In Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1* (pp. 121–130).

Chiasson, S., Forget, A., Stobert, E., Van Oorschot, P. C., & Biddle, R. (2009). Multiple Password Interference in Text and Click-Based Graphical Passwords. *In Proceedings of the 16th ACM conference on Computer and communications security* (pp. 500–511).

Chiasson, S., Van Oorschot, P. C., & Biddle, R. (2007). Graphical password authentication using cued click points. *Lecture Notes in Computer Science*, *4734/2007*, 359–374.

Craik, F., & McDowd, J. (1987). Age Differences in Recall and Recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 474–479.

Davis, D., Monrose, F., & Reiter, M. (2004). On user choice in graphical password schemes. *In Proceedings of the 13th conference on USENIX Security Symposium-Volume 13* (p. 11). USENIX Association Berkeley, CA, USA.

DeAngeli, A., Coutts, M., Coventry, L., Johnson, G., Cameron, D., & Fischer, M. (2002). VIP: a visual approach to user authentication. *In Proceedings of the Working Conference on Advanced Visual Interfaces* (pp. 316–323).

Derwinger, A., Stigsdotter Neely, A., MacDonald, S., & Bäckman, L. (2005). Forgetting numbers in old age: strategy and learning speed matter. *Gerontology*, *51*(4), 277–84.

Dhamija, R., & Perrig, A. (2000). Deja vu: A user study using images for authentication. *In Proceedings of the 9th USENIX Security Symposium* (pp. 45–48).

Dirik, A. E., Memon, N., & Birget, J.-C. (2007). Modeling user choice in the PassPoints graphical password scheme. *In Proceedings of the 3rd symposium on Usable privacy and security* (pp. 20–28).

Dobbs, A., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, *4*(4), 500–503.

Dourish, P., Grinter, R. E., Delgado de la Flor, J., & Joseph, M. (2004). Security in the wild: User strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, *8*(6), 391–401.

Dunphy, P, & Yan, J. (2007). Do background images improve “draw a secret” graphical passwords? *In Proceedings of the 14th ACM conference on Computer and communications security* (pp. 36–47).

Dunphy, P., Nicholson, J., & Olivier, P. L. (2008). Securing Passfaces for Description. *In Proceedings of 4th symposium on Usable privacy and security* (pp. 24–35).

Dunphy, P., & Olivier, P. L. (2012). On Automated Image Choice for Secure and Usable Graphical Passwords. *In Proceedings of the 28th Annual Computer Security Applications Conference (ACSAC ’12)*.

Ebner, N. C. (2008). Age of face matters: age-group differences in ratings of young and old faces. *Behavior research methods*, *40*(1), 130–136.

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES--a database of facial expressions in young, middle-aged, and older women and men: development and validation. *Behavior Research Methods*, *42*(1), 351–62.

Everitt, K. M., Bragin, T., Fogarty, J., & Kohno, T. (2009). A comprehensive study of frequency, interference, and training of multiple graphical passwords. *In Proceedings of the 27th international conference on Human factors in computing systems* (pp. 889–898). ACM New York, NY, USA.

Fisk, A.D., & Rogers, W. A. (1991). Toward an understanding of age-related memory and visual search effects. *Journal of Experimental Psychology: Generalsychology. General*, *120*(2), 131–149.

Fisk, Arthur D., Rogers, W. A., Charness, N., & Czaja, S. J. (2004). *Designing for Older Adults: Principles and Creative Human Factors Approaches*. Florida, USA: CRC Press.

Florencio, D., & Herley, C. (2007). A large-scale study of web password habits. *In Proceedings of the 16th international conference on World Wide Web - WWW ’07* (pp. 657–666). New York, New York, USA: ACM Press.

Fulton, A., & Bartlett, J. C. (1991). Young and old faces in young and old heads: the factor of age in face recognition. *Psychology and aging*, *6*(4), 623–30.

Gaw, S., & Felten, E. W. (2006). Password management strategies for online accounts. *In Proceedings of the second symposium on Usable privacy and security - SOUPS '06*, 44.

Gazzaley, A., Clapp, W., Kelley, J., McEvoy, K., Knight, R. T., & D'Esposito, M. (2008). Age-related top-down suppression deficit in the early stages of cortical visual memory processing. *In Proceedings of the National Academy of Sciences of the United States of America*, *105*(35), 13122–13126.

Goodin, D. (2012a). Why passwords have never been weaker—and crackers have never been stronger. *Ars Technica*. Retrieved November 10, 2012, from http://arstechnica.com/security/2012/08/passwords-under-assault/

Goodin, D. (2012b). Hackers expose 453,000 credentials allegedly taken from Yahoo service. *Ars Technica*. Retrieved from http://arstechnica.com/security/2012/07/yahoo-service-hacked/

Goodin, D. (2012c). 8 million leaked passwords connected to LinkedIn, dating website. *Ars Technica*. Retrieved from http://arstechnica.com/security/2012/06/8-million-leaked-passwords-connected-to-linkedin/

Goodin, D. (2012d). 11 million passwords from hacked game website dumped online. *Ars Technica*. Retrieved from http://arstechnica.com/security/2012/07/passwords-from-hacked-game-site-dumped-online/

Grady, C. L., & Craik, F. (2000). Changes in memory processing with age. *Current Opinion in Neurobiology*, *10*(2), 224–31.

Grady, C., McIntosh, A., Horwitz, B., & Maisog, J. (1995). Age-related reductions in human recognition memory due to impaired encoding. *Science*, *269*(5221), 218.

Grawemeyer, B., & Johnson, H. (2011). Using and managing multiple passwords: A week to a view. *Interacting with Computers*, *23*(3), 256–267.

Hancock, P. J. ., Bruce, V., & Burton, M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. *Categorical perception: The groundwork of cognition* (pp. 1–52). Cambridge University Press.

Harrison, V., & Hole, G. J. (2009). Evidence for a contact-based explanation of the own-age bias in face recognition. *Psychonomic Bulletin & Review*, *16*(2), 264–269.

Hart, T., Chaparro, B., & Halcomb, C. (2008). Evaluating websites for older adults: adherence to "senior-friendly" guidelines and end-user performance. *Behaviour & Information Technology*, *27*(3), 191–199.

Hasher, L., & Zacks, R. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of Learning and Motivation*, *22*, 193–225.

Hayashi, E., & Hong, J. (2011). A diary study of password usage in daily life. *In Proceedings of CHI 2011* (pp. 2627–2630).

Herley, C., Oorschot, P. C. van, & Patrick, A. (2009). Passwords: If We're So Smart, Why Are We Still Using Them? In R. Dingledine & P. Golle (Eds.), *Financial Cryptography and Data* (pp. 230–237). Springer-Verlag Berlin, Heidelberg.

Hockley, W. E. (2008). The picture superiority effect in associative recognition. *Memory & Cognition*, *36*(7), 1351–1559.

Holt, L. (2011). Increasing real-world security of user IDs and passwords. *Proceedings of the 2011 Information Security Curriculum Development Conference on - InfoSecCD '11* (pp. 34–41). New York, New York, USA: ACM Press.

Inglesant, P., & Sasse, M. A. (2010). The True Cost of Unusable Password Policies : Password Use in the Wild. *In Proceedings of the 28th international conference on Human factors in computing systems* (pp. 383–392).

Intraub, H., & Nicklos, S. (1985). Levels of processing and picture memory: The physical superiority effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 284–298.

Ives, B., Walsh, K. R., & Schneider, H. (2004). The domino effect of password reuse. *Communications of the ACM*, *47*(4), 75–78.

Jenkins, R., White, D., Van Montfort, X., & Mike Burton, a. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–23.

Jermyn, I., Mayer, A., & Monrose, F. (1999). The design and analysis of graphical passwords. *In Proceedings of the 8th USENIX Security Symposium*.

Just, M. (2004). Designing and evaluating challenge-question systems. *IEEE Security & Privacy Magazine*, *2*(5), 32–39.

Just, M. (2005). Designing authentication systems with challenge questions. *Security and Usability: Designing Secure Systems* (pp. 147–160).

Just, M., & Aspinall, D. (2009). Personal choice and challenge questions: a security and usability assessment. *In Proceedings of SOUPS 2009*.

Kausler, D. H., Salthouse, T. A, & Saults, J. S. (1988). Temporal memory over the adult lifespan. *The American Journal of Psychology*, *101*(2), 207–215.

Kay, H. (1951). Learning of a serial task by different age groups. *Quarterly Journal of Experimental Psychology*, *3*(4), 166–183.

Keith, M., Shao, B., & Steinbart, P. (2009). A Behavioral Analysis of Passphrase Design and Effectiveness. *Journal of the Association for Information Systems*, *10*(2), 63–89.

Keith, M., Shao, B., & Steinbart, P. J. (2007). The usability of passphrases for authentication: An empirical field study. *International Journal of Human-Computer Studies*, *65*(1), 17–28.

Kinisch, W. (1970). Models of Free Recall and Recognition. In D. A. Norman (Ed.), *Models of Human Memory*. New York: Academic Press.

Komanduri, S., & Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., Cranor, L. F., et al. (2011). Of passwords and people: Measuring the effect of password-composition policies. *In Proceedings of CHI 2011* (pp. 2595–2604).

Koutstaal, W., & Schacter, D. (1997). Gist-Based False Recognition of Pictures in Older and Younger Adults. *Journal of Memory and Language*, *37*(4), 555–583.

Kuo, C., Romanosky, S., & Cranor, L. F. (2006). Human Selection of Mnemonic Phrase-based Passwords Human Selection of Mnemonic Phrase-based Passwords. *In Proceedings of SOUPS 2006*.

Lamont, A., Stewart-Williams, S., & Podd, J. (2005). Face Recognition and Aging: Effects of Target Age and Memory Load. *Memory & Cognition*, *33*(6), 1017–1024.

Levin, D. T. (2000). Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology. General*, *129*(4), 559–574.

Lovén, J., Herlitz, A., & Rehnman, J. (2011). Women's own-gender bias in face recognition memory. *Experimental Psychology*, *58*(4), 333–340.

Macdonald, S. W. S., Stigsdotter-Neely, A., Derwinger, A., & Bäckman, L. (2006). Rate of acquisition, adult age, and basic cognitive abilities predict forgetting: new views on a classic problem. *Journal of Experimental Psychology: General*, *135*(3), 368–390.

Madden, D. J. (1983). Aging and distraction by highly familiar stimuli during visual search. *Developmental Psychology*, *19*(4), 499–507.

Maylor, E., Vousden, J. I., & Brown, G. D. (1999). Adult age differences in short-term memory for serial order: data and a model. *Psychology and aging*, *14*(4), 572–94.

Meissner, C. a., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35.

Merriam, S. B., & Cunningham, P. M. (1989). *Handbook of adult and continuing education*. *Jossey-Bass*.

Michel, C., Rossion, B., Han, J., Chung, C.-S., & Caldara, R. (2006). Holistic processing is finely tuned for faces of one's own race. *Psychological Science*, *17*(7), 608–15.

Mitchell, D. (1986). Semantic activation and episodic memory: Age similarities and differences. *Developmental Psychology*, *22*, 86–94.

Mitchell, K. J., Johnson, M. K., Raye, C. L., & D'Esposito, M. (2000). fMRI evidence of age-related hippocampal dysfunction in feature binding in working memory. *Cognitive Brain Research*, *10*(1-2), 197–206.

Moncur, W., & LePlâtre, G. (2007). Pictures at the ATM - Exploring the usability of multiple graphical passwords. *In Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 887–894).

Morris, R., & Thompson, K. (1979). Password security: A case history. *Communications of the ACM*, *22*(11), 594–597.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1170–1187.

Naveh-Benjamin, M., Brav, T. K., & Levy, O. (2007). The associative memory deficit of older adults: the role of strategy utilization. *Psychology and Aging*, *22*(1), 202–208.

Naveh-Benjamin, M., Hussain, Z., Guez, J., & Bar-On, M. (2003). Adult age differences in episodic memory: further support for an associative-deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 826–837.

Nelson, D., Reed, V., & Walling, J. (1976). Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 523–528.

Ng, K., Hon, K., & Lee, T. (2007). Ageing effect on face recognition. *Asian Journal of Gerontology & Geriatrics*, *2*(2), 93–98.

Nicholson, J., Dunphy, P., Coventry, L., Briggs, P., & Olivier, P. L. (2012). A security assessment of Tiles: a new portfolio-based graphical authentication system. *CHI 2012 Works-in-Progress*. Texas, USA.

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *October*, *19*(2), 155–160.

Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: a meta-analysis. *Psychology and Aging*, *23*(1), 104–118.

Park, D. C., Puglisi, J. T., & Sovacool, M. (1983). Memory for pictures, words, and spatial location in older adults: evidence for pictorial superiority. *Journal of gerontology*, *38*(5), 582–588.

Park, D. C., Smith, a D., Morrell, R. W., Puglisi, J. T., & Dudley, W. N. (1990). Effects of contextual integration on recall of pictures by older adults. *Journal of Gerontology*, *45*(2), 52–57.

Park, D. C., Puglisi, J., & Smith, A. (1986). Memory for pictures: Does an age-related decline exist? *Journal of Psychology and Aging*, *1*(1), 11–17.

Park, D. C., Royal, D., Dudley, W., & Morrell, R. (1988). Forgetting of Pictures Over a Long Retention Interval. *Psychology and Aging*, *3*(1), 94–95.

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

Pike, G., Kemp, R., & Brace, N. (2000). The psychology of human face recognition. *IEE Colloquium on Visual Biometrics* (pp. 11–17).

Proctor, R. W., Lien, M.-C., Vu, K.-P. L., Schultz, E. E., & Salvendy, G. (2002). Improving computer security for authentication of users: influence of proactive password restrictions. *Behavior Research Methods, Instruments, & Computers*, *34*(2), 163–169.

Raaijmakers, J. G. W., & Schiffrin, R. M. (1992). Models for recall and recognition. *Annual Review of Psychology*, *43*(1), 205–234.

Ragan, S. (2012). Formspring Hacked - 420,000 Passwords Leaked. *Security Week*. Retrieved from http://www.securityweek.com/formspring-hacked-420000-passwords-leaked

Rasmussen, M., & Rudmin, F. W. (2010). The coming PIN code epidemic : A survey study of memory of numeric security codes. *Electronic Journal of Applied Psychology*, *6*(2), 5–9.

Renaud, K. (2005). A visuo-biometric authentication mechanism for older users. *In Proceedings of HCI 2005* (pp. 167–182).

Renaud, K., & De Angeli, A. (2004). My password is here! An investigation into visuo-spatial authentication mechanisms. *Interacting with Computers*, *16*(6), 1017–1041.

Renaud, K., & Ramsay, J. (2007). Now what was that password again? A more flexible way of identifying and authenticating our seniors. *Behaviour & Information Technology - Designing Computer Systems for and with Older Users*, *26*(4), 309–322.

Renaud, K. (2009). Guidelines for designing graphical authentication mechanism interfaces. *International Journal of Information and Computer Security*, *3*(1), 60–85.

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: a meta-analytic and theoretical review. *Psychological Bulletin*, *138*(1), 146–174.

Rose, B. (n.d.). *Science Behind Passfaces*. Retrieved from http://www.realuser.com/enterprise/about/about_passfaces.htm

Salthouse, T. A. (1991). Mediation of Adult Age Differences in Cognition By Reductions in Working Memory and Speed of Processing. *Psychological Science*, *2*(3), 179–183.

Sarno, J. A., & Alley, T. R. (1997). Attractiveness and the memorability of faces : Only a matter of distinctiveness? *The American Journal of Psychology*, *110*(1), 81–92.

Sasse, M. A., Brostoff, S., & Weirich, D. (2001). Transforming the “weakest link” — a human/computer interaction approach to usable and effective security. *BT Technology Journal*, *19*(3), 122–131.

Sayago, S., & Blat, J. (2009). About the relevance of accessibility barriers in the everyday interactions of older people with the web. *In Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibililty* (pp. 104–114).

Schonfield, D., & Robertson, B. a. (1966). Memory storage and aging. *Canadian Journal of Psychology*, *20*(2), 228–236.

Searcy, J. H., Bartlett, J. C., & Memon, N. (1999). Age differences in accuracy and choosing in eyewitness identification and face recognition. *Memory & Cognition*, *27*(3), 538–552.

Shepard, R. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 156–163.

Shepherd, J. W., & Ellis, H. D. (1973). The effect of attractiveness on recognition memory for faces. *The American Journal of Psychology*, *86*(3), 627–633.

Sjöberg, L., & Fromm, J. (2001). Information Technology Risks as Seen by the Public. *Risk Analysis*, *21*(3), 427–442.

Slamecka, J. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604.

Smith, A., Park, D. C., Cherry, K., & Berkovsky, K. (1990). Age differences in memory for concrete and abstract pictures. *Journal of gerontology*, *45*(5), 205–209.

Smith, A., & Winograd, E. (1978). Adult age differences in remembering faces. *Developmental Psychology*, *14*(4), 443–444.

Sporer, S. L. (1991). Deep--deeper--deepest? Encoding strategies and the recognition of human faces. *Journal of experimental psychology. Learning, memory, and cognition*, *17*(2), 323–33.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, *19*(2), 73–74.

Stanton, J., Stam, K., Mastrangelo, P., & Jolton, J. (2005). Analysis of end user security behaviors. *Computers & Security*, *24*(2), 124–133.

Suo, X., Zhu, Y., & Owen, G. (2005). Graphical passwords: A survey. *In Proceedings of the 21st Annual Computer Security Applications Conference* (pp. 463–472).

Tullis, T. S., Tedesco, D. P., & McCaffrey, K. E. (2011). Can users remember their pictorial passwords six years later. *In Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (pp. 1789–1794). ACM.

Tulving, E., & Thomson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, *80*(5), 352–373.

Tulving, E., & Watkins, M. (1973). Continuity Between Recall and Recognition. *The American Journal of Psychology*, *86*(4), 739–748.

Valentine, T. (1998). *An evaluation of the Passface personal authentication system*. *Technical Report, Goldsmith College University*.

Valentine, T. (1999). *Memory for Passfaces after a long delay*. *Technical Report, Goldsmith College University*.

Valentine, T., & Bruce, V. (1986). The effect of race, inversion and encoding activity upon face recognition. *Acta Psychologica*, *61*(3), 259–273.

Verhaeghen, P., Marcoen, A., & Goossens, L. (1993). Facts and fiction about memory aging: A quantitative integration of research findings. *Journal of Gerontology*, *48*(4), 157–171.

Vines, J., Blythe, M., Dunphy, P., & Monk, A. (2011). Eighty Something : Banking for the older old. *In Proceedings of BCS HCI Conference 2011*.

Vu, K., Proctor, R., Bhargavspantzel, A., Tai, B., Cook, J., & Eugeneschultz, E. (2007). Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies*, *65*(8), 744–757.

Walker, N., Philbin, D. a, & Fisk, a D. (1997). Age-related differences in movement control: adjusting submovement structure to optimize performance. *The Journals of Gerontology*, *52*(1), P40–52. R

Walker, P. M., & Tanaka, J. W. (2003). An encoding advantage for own-race versus other-race faces. *Perception*, *32*(9), 1117–1126.

Watkins, M. (1979). An appreciation of generate-recognize theory of recall. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 687–704.

Weir, C., Douglas, G., Carruthers, M., & Jack, M. (2009). User perceptions of security, convenience and usability for ebanking authentication tokens. *Computers & Security*, *28*(1-2), 47–62.

Weir, M., Aggarwal, S., Collins, M., & Stern, H. (2010). Testing metrics for password creation policies by attacking large sets of revealed passwords. *In Proceedings of CCS 2010* (pp. 162–175).

Weiss, R., & De Luca, A. (2008). PassShapes: Utilizing stroke based authentication to increase password memorability. *In Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges* (pp. 383–392). ACM.

Weldon, M. (1987). Altering retrieval demands reverses the picture superiority effect. *Memory & Cognition*, *15*(4), 269–280.

West, R. (2008). The psychology of security. *Communications of the ACM*, 34–40.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*(1), 48–64.

Wickham, L. H. ., & Morris, P. E. (2003). Attractiveness, distinctiveness, and recognition of faces: attractive faces can be typical or distinctive but are not better recognised. *American Journal of Psychology*, *116*(3), 455–468.

Wiedenbeck, S., Waters, J., Sobrado, L., & Birget, J. (2006). Design and evaluation of a shoulder-surfing resistant graphical password scheme. *In Proceedings of the Working Conference on Advanced Visual Interfaces* (pp. 177–184).

Wiedenbeck, S., & Waters, J. (2005). Authentication Using Graphical Passwords: Effects of Tolerance and Image Choice. *Proceedings of SOUPS 2005*.

Wiedenbeck, S., Waters, J., Birget, J.-C., Brodskiy, A., & Memon, N. (2005). PassPoints: Design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies*, *63*(1-2), 102–127.

Wiese, H., Schweinberger, S. R., & Hansen, K. (2008). The age of the beholder: ERP evidence of an own-age bias in face memory. *Neuropsychologia*, *46*(12), 2973–85.

Winograd, E., Smith, A., & Simon, E. (1982). Aging and the Picture Superiority Effect in Recall. *Journal of Gerontology*, *37*(1), 70–75.

Yan, J., Blackwell, A., Anderson, R., & Grant, A. (2004). Password memorability and security: Empirical results. *Security & Privacy, IEEE*, *2*(5), 25–31.

van Oorschot, P. C., Salehi-Abari, A., & Thorpe, J. (2010). Purely Automated Attacks on PassPoints-Style Graphical Passwords. *IEEE Transactions on Information Forensics and Security*, *5*(3), 393–405.