Usability and Security of Human-Interactive Security Protocols

Ronald Kainda
St. Cross College
University of Oxford

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Trinity Term 2011

Abstract

We investigate the security and usability of Human-Interactive Security Protocols (HISPs); specifically, how digests of 4 or more digits can be compared between two or more systems as conveniently as possible while ensuring that issues such as user complacency do not compromise security. We address the research question: *given different association scenarios and modes of authentication in HISPs, how can we improve on existing, or design new, empirical channels that suit human and contextual needs to achieve acceptable effective security?* We review the literature of HISPs, proposed empirical channels, and usability studies of HISPs; we follow by presenting the methodology of the research reported in this thesis. We then make a number of contributions discussing the effectiveness of empirical channels and address the design, analysis, and evaluation of these channels. In Chapter 4 we present a user study of pairwise device associations and discuss the factors affecting effective security of empirical channels in single-user scenarios. In Chapter 5 we present a user study of group device associations and discuss the factors affecting effective security of empirical channels in multi-user scenarios. In Chapter 6 we present a set of principles for designing secure and usable empirical channels. We demonstrate the effectiveness of these principles by proposing new empirical channels. In Chapter 7 we present a framework designed for researchers and system designers to reason about empirical channels in HISPs. The framework is grounded in experimental data, related research, and validated by experts. Finally, in Chapter 8 we present a methodology for analysing and evaluating the security and usability of HISPs. We validate the methodology by applying it in laboratory experiments of HISPs.

Mum, you will always be my hero

Acknowledgements

In no chronological order, I would like to take this opportunity to express my sincere gratitude to my supervisors Professor Andrew William Roscoe and Dr. Ivan Flechaìs for their support and guidance through out my DPhil journey. Many thanks to Professor A.W. Roscoe for securing funding for my final year of my research. I would also like to thank Research In Motion for the financial support without which my studies would have not been completed on time.

I would also like to thank Dr. Andrew Martin (for being my transfer, confirmation, and internal examiner), Professor Dusko Pavlovic (transfer examiner), Dr. Marina Jirotika (confirmation examiner), and Professor Karen Renaud (external examiner) for their feedback that helped improve my work greatly.

Thanks also to my office mate (only other student member of the 'Centre for Usable Security') Shamal Faily for all the discussions on usable security and Bangdao Chen and Long Nguyen for insightful discussions on the subject of my thesis. I would also like to acknowledge members of the security reading group for the enlightening discussions on various subjects that helped me broaden my knowledge of security and related subjects.

I would also like to thank the Rhodes Trust for awarding me a life time opportunity of a scholarship to study in Oxford. Many thanks also to the staff at Rhodes House who made my life easier in Oxford, specifically Mary Eaton, Catherine King, Sheila Partridge, Bob Wyllie, and the former Warden Sir Colin Lucas.

I would like to thank a whole list of special people: my siblings whose love is ever enduring, Ola for being so supportive, supplying an unlimited amounts of coffee and tea, and taking the time to proof read all the chapters - thank you. The funniest guy I have ever met, Nkisu Kainda - you always made me smile and gave me a reason to live another day. Precious, for having the love and patience to take care of Master Nkisu. Danny Leza, you have always been there for me. Thomas Simbule, for helping me focus on my studies by taking care of my other responsibilities.

Contents

Abstract i

Dedication ii

Acknowledgements iii

Contents iv

List of figures ix

List of tables x

1 Introduction 1
- 1.1 Human-Computer Interactions and Security . . . 1
- 1.2 Human-Interactive Security Protocols (HISPs) . . . 3
- 1.3 The need for Security and Usability in HISPs . . . 4
- 1.4 Security of HISPs: a socio-technical approach . . . 6
  - 1.4.1 Technical security of HISPs . . . 6
  - 1.4.2 Effective security of HISPs . . . 6
  - 1.4.3 HISPs as socio-technical systems . . . 7
  - 1.4.4 Security and usability evaluation of HISPs . . . 8
- 1.5 Research problem . . . 9
- 1.6 Research contributions . . . 10
- 1.7 Thesis overview . . . 11

2 Background and literature review 13
- 2.1 Definitions . . . 13
- 2.2 Human-Interactive Security Protocols . . . 15
  - 2.2.1 Symmetrised Hash Commitment Before Knowledge (SHCBK) protocol . . . 15
- 2.3 Empirical channels . . . 17
  - 2.3.1 Manual comparison (MC) . . . 17
  - 2.3.2 Manual copying and entering (MCE) . . . 18
  - 2.3.3 Auxiliary device methods (ADM) . . . 18
  - 2.3.4 Short-range directed channels (SDC) . . . 19
  - 2.3.5 Timing methods (TM) . . . 20
  - 2.3.6 Limitations of current empirical channels . . . 20
  - 2.3.7 Summary . . . 22
- 2.4 User studies . . . 22
  - 2.4.1 Usability Analysis of Secure Pairing Methods . . . 23
  - 2.4.2 Analysis of Bluetooth Simple Pairing and Wi-Fi Protected Setup . . . 23
  - 2.4.3 Human Factors in HCBK protocols . . . 24
  - 2.4.4 A Comparative Usability Study of Secure Device Pairing Methods . . . 24
  - 2.4.5 Other studies . . . 24

2.4.6 Summary . . . 25
2.5 Conclusion . . . 25

3 Methodology **27**
3.1 Introduction . . . 27
3.2 Laboratory experiments . . . 28
3.2.1 Reliability . . . 28
3.2.2 Internal validity . . . 29
3.2.3 External validity: context . . . 29
3.2.4 Statistical analysis . . . 30
3.2.5 Sample sizes . . . 30
3.3 Data collection methods . . . 30
3.3.1 Questionnaires . . . 30
3.3.2 Interviews . . . 31
3.3.3 Observation . . . 31
3.4 Research approach . . . 32
3.4.1 Laboratory experiments . . . 32
3.5 Validity of research . . . 33
3.6 Conclusion . . . 34

4 HISPs: Security and Usability in Single-User Scenarios **35**
4.1 Introduction . . . 35
4.2 Device association in single-user scenarios . . . 35
4.2.1 Association scenarios . . . 35
4.2.2 Definitions . . . 37
4.3 Empirical evaluation of single-user association scenarios . . . 38
4.3.1 Participants . . . 38
4.3.2 Material and apparatus . . . 38
4.3.3 Methods tested . . . 40
4.3.4 Participant tasks . . . 45
4.3.5 Hypotheses tested . . . 45
4.4 Results . . . 46
4.4.1 Objective results . . . 46
4.4.2 Subjective results . . . 50
4.4.3 Hypotheses validation . . . 52
4.5 Effectiveness of empirical channels in single-user scenarios . . . 53
4.6 Factors affecting security and usability of empirical channels in single-user scenarios 57
4.6.1 User conditioning . . . 57
4.6.2 User motivation . . . 58
4.6.3 Attentiveness . . . 58
4.6.4 Device affordances . . . 59
4.6.5 Social contexts . . . 60
4.6.6 Personal variables . . . 60
4.7 Summary and conclusion . . . 61

5 HISPs: Security and Usability in Group Scenarios **63**
5.1 Introduction . . . 63
5.2 Device association in groups . . . 63
5.3 Security and Usability challenges of group scenarios . . . 66
5.4 Empirical evaluation of group association scenarios . . . 69
5.4.1 Participants . . . 70
5.4.2 Materials and apparatus . . . 70
5.4.3 Methods tested . . . 72
5.4.4 Participant tasks . . . 75

5.4.5 Hypotheses tested . . . 76
5.5 Results . . . 76
5.5.1 Analysis by age . . . 77
5.5.2 Analysis by method . . . 78
5.5.3 Hypotheses validation . . . 83
5.6 Discussion . . . 84
5.7 Factors affecting security and usability of empirical channels in group scenarios . . . 85
5.7.1 Trial and error . . . 85
5.7.2 Context . . . 86
5.7.3 Sum of efforts . . . 88
5.7.4 User conformity . . . 88
5.7.5 Interpretation of security . . . 89
5.7.6 Perception of security . . . 89
5.8 Summary and conclusion . . . 90

6 Principles for designing empirical channels 92
6.1 Introduction . . . 92
6.2 Principles for designing empirical channels . . . 93
6.2.1 Principle of commitment . . . 93
6.2.2 Principle of unpredictability . . . 94
6.2.3 Principle of single interaction path . . . 95
6.2.4 Principle of design by context . . . 97
6.3 Demonstration of design principles . . . 98
6.3.1 Example 1: Word-matching and number-typing . . . 98
6.3.2 Example 2: Repeated numeric comparison . . . 100
6.4 Usability evaluation of the empirical channels . . . 101
6.4.1 Experimental design . . . 101
6.4.2 Results . . . 102
6.4.3 Comparison with earlier methods . . . 103
6.4.4 Applications scenarios of proposed methods . . . 104
6.5 Summary and conclusion . . . 104

7 HISPs framework for reasoning about empirical channels 106
7.1 Introduction . . . 106
7.2 HISPs framework . . . 108
7.2.1 Technical and contextual factors . . . 109
7.2.2 Human factors . . . 114
7.2.3 Empirical channels . . . 116
7.3 Application of HISPs framework . . . 117
7.4 Application of HISPs framework: example . . . 118
7.4.1 Context . . . 119
7.4.2 Technical security . . . 120
7.4.3 Human factors . . . 120
7.4.4 Empirical channel . . . 121
7.4.5 Recommendation of empirical channel . . . 122
7.4.6 Discussion . . . 122
7.5 Expert validation of HISPs framework . . . 123
7.5.1 Benefits . . . 124
7.5.2 Criticisms . . . 124
7.6 Summary and conclusion . . . 128

8 HISPs: Analysis and Evaluation of Security and Usability 129
8.1 Introduction . . . 129
8.2 Current practice in security and usability evaluation . . . 130
8.3 Security-usability threat model . . . 132
8.3.1 Usability . . . 133
8.3.2 Security . . . 134
8.3.3 Measurable metrics . . . 136
8.4 Security-usability evaluation of secure systems . . . 138
8.4.1 Identify usage scenarios . . . 139
8.4.2 Identify threat scenarios . . . 139
8.4.3 Assess difficulty-of-use of usage scenarios . . . 140
8.4.4 Assess ease-of-use of threat scenarios . . . 141
8.4.5 Make recommendations . . . 142
8.5 Application and Validation . . . 142
8.5.1 Case Study: Security and usability study of HISPs in group scenarios . . . 143
8.6 Summary and conclusion . . . 144

9 Conclusion and future work 146
9.1 The research problem restated . . . 146
9.2 Research contribution . . . 149
9.2.1 Analysis of the effectiveness of empirical channels for mobile device associations 149
9.2.2 Design principles . . . 150
9.2.3 Framework for reasoning about empirical channels . . . 151
9.2.4 Model and process for evaluating usability and security of empirical channels 152
9.3 Future directions . . . 153

REFERENCES 154

APPENDICES 159

A HISPs framework validation - expert feedback 160

B Proforma for collection of expert feedback 162

C ISUT: A Tool for Security and Usability Testing of Device Association Protocols 163
C.1 Introduction . . . 163
C.2 ISUT main components . . . 163
C.2.1 Protocol layer . . . 164
C.2.2 Framework layer . . . 164
C.2.3 Evaluation layer . . . 165
C.2.4 Application layer . . . 165
C.2.5 Configuration manager . . . 165
C.3 Main features of ISUT . . . 166
C.3.1 Workload management . . . 166
C.3.2 Usability testing within context . . . 166
C.3.3 Large scale testing . . . 166
C.3.4 Protocol performance analysis . . . 166
C.4 Limitations . . . 167
C.5 Related work . . . 167

D Sample test plan **168**
D.1 Introduction . . . 168
D.2 Test plan . . . 168
D.2.1 Purpose of study . . . 168
D.2.2 Problem statement . . . 168
D.2.3 User profiles . . . 169
D.2.4 Study design . . . 169
D.2.5 Participant tasks . . . 171
D.2.6 Task list . . . 172
D.2.7 Data collection and analysis . . . 173

E Questionnaires **174**
E.1 After Scenario Questionnaire . . . 174
E.2 After experiment questionnaire . . . 174
E.3 Enrolment questionnaire . . . 176

List of Figures

1.1 Human-Interactive Security Protocol . . . 3
1.2 Thesis overview . . . 12

2.1 Auxiliary device method using 2D barcode . . . 19

4.1 Single-user scenarios . . . 36
4.2 Images used in the study . . . 42
4.3 Compare and select: Completion times . . . 48
4.4 Manual copying and entering: Completion times and failures . . . 49
4.5 Participants ASQ scores . . . 51
4.6 Participants' choices: Easy and preferred methods . . . 52
4.7 Participants' choices: Difficult and unpreferred. . . . 52

5.1 Word-matching and number-typing method . . . 75
5.2 Task sequence . . . 77
5.3 Preferences . . . 83

6.1 Principle of single interaction path . . . 96
6.2 Screen shot of Word-matching and number-typing . . . 99

7.1 HISPs framework . . . 108
7.2 User-Centered Design process . . . 118

8.1 Security-usability threat model . . . 133
8.2 Process for security-usability analyses . . . 140

C.1 ISUT: tool support for security and usability testing . . . 164

List of Tables

1.1 Scenario variables 5

2.1 Empirical channels and scenarios 22

4.1 Participant demographics 38
4.2 Manual comparison: Security and non-security failures 47
4.3 Manual comparison: Completion times 47
4.4 Compare and select: Security and non-security failures 48
4.5 Barcode: Completion times and failures 50
4.6 Ranking based on SUM 54
4.7 Ranking based on security failures 56

5.1 Participant demographics 70
5.2 Performance by age 78
5.3 Security and non-security failures 79
5.4 Group members' completion times (in seconds) 80
5.5 Initiators' completion times 80
5.6 Group members' rating scores 81
5.7 Initiators' rating scores 82

7.1 Summary of scenarios showing candidate empirical channels 121
7.2 Experts' areas of specialisation 123
7.3 Summarised expert feedback - strengths, contribution, and practicality 124
7.4 Snapshot of expert comments 125

8.1 Measurable metrics 137

Chapter 1

Introduction

1.1 Human-Computer Interactions and Security

Security and usability are usually at odds. Improving one may affect the other in a negative way. For example, machine generated passwords are theoretically more secure than user chosen ones but the gained security comes at a cost of usability [132]. Yee [134] attributed this conflict to system implementers who treat security or usability as an add-on to a finished product. In addition, security and usability are at odds due to a conflict of interests that may exist between a system's owner and its users. In the music industry, for example, some implementations of Digital Rights Management (DRM) have caused concern from genuine customers because the latter cannot move their music between the gadgets that they use [116]. Moreover, security is sometimes considered as part of non-functional requirements [33] that system designers consider only after implementation of other requirements.

While the conflict between security and usability exists in many systems, there has been a realisation that improving one may help improve the other. Flechais [31], for example, argued that since security is aimed at making undesirable actions more difficult for users to engage in while usability aims at making desirable actions easily accessible, improving one should help improve the other. A usable system will minimise unintentional errors that users make while a secure system will ensure that undesirable actions are difficult, if not inaccessible, and make desirable ones easier to perform.

Research in Human-Computer Interaction (HCI) started as early as 1975 [6] focusing on improving the usability of software through a systematic approach to design. However, despite its long existence, Balfanz *et al.* [6], as late as 2004, pointed out that very little work was focusing on usability of secure systems. Secure systems continue to be poorly designed and, in some cases, forcing users to find alternative interactions with the system or avoid it completely [4]. Flechais [31] pointed out that the field of Human-Computer Interaction Security (HCISec) is focusing nearly exclusively on improving the user interface of systems. While the importance of improving a user interface in making a secure system more usable cannot be ignored, studies such as [129] show that this alone is not sufficient.

HCISec is a comparatively young and emerging field with a lot to learn from HCI. HCI has developed methodologies, concepts, and processes that ensure systematic system designs that focus on users' needs and requirements. One such process is User Centred Design [131]. Rubin [95] identified four factors that are necessary to consider during the UCD process: context, objectives, goals, and environment. A product's goals, objectives, context, and environment must be derived from users' perspective. In fact, Woodson [131] defined UCD as *the practice of designing products so that users can perform required use, operation, service and supportive tasks with a minimum of stress and a maximum of efficiency.*

While HCI focusses on tasks and how users accomplish them, security tasks are usually secondary for most users [129]. Users may view security tasks as obstacles to accomplishing their primary goals and, as a consequence, secure systems have a habit of being broken by their own users. Whether this is deliberate (like insider attacks) or accidental (mistakes or unusable security mechanisms), it is now commonly accepted that *people are the weakest link in the security chain* [105].

Labelling users as *weakest link*, however, has both negative and positive consequences depending on one's interpretation of the term. System designers and developers may interpret the term as implying that users are 'stupid' and should be taken out of the security chain. This interpretation results in systems that focus entirely on technical security countermeasures (rather than socio-technical) and alienating the very intended users of the system. On the other hand, the term has positive impact on the design and implementation of a secure system if it is interpreted that users have specific needs to be met and systems should be designed to meet these needs within the context of their application. In order to meet these needs, a socio-technical approach to design must be considered.

In this thesis, we adopt the second interpretation of *weakest link* because it recognises users' needs and limitations.

1.2 Human-Interactive Security Protocols (HISPs)

Security weaknesses in the Bluetooth pairing protocol [46] prompted a search for new and more secure protocols (e.g. [7,15,36,71,79,100,124]) that potentially may replace it or be used in scenarios where greater security is required. One proposed approach is using two channels: a high bandwidth (normal) channel, which is subject to the Dolev-Yao attack model[1] and a low bandwidth Out-Of-Band (OOB) channel. Under the Dolev-Yao attack model, messages can be modified or spoofed by an attacker whereas messages on the OOB channel cannot.

The OOB channel has low bandwidth but is not vulnerable to Man-in-The-Middle (MiTM) attacks. It is only appropriate for exchanging limited amounts of information such as cryptographic fingerprints. One interesting example of the OOB channel is direct human communication, which naturally allows certain levels of trust to be established amongst communicants. With the right security protocol, this trust can be transferred to devices that belong to the users — enabling two or more devices to establish a trusted communication that reflects the existing trust their users place on one another. In this thesis, we refer to protocols that require human action to establish a secure communication between devices as *Human-Interactive Security Protocols* (HISPs). We also refer to the OOB channel as the *empirical channel* following Roscoe *et al.* [93, 94] since it provides users with some form of empirical evidence about the security established between devices. Figure 1.1 is a graphic representation of a HISP showing the empirical (E) and normal channels.

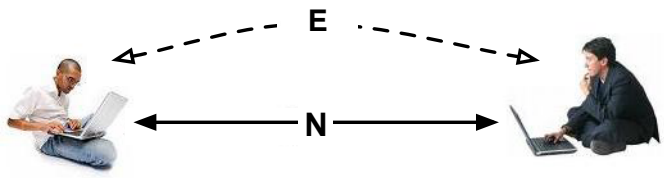

Figure 1.1: Human-Interactive Security Protocol

[1]The Dolev-Yao threat model represents an attacker that can overhear, intercept, and synthesise any message and is only limited by the constraints of the cryptographic methods used [24].

A major strength of the empirical channel is authentication and message integrity, as opposed to secrecy. In other words, two people interacting over an OOB channel (such as a face to face discussion) have good assurances of the others' identity and obvious guarantees that their conversation is not modified or otherwise tampered with, but no assurance that they are not being overheard. Thus, users can exchange messages over the empirical channel by conversation in the presence of others who are not part of the protocol (which might include a potential attacker) but the latter have no control over these messages as they cannot block or modify them. The protocol allows the authenticity and integrity of the messages exchanged over the normal channel to be verified by messages exchanged over the empirical channel.

For example, devices can exchange RSA public keys together with other data using the normal channel and independently calculate a cryptographic hash (digest) value from this information. The digest is then transmitted from one device to the other in order to verify that both have the same value, indicating that information exchanged on the normal channel has not been altered. In these protocols, transfer of digests is dependant on human users. Either both devices display the digest value which users can then compare, or one device displays its digest value and a user copies it to the other device which then checks the value. Should the values match, the public keys that were exchanged over the normal channel are thus authenticated, allowing the devices to set up secure communications (such as Secure Socket Layer/Transport layer Security (SSL/TLS), for example).

1.3 The need for Security and Usability in HISPs

HISPs present a number of challenges. First, they require users to carry out security critical tasks — creating extra work for them. This may not be easy to implement in practice, given that these protocols may run on devices that have limited input/output interfaces. Moreover, users generally experience difficulties when using security mechanisms [13, 129]. Another issue is users' understanding of the need to have a more secure protocol. The current Bluetooth pairing is relatively straightforward but given that attacks have been found against this [46], users need to understand the importance of the information they want to protect in order to buy into the need for additional security. Lacking this incentive to adopt stronger security, should a new security mechanism require more effort than currently is being expended on the knowledge-based mechanism, serious adoption issues may arise.

Moreover, different association scenarios and modes of authentication affect the security and usability of HISPs. First, authentication can be either one-way or mutual. One-way authentication requires a user to confirm authenticity of information on a subset of devices only whereas in mutual authentication confirmation must be done on all devices involved in an association. Either mode of authentication places different demands on users and understanding what these differences are, together with their effect on the usability and security of HISPs, is critical to designing empirical channels that are effectively secure. Second, different association scenarios also have a bearing on the security and usability of HISPs. The three categories of association scenarios are number of devices (pairwise or group), device proximity (close or distant), and number of users (single or multi-user). An understanding of how a combination of these scenarios, together with a mode of authentication, affect security and usability of HISPs is invaluable to designing empirical channels that work in real world settings. Table 1.1 summarises modes of authentication and association scenarios.

Mode of authentication		**Association scenarios**		
		Number of Devices	Number of Users	Device proximity
One way		2	1	Close
Mutual		>2	>1	Distant

Table 1.1: Scenario variables

Finally, the context in which a protocol is used impacts its security or usability or both. For example, a protocol may be used in sensitive situations such as in military operations, peer-to-peer electronic payment systems, online payment systems, or in group settings such as meetings. It has been found that users change behaviour depending on the social context, whether in a crowded or private environment [35, 45]. Moreover, physical environmental variables such as light and noise levels can affect usability and security of HISPs. How the empirical channel is designed to run in each of these contexts has an effect on its viability .

HISPs, however, present an opportunity. They enable users to effectively manage their own security by allowing device owners to actively participate in device associations and ensuring that only devices (or whose owners) they trust take part. This is crucial in scenarios where there is no infrastructure or third party to support device association, users do not trust a third party, or where users want to retain control of the security of associations in which their devices participate. This can apply in applications involving exchange of sensitive information such as payment and medical data.

In summary, technical security (see Section 1.4.1) in HISPs appears to run directly against certain human factors (e.g. limited accuracy, memory constraints, and issues of motivation). The larger the size of information required to authenticate an association, the more difficult it is likely to be for users to compare or transfer it. This may lead to errors that could compromise the security of the protocol, and certainly affect its performance. The critical success factor of any HISP's design is in integrating technical security requirements with those of usability requirements from the onset. Efforts in designing empirical channels need to consider both social and technical components of the system and the context in which these protocols would work.

1.4 Security of HISPs: a socio-technical approach

1.4.1 Technical security of HISPs

The primary concern of designers of security protocols is about technical security, that is, security based on mathematical assertions. As such it is not surprising that, in many instances, a protocol is said to be secure if the technical security asserts that it cannot be broken under certain conditions. The technical security of an optimal HISP is such that the odds of being successfully attacked are 1 in 2^b, where b is the size of the string that users transfer using the empirical channel. In this thesis we will refer to this string as a *digest*[2]. Choosing the value of b that is sufficiently large for a given application gives guarantees about technical security. Technical security requirements, however, may run against usability which may result in a difficult-to-use system resulting in user behaviour defeating technical security. The challenge in achieving acceptable security is finding an optimal digest size that maximises technical security without negatively impacting usability.

1.4.2 Effective security of HISPs

Security is a process that depends on all interconnected components that are part of one system [106]. It, therefore, depends on all components working correctly and mathematical assertions are just part of the whole system. There are several examples where technical components of secure systems have

[2]A digest is a cryptographic function related to a *universal hash function*. It has two arguments, namely a key and data to be digested. It should be designed so that *inter alia* the likelihood (as the key k varies) that $digest(k, A) = digest(k, B)$ is minimised for all $A \neq B$ [79].

been formalised and proved secure (mathematically) but have been broken by their own users. The Russian army in World War I, for example, found its cipher system too hard to use and reverted to simpler systems that were easily broken by the Germans [4]. Schneier [106] has argued that security is not just mathematics as it also involves people and, as such, secure systems could be broken due to improper use rather than just mathematical weakness. We, therefore, define *effective security* as the resultant security taking into account technical security and the threats from user interactions in a social-technical system.

Effective security depends on the weakest link in the chain. In HISPs, such a link may be the very users, intended for the system, who have no intention of breaking their own system but may do so through improper use or by resorting to insecure behaviours because they find systems that are to protect them difficult to use. Assessing effective security of a system requires a holistic approach to understand the different components involved, how they interact, and what the impact of those interactions is.

Human-Interactive Security Protocols require users to carry out security critical tasks with high degrees of accuracy. Users, however, are usually faced with competing tasks and security critical tasks are not their primary goals. As a result, users are unmotivated to carry out these tasks with required attention and accuracy. Lack of motivation and attention may cause security failures[3] or usability problems (or both).

1.4.3 HISPs as socio-technical systems

Secure systems are socio-technical in nature. HISPs, in particular, require users to carry out security critical tasks that users may regard as secondary. Users' actions in these protocols bring about threats that cannot be covered by mathematical assertions. For example, will users compare digests accurately? Will they bother to compare and not skip this step? Can they be duped into associating devices whose digests do not match? How large a value of b can they effectively deal with? The success of HISPs to a great extent depends on fitting into the social contexts in which they may operate. This requires an understanding of the impact of human behaviour in different contexts and designing protocols that fit user needs for different scenarios.

[3]Security failures are errors that could compromise the security of the protocol

Attempts to minimise human effort needed to transfer or compare digests focus on user interfaces even though this alone does not necessarily achieve the desired outcome or improve overall usability [31, 129]. They also focus on devices of similar capabilities such as devices with very limited input and output interfaces (e.g. single button devices [113]), or devices with reasonable input and output interfaces such as camera phones [72]. There are, however, many situations where devices of differing capabilities need secure association, and these may in fact be the common ones. Examples include a Bluetooth hands free set and mobile phone, a PDA and printer or external storage media.

Attempts to improve usability of empirical channels should cover a wide range of scenarios rather than merely claiming to find a universal solution that covers *all* use scenarios. The pervasiveness of mobile device interactions demand empirical channels that can apply to a wide range of association scenarios, physical and social contexts. Designing such empirical channels requires an understanding of the different factors that affect them in different scenarios and contexts. Such an understanding can only be developed if specific scenarios and contexts are analysed rather than focusing on a universal solution — which is likely to overshadow crucial factors that may only be uncovered when a specific scenario is considered before generalising to other scenarios. For example, a factor such as *proximity* may be important in one application scenario (such as telephony) and not in another. Specific scenarios, such as this, provide insights that are crucial to designing secure and usable empirical channels that are applicable to a wide range of contexts.

1.4.4 Security and usability evaluation of HISPs

Usability evaluations of secure systems require procedures that deviate from standard HCI techniques. Whitten [128] highlighted the differences between secure software and other software and why usability evaluation of secure software is difficult. In addition to encompassing main elements of education software (such as learnability), safeware (no undo for dangerous errors), and general consumer software (all kinds of users, goals set by users, no training), she highlights properties that make security difficult. These include the secondary goal (unmotivated user) property, hidden failure property, barn door property, abstraction property, and weakest link property. A usability evaluation of secure software should not focus on usability to the exclusion of security: in certain cases it is necessary, for the purposes of security, to include behaviour that is complex. Conversely it is possible to weaken the security of a system by simplifying or automating certain elements, which

usually improve usability. Usability and security have a closely tied relationship, it is important to consider both factors when evaluating a system.

In HISPs, the pervasive nature and frequency of mobile interactions transverse elements of education, safeware, and general purpose software. Evaluation of association methods focus on usability. In this regard, insecure association methods are recommended for use in HISPs. For example, Uzun *et al.* [123] recommended using manual comparison as a method for device pairing. This is despite the fact that this method is susceptible to security failures. In another study, Kobsa *et al.* [60] recommended using PIN comparison for devices with interfaces and audio comparison for devices without displays. Both of these methods are susceptible to security failures and are affected by environmental variables such as noise and lighting conditions. The failure to recognise the seriousness of security failures in these methods is due to lack of a methodology for identifying elements that may affect not only usability but also security of a system and a process of how to assess these.

1.5 Research problem

HISPs rely on users to transfer digests among devices. Thus, these protocols are actively constrained by the fact that the level of security offered may depend on the amount of human effort expended in transferring or comparing digests. We, therefore, in this thesis concern ourselves with the process of exchanging digests by users as this is critical to achieving desired effective security. We aim to investigate how digests of 4 or more digits can be compared between two or more systems, in a way that enables as much data as possible to be compared reliably, as conveniently as possible, while ensuring that issues such as user complacency do not compromise security.

We address the following research question:

> *Given the different association scenarios and modes of authentication in Human-Interactive Security Protocols, how can we improve on existing, or design new, empirical channels that suit human and contextual needs to achieve acceptable effective security?*

Answering the above research question requires an understanding and analysis of association scenarios, modes of authentication, contexts, and effective security. In addition, an evaluation of existing empirical channels to assess their effectiveness is crucial. The integration of technical security and

usability must achieve *acceptable* effective security for a given application. This means that the quest to improve usability should not unduly compromise technical security, while the quest for attaining high technical security should not make the protocol unusable. While a number of methods have been proposed, no research has looked at the design of empirical channels that integrate usability and technical security, and how currently proposed methods may compromise effective security.

The above research question is decomposed into four specific research problems that this thesis addresses:

1. Which methods for transferring or comparing digests are effectively secure? This question aims at assessing effectiveness of currently proposed empirical channels through identification of their weaknesses and strengths for both security and usability.

2. In what association scenarios and modes of authentication can these methods work while achieving acceptable effective security? A method that is acceptable and secure in one mode of authentication and association scenario may not be in another. This research question is aimed at identifying human, environmental, and contextual factors for the different association and modes of authentication scenarios that may have an impact on the security and usability of empirical channels. We address this and the previous research question in Chapters 4 and 5.

3. How can we design or improve existing empirical channels to achieve acceptable effective security? This research problem is aimed at developing design principles and methodologies for developing new, and improving existing, effectively secure empirical channels. Chapters 6 and 7 attempt to address this problem.

4. How can we improve methodologies and procedures for usability and security evaluation of HISPs? Given the differences between systems evaluated (and goals of evaluation) under HCI and those under HCISec, either HCI methodologies and procedures need to be adapted to meet HCISec requirements or new methodologies and procedures specific to HCISec be developed. We address this problem in Chapter 8.

1.6 Research contributions

This thesis addresses the above problems and presents the following original research contributions:

1. **Analysis of the effectiveness of empirical channels for mobile device associations**

   An evaluation of the effectiveness of empirical channels in both single-user pairwise and multi-user group association scenarios. This thesis examines the effectiveness of empirical channels and develops recommendations based on results of empirical user studies. It identifies crucial factors affecting usability and security of empirical channels and also the subtle differences between single user and group association scenarios.

2. **Design principles**

   This work develops principles for designing and implementing empirical channels. These principles are demonstrated with new proposals for empirical channels.

3. **Framework for reasoning about empirical channels**

   There have been several proposals for empirical channels all of which take a single faceted approach that ignores crucial factors rendering proposed methods insecure or unusable in common use scenarios. The framework is aimed at helping both researchers and system designers in reasoning about empirical channels and developing/choosing methods that are secure and usable.

4. **Model and process for evaluating usability and security of empirical channels**

   Following HCI standard procedures and methodologies in conducting security usability studies has resulted in having results that focus entirely on usability issues (ignoring security) or results that lack external validity because participants carry out security tasks as primary tasks or both. An original contribution of this thesis regarding this problem is a model for identifying security and usability issues that may affect a specific empirical channel and a process for evaluating security and usability. Both the model and process for evaluating usability and security are generalised for application to secure systems in general rather than just empirical channels.

1.7 Thesis overview

Chapter 2 reviews the literature on HISPs, empirical channels, and studies conducted to analyse security and usability of HISPs. Chapter 3 discusses research methodologies used during this research. Chapters 4 and 5 present empirical evaluations of empirical channels in single-user and group

scenarios respectively and identify factors that affect usability and security of HISPs in both scenarios. Chapter 6 presents a set of principles for designing empirical channels. Chapter 7 presents a framework for reasoning about empirical channels. Chapter 8 presents a model for identifying security and usability factors and a process for evaluating them and Chapter 9 concludes the thesis and proposes future work. Figure 1.2 summarises the thesis overview.

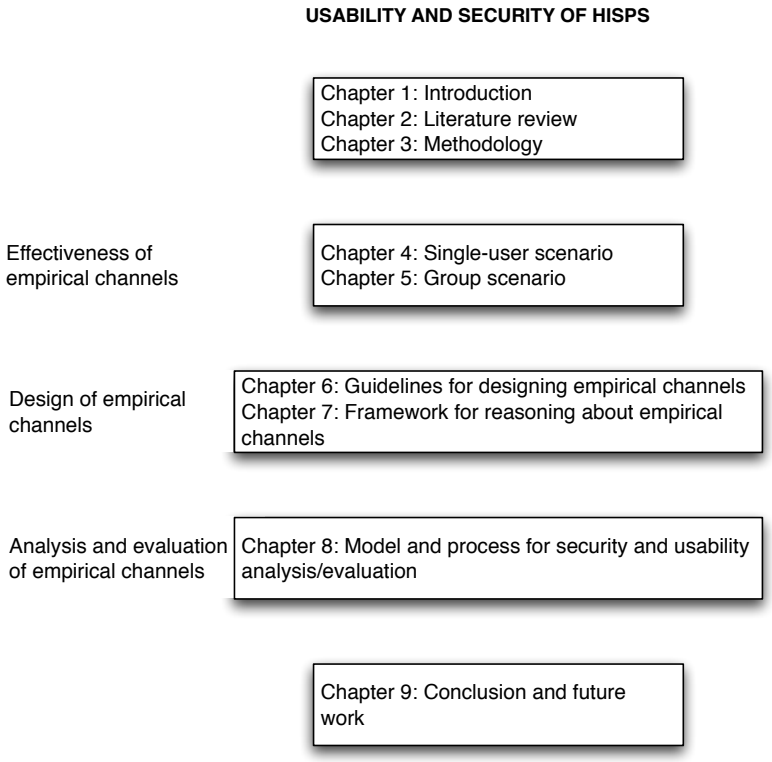

Figure 1.2: Thesis overview

Chapter 2

Background and literature review

2.1 Definitions

Usability: The International Organisation for Standards (ISO) [86] defines usability as *the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.* This definition focusses on users' goals (effectiveness), the speed with which goals are achieved (efficiency), and users' satisfaction with the system within a specified context. The definition implies that usability is contextual; a system deemed usable in one context may not be so in another. Other definitions of usability include other elements such as learnability [110] and memorability [82]. Consolidating various definitions, usability then consists of effectiveness, efficiency, satisfaction, learnability, and memorability. A usability evaluation of a system, therefore, focusses on one or more of these elements.

Whitten and Tygar [129] made the first attempt in defining usability for security. In their definition, security software is usable if people who are expected to use it:

1. are reliably made aware of the security tasks they need to perform;
2. are able to figure out how to successfully perform those tasks;
3. do not make dangerous errors; and
4. are sufficiently comfortable with the interface to continue using it.

In HISPs, users must be aware of the tasks they are required to perform and should be capable of performing those tasks successfully. Users should be aware of their need to securely connect two mobile devices (and transfer files between them, for example) and more importantly must be able to accomplish an association with minimum effort without compromising security.

Security: Definitions of security revolve around attackers — who are typically regarded as agents with malicious intent rather than legitimate and trustworthy users of a system. For example, Gollmann [37] defines computer security as *concerned with the measures we take to deal with* intentional *actions by parties behaving in some unwelcome fashion.* This definition implies that security should only be concerned with malicious intent and ignores threats from users who have no intention of harming the system or owner. Flechais [31], however, points out why non-malicious users may break security of a system; a user may not perceive the interaction to be detrimental to the system or has a greater incentive to engage in an insecure interaction. Deviating from standard definitions of security, Dourish [25] points out that systems must not only be secure but usably and practically secure. This statement highlights the need for systems to be secure within the context in which they operate. In this thesis, following Gollmann's definition, the term security is used to mean measures we take to deal with *intentional and unintentional* actions by parties behaving in some unwelcome fashion.

Socio-technical systems: Beynon-Davies [10] defines a socio-technical system as a system of technology used within a system of activity, such as an Information System while the Oxford English Dictionary defines socio- as *relating to society (the aggregate of people living together...*). This thesis uses the term to refer to a piece of technology used by people to achieve specific goals.

Context: The emergence of ubiquitous computing saw the term *context* coming into the computing science domain. The most widely accepted definition of the term in the computing domain is that offered by Dey [21] who defines context as: *...any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.*

The above definition implies that we can draw out specific information about the environment in which a user and application interact. It also means that this information is not finite and a designer or modeller of a system must understand what is relevant to the application at hand. Information about the mode of authentication, physical and social environment, proximity of devices, association

scenarios, device affordances, and number of users involved can be used to characterise the situation of HISPs. In the context of this thesis, therefore, the term context is specifically used to mean variables related to mode of authentication, environment, device affordances, proximity, association scenario, and number of users that may affect a user's interaction with HISPs.

Human factors: Human factors research is defined as a *multidisciplinary field that draws on the methods, data, and principles of the behavioural and social sciences, engineering, physiology, anthropology, biomechanics, and other disciplines to design systems that are compatible with the capabilities and limitations of the people who will use them* [77]. In this work the term human factors is used to denote users' characteristics that may help or hinder users achieving their goals in a secure and efficient manner.

2.2 Human-Interactive Security Protocols

2.2.1 Symmetrised Hash Commitment Before Knowledge (SHCBK) protocol

Nguyen and Roscoe [79, 80] proposed the SHCBK protocol that uses a digest exchanged over the empirical channel to authenticate information exchanged over the normal channel. The protocol bootstraps security in *ad hoc* networks from zero assumptions. As pointed out earlier, it is common in these environments that devices have no pre-shared information or a common third party that can authenticate them. The SHCBK protocol aims to achieve a high level of authenticity of the information that devices exchange. Unlike most of the proposed protocols for *ad hoc* networks, SHCBK covers a wide variety of scenarios including single user pairwise or multi-device associations, group associations, and close or distant device associations. The SHCBK protocol achieves the goal of authentication through the use of a combination of the normal Dolev-Yao channel and empirical or human mediated channel. The protocol is presented below.

1. $\forall A \longrightarrow_N \forall A' : A, INFO_A, longhash(A, k_A)$
2. $\forall A \longrightarrow_N \forall A' : k_A$
3. $\forall A \longrightarrow_E \forall A'$:users compare *Digest(k*,INFOs)*
   where k* is the XOR of all the $k'_A s$ for A $\in$ G

During the protocol run, each device sends its identity A, any information it wants to be authenticated including its public key $INFO_A$ and a *longhash*[1] of its identity and a key k_A. In the second message each device discloses its key to all the other devices. Using the key and previously exchanged information, devices independently calculate a digest of the XOR of all the keys k_As (k*) and *INFOs* and users compare the digest.

In the SHCBK protocol, we see that Messages 1 and 2 are transmitted on the normal Dolev-Yao channel ($\longrightarrow_N$) where messages can be overhead, intercepted, and synthesised. Message 3 is transmitted over the empirical channel ($\longrightarrow_E$) that guarantees authentication and integrity but not secrecy. By using the SHCBK, devices can exchange public keys, long term or ephemeral, together with other data using the normal channel and independently calculate a digest from this information. The digest is then transmitted from one device to the other in order to verify that both have the same value, indicating that information exchanged on the normal channel has not been altered. In the SHCBK protocol, transfer of digests is dependant on human users. Either both devices display the digest which users can then compare, or one device displays its digest value and a user copies it to a personal device which then checks the value. Should the values match, the public keys that were exchanged over the normal channel are thus authenticated, allowing the devices to set up secure communications such as Secure Socket Layer(SSL) and Transport Layer Security (TLS) if desired.

The exchange of information in the protocol does not require devices to be close together. Participants in the protocol could be far away from each other and use Internet to exchange public information that requires authentication. Once a digest is calculated at both ends, it can be exchanged using an alternative channel, say by telephone. This brings a whole range of applications that the SHCBK protocol may be applied to including electronic payments, bootstrapping secure sensor networks, secure voice over IP, and secure online or group meetings.

An intruder may attempt to attack the SHCBK protocol in two ways. First, an intruder can try a combinatorial attack where participants share differing information but end up with the same digest. The SHCBK, however, forces participants to commit a final digest (by sending $longhash(A, k_A)$) without knowing the its value. In doing so, an intruder is limited to a single random guess that is no better that 2^{-b} where b is the size of the digest. The proof of the security of the protocol against a combinatorial attack is given in [79, 80]. In this thesis, we focus on the second attack where an intruder can dupe users to agree to non-matching digests. In this attack, an intruder

[1] *longhash* is a strongly collision-resistant and inversion-resistant hash function

passively participates in an association and relies on users to agree to digests that are non-matching. For example, a user associating two mobile devices may pair each of the devices to a third that is unintended if digests are not compared accurately.

Other proposed protocols include MANA (MANual Authentication) [36], Seeing-is-Believing [72], Visual Channel Mutual Authentication Protocol (VIC) [100], Human-Verifiable Protocol [71], and Basic Pre-authentication Scheme [7]. A common factor in all HISPs is that human effort (both mental and physical) is required either during transfer of data between devices or during confirmation of an *accept* or *reject* on one or more devices. There are, however, differences between protocols in terms of human effort required. For example, MANA I and II protocols require between 32 and 40 bits for the digest while SHCBK and MANA III require 16 bits. However, due to differences in efficiency, different protocols provide varying levels of technical security for a given number of bits. For example, the SHCBK provides technical security equivalent to the size of the digest while MANA protocols only provide security equivalent to half the size of the digest. Moreover, there are also differences on the properties that an empirical channel must possess. SHCBK and others require only authentication while MANA protocols require secrecy too.

2.3 Empirical channels

2.3.1 Manual comparison (MC)

MC requires a user to compare digests displayed on two or more devices and indicate, on the devices, whether they match or not. A digest is encoded into one of a variety of different forms including images [88], sentences [39], digits [40], and sound [114]. These methods rely on the user's ability to compare two data items accurately and indicate correctly on the devices the result of the comparison. In addition to user mental and physical effort, manual comparison requires devices with a display (or speaker for sound) and an input interface through which a user can indicate a match or disparity of compared digests. Furthermore, these methods require a sufficient level of light when comparing strings or images, and comparing sounds is obviously dependent on environmental noise levels. *Manual comparison's* potential is tempered by security failures. It, however, only requires hardware features that basic devices usually have and studies have found it more usable compared

to other methods [53, 60, 123]. In addition, it is appropriate in single as well as group scenarios and to both close and distant device scenarios.

2.3.2 Manual copying and entering (MCE)

Manual copying and entering involves having one device display its digest for a user to copy to another device which then compares the entered value with its own. A device then indicates whether the digests match or otherwise. This seems promising for two reasons; first, the popularity of Short Message Service (SMS) [121] means that many users are comfortable with entering text on devices with limited input interfaces and, second, it is more efficient and effective to have devices compare digests than it would be for users. These methods, however, require one device with a display, another with a keypad and a way of indicating to the user a match or disparity of the comparison. They also require sufficient light or lit display for users to read strings without much difficulty. Human effort required in *manual copying and entering* poses a significant challenge. For example, studies including [53,60,123] show that users find this method less usable compared to MC. Moreover, human effort increases linearly with the size of digest and number of devices (in single-user scenarios). Other factors such as device affordances[2], the format in which the digest is represented, and association scenario affect its usability. In addition, personal factors such as experience with mobile devices, especially texting in cases where digests comprise alphanumeric characters, affect usability of the method.

2.3.3 Auxiliary device methods (ADM)

Auxiliary device methods use devices or hardware as empirical channels. An auxiliary device is used either only to transfer data or to also verify the authenticity of the exchanged data. For example, devices with a display may encode a digest into a 2D-barcode (see Figure 2.1) and an extra device (not participating in the association) with a digital camera and appropriate software reads (equivalent to transferring information) all barcodes and compares (equivalent to verification) the decoded information.

[2] Affordances are the means through which a user interacts with a device [16]

Figure 2.1: Auxiliary device method using 2D barcode

A user indicates on associating devices the result of the comparison (confirmation). In some cases, one or more devices participating in an association may have the capability to act as an auxiliary device and thereby avoid the necessity of an extra device. Proposed devices include a data cable, digital camera, microphone, speakers, and external storage media such as memory cards. *Auxiliary device methods* are hampered by the need for extra hardware and devices with standardised interfaces. In addition, they are also constrained by proximity requirements — they can only work when devices are close together — and may be practical only for pairwise scenarios.

2.3.4 Short-range directed channels (SDC)

These methods require users to align associating devices to facilitate exchange of information using short-range location limited channels such as infra-red [29], light [71, 100], and integrity regions [15]. They have the advantage of not relying on users to compare, enter strings, or carry extra devices, hence, place fewer demands on users. However, lack of human verification of an association can be exploited by attackers [114] as users are unable to detect anomalies during the process. In addition, security weaknesses in these technologies [114] and user inattentiveness [15] may allow unexpected devices to participate in an association undetected. SDC channels are also only practical for pairwise association of devices that are physically close to each other. Moreover, the need for hardware such as laser beamers and readers (proposed in [71]), that are non default in most devices, make these methods impractical.

2.3.5 Timing methods (TM)

Timing methods rely on transmission of cryptographic information within well timed intervals such that an intruder finds it hard to synchronise and successfully attack an association. Proposed methods include shaking devices together [70] and button pressing [101]. These methods provide some guarantee to users that devices touched are the ones communicating at that particular instance and the communication is a result of that particular user action. The direct communication between devices reduces security failures that may result from user action. *Timing methods*, nevertheless, are only appropriate for single-user pairwise scenarios where devices are in close proximity. They are impractical in single-user multi-device, group, and distant device scenarios as synchronising devices in these contexts is difficult. They also vulnerable to being watched, that is, need privacy and may require specific devices or affordances. For example, shaking devices relies on accelerometers and devices that a user can shake without difficulty. This excludes a whole range of applications in which HISPs may be applied including association between mobile device and laptop, desktop, or vending machine.

2.3.6 Limitations of current empirical channels

2.3.6.1 A choice between security and usability

Currently proposed empirical channels force one to choose between security and usability. *Manual comparison* has been found both more usable (compared to other methods) and preferred by users but also susceptible to security failures [53, 60, 123]. Other methods such as timing and auxiliary device methods are also susceptible to security failures [101, 113] though they do not offer the same level of usability as *manual comparison*. On the other hand, *manual copying and entering* is not susceptible to security failures but requires users to type at least 4 digits. This is particularly problematic in applications where a user needs to carry out device associations several times a day. Both usability and security should be designed into an empirical channel such that users do not have to choose between them.

2.3.6.2 Limited application

Even though mobile interactions are ephemeral and context may change from one interaction to the next, current proposals seem to focus on specific unchanging contexts. For example, timing methods and short range directed channels may work well in pairwise associations but may be difficult to extend to group scenarios or where devices are a certain distance apart. In addition, the range of affordances on mobile devices is diverse, ranging from netbooks and smart phones to Bluetooth hands-free sets. Current proposals of empirical channels focus on devices of similar affordances such as cameras [72], accelerometers [70], laser beamers and readers [71,100]. There is anecdotal evidence that shows that the majority of *ad hoc* mobile interactions occurs either between devices of sufficient input/output capabilities or between an input/output rich device and an input/output constrained device. With this in mind, it is possible to exploit devices with rich input/output interfaces to leverage limitations of input/output constrained devices. Current proposals, however, fail to take this into account and have led to an assumption that most mobile interactions occur between devices with poor input/output interfaces. Empirical channels should cover as many scenarios as possible and, by taking advantage of the fact that one device in the association is likely to have rich input/output interfaces, better methods that also apply to interface constrained devices can be developed.

2.3.6.3 Unmotivated user property

Users tend to avoid, whenever possible, security tasks because they are not their primary goals in most scenarios. While empirical channels rely on users to accurately perform security critical tasks, many do not compel users to perform required tasks. For example, in *manual comparison*, users can ignore comparing but simply assume that the digest is matching. Even vigilant users may be distracted, accidentally press a wrong button, or merely miss the difference. In fact, in the study of pairing methods reported in Chapter 4, some of the participants were concerned that with *manual comparison* it was easy for a security failure to occur because the method did not force them to compare accurately.

2.3.7 Summary

Currently, proposed empirical channels are limited in the scenarios to which they can be applied. This is because proponents of empirical channels fail to take into account factors that are central to designing methods that are both secure and usable in different scenarios and use contexts. Table 2.1 summarises empirical channels showing the scenarios to which they can be applied to and their resistance to security failures.

	Scenario	Manual Comparison	Manual Copying and Entering	Auxiliary Device Methods	Short-range Directed Channels	Timing Methods
Proximity	Close	✓	✓	✓	✓	✓
	Distant	✓	✓	×	×	×
Authentication	Asymmetric	✓	✓	✓	✓	✓
	Symmetric	✓	✓	✓	✓	✓
# of users	Single	✓	✓	✓	✓	✓
	Group	✓	✓	✓	×	×
# of devices	Pairwise	✓	✓	✓	✓	✓
	Multiple	✓	✓	✓	×	×
Resistant to security failures	User	×	×	×	✓	×
	Technical	✓	✓	✓	×	×

Table 2.1: Empirical channels and scenarios to which they may be applied.

An important point to note about Table 2.1 is that each scenario is considered independent of another. For example, *timing methods* are shown to work for symmetric authentication but this is only the case in pairwise as it is almost impossible for a user to synchronise actions on more than two devices. Nevertheless, scenarios are not mutually exclusive in practice and a combination of two or more scenarios puts further constraints on requirements of empirical channels.

2.4 User studies

A number of user studies of secure systems have been conducted including authentication systems [1,12,13,63,130,130], secure email [129], security tools [17], and Identity-Management software [48]. A summary of some of the usability studies of HISPs is presented below.

2.4.1 Usability Analysis of Secure Pairing Methods

Uzun *et al.* [123] evaluated the usability of the Bluetooth pairing protocol using manual comparing and copying of short strings (6 digits). In the study, participants found copying of strings more difficult than comparing, even though they felt the former was more professional. As a result, the authors of the study recommended comparison as the best method, ignoring the fact that it is subject to security failures. In addition, the authors recommended that checksums and passkeys should not be longer than 7 digits — based on Miller's findings [73] that the maximum number of chunks human working memory can hold at any one time is 7. In these protocols, however, users do not need to memorise any strings and it is also possible to split the string into smaller chunks for easier comparison or copying.

2.4.2 Analysis of Bluetooth Simple Pairing and Wi-Fi Protected Setup

Kuo *et al.* [64] conducted a usability evaluation of Bluetooth Simple Pairing and Wi-Fi Protected Setup and concluded that the issues uncovered are as a result of the existence of multiple methods for each technology. The Bluetooth Special Interest Group's (SIG) Simple Pairing [40] specifies four different pairing models: "Just Works", "Passkey Entry", "Numeric Comparison", and "Out of Band". The Wi-Fi Alliance's Protected Setup specification [3] specifies three different ways of pairing: "Push Button Configuration", "PIN entry", and "Out-of-Band". Kuo's *et al.* recommendation is that it is necessary to have a common baseline across hardware features and a consistent, interoperable user experience across devices. While adopting a smaller set of methods in pairing protocols can reduce user burden and improve overall security (by reducing possibility of failures), standardising hardware is expensive. The proposal to use physical channels such as cable and external storage media, for example, has failed because it required standardised interfaces across devices. In addition, the differences in affordances in mobile devices make it difficult to use one method across all devices. For example, typing digits may work on mobile phones but may not on Bluetooth headsets.

2.4.3 Human Factors in HCBK protocols

A usability analysis of the Hash Commitment Before Knowledge [92] protocol was carried out to identify usability factors that may compromise security or cause major adoption issues [52]. The study concluded that security provided by a particular security protocol is not an all-or-nothing but a total sum of different factors and all components that interact with it. In addition, it concluded that the major problem with comparing strings is that users are not compelled to pay attention and carry it out accurately and devices have no way of detecting when a user has not taken care. Nevertheless, the method has the flexibility to accommodate a variety of strings as well as images and sounds.

2.4.4 A Comparative Usability Study of Secure Device Pairing Methods

Another notable usability study on device pairing was conducted by Kobsa *et al.* [60]. Several empirical channels were evaluated including comparing images, sounds, digits, and sentences. The study recommended using comparing digits, sentences, and images (in that order) for devices with a display and comparing sound for devices without a display. There are two issues with this recommendation. First, the study did not evaluate security of any of the methods tested and, therefore, did not consider the fact that comparing is subject to security failures. Second, the study failed to appreciate that environmental differences between a laboratory and where device pairing may occur have significant effects on some of the methods; comparing sound for example. The second recommendation generalises the results to the population despite having only 22 (11 male) participants. For such a small sample size (worse still when split into 3 different age groups of 18-25, 25-40, and 40+), it is statistically incorrect to generalise results to a population of millions of potential users. Such declarations of external validity require a sample that is representative of the population and of sufficient size. In addition, participants must do representative tasks as they would in the real world as opposed to security tasks only.

2.4.5 Other studies

There have been a number of studies that have focused on one or two types of empirical channels only. For example, a study of device pairing using accelerometers requiring users to shake devices [70]

and another on stimulus-response requiring users to press a button in response to device vibrations or change in lighting [101]. Unfortunately these studies never compared their results with any other empirical channel evaluation and seem to ignore contextual factors that may hinder these methods. For example, shaking devices is not feasible if one wishes to connect to a vending machine and pressing a button in response to changes in light may only be possible if a single user is in control of both devices.

More recently, a number of studies have been conducted including comparative studies of two-user [61] and group scenarios [85], usability study of distance bounding method [125], a comparative study of single user scenarios [62], and method specific studies [90,102,103]. A conclusion that can be drawn from these studies is that different contexts impose specific challenges on empirical channels and that a universal solution to the device association problem is yet to be found. These studies, however, fail to acknowledge the different factors that influence usability and security of empirical channels. A study conducted by Ion *et al.* [45], however, is noteworthy. It is the only study in the literature that investigated users' perceptions, security needs, and social factors and how they influence choice of a pairing method. This study found that users will make different choices in different social and physical environmental contexts.

2.4.6 Summary

Usability studies are not only necessary but crucial in identifying vulnerabilities of empirical channels when used in a socio-technical context. These studies should not aim to identify usability problems only, but also the impact of user actions on effective security. By identifying both usability and security problems, only then will new and better methods be proposed. There is, however, a common failure in methodology through limitedness to specific devices, association scenarios, or modes of authentication. In addition, studies focus on usability resulting in recommendations that fail to appreciate impact of user actions on security.

2.5 Conclusion

HISPs face both security and usability challenges because technical security requirements run again usability. Increasing the size of the digest may result in poor usability forcing users to resort to

insecure behaviour while improving usability may mean compromising on technical security. Currently proposed empirical channels may fail to compel users to carry out security tasks correctly, apply only to one mode of authentication or association scenario, employ expensive hardware, or be subject to security failures. These proposals also fail to acknowledge the impact of different physical and social environmental contexts. Studies to evaluate usability of empirical channels focus on usability, ignoring the impact of user action on effective security. This demonstrates the weaknesses of the current state of the art of applying HCI methods, that focus on evaluating primary tasks, to secure applications. In this thesis, we address these issues by first developing an understanding of the issues that affect security and usability of empirical channels in different scenarios. We then build on these findings to develop principles for designing empirical channels. Third, we develop a framework to help designers and researchers in reasoning about empirical channels and help them choose or develop methods that are secure and suit human and contextual needs. Finally, we propose a model for identifying elements that may affect the security and usability of HISPs and a process for evaluating security and usability.

Chapter 3

Methodology

3.1 Introduction

Real-world evaluation of the security and usability of Human-Interactive Security Protocols remains a challenge given the lack of deployment of these protocols. The protocol closest to HISPs that is deployed in the real-world is Bluetooth. Bluetooth pairing, however, usually occurs in private spaces. This is because there is no deployment of the protocol for publicly conducted tasks such as point of sale payments and vending machine transactions. This makes it difficult to observe Bluetooth pairing. Moreover, the Bluetooth protocol is limited to only pairwise associations. Deploying HISPs for usability and security evaluation would be prohibitively expensive as it requires infrastructure to support the various interaction scenarios. For example, Point Of Sale terminals and vending machines should support such interactions for payment applications. Laboratory experiments, however, have been used in both academia and industrial studies. These studies have contributed to the body of knowledge in research and have shaped the design of products in industry (See [17] for examples). In fact, some of the most cited HCISec usability studies, including [129], are laboratory based experiments. Given the above constraints to conducting real-world studies of HISPs and the validity of laboratory experiments as demonstrated by previous research, the latter was used to gather empirical data in this research.

Evaluating the effectiveness of empirical channels calls for collecting and analysing both quantitative and qualitative data. Quantitative data is essential for performance measures such as completion times and failure rates. Qualitative data, on the other hand, affords information on users' understanding and perceptions of evaluated methods. To collect both types of data, we logged performance data and employed interviews, questionnaires, and observations. In addition, we used triangulation (to identify and deal with inconsistent data), pilot studies (to evaluate study procedures and tools), and test plans (to ensure consistency in conducting a study). In this chapter, we present methodologies used in this research. We begin by a discussion of laboratory user studies in Section 3.2 and follow on with a discussion of data collection methods in Section 3.3. We then present the research approach in Section 3.4 and discuss the validity of the research methodology in Section 3.5.

3.2 Laboratory experiments

Also referred to as the *experimental method, user experiments,* or *controlled experiments* [30], laboratory experiments are the basic approach for researchers whose aim is to generate knowledge from findings and generalise results beyond the environment in which the study is conducted [111]. In a user experiment, a hypothesis should either be proved or refuted and the results obtained used to build on the theory of usability. In addition, experiments should employ statistically significant sample sizes and the results statistically analysed. This is in contrast to informal or usability testing studies [111] where the aim is to uncover as many usability problems in a product as possible in the shortest possible time. Laboratory experiments are widely accepted in the scientific community, HCISec included. They have been conducted on secure systems including authentication systems [13, 63, 130], secure email [129], HISPs [53, 56, 60, 123], and identity management software [48]. User experiments produce results that are reliable and valid only when attention is paid to the following:

3.2.1 Reliability

Reliability of an instrument means producing the same measurement every time an instrument is used under similar conditions [66, 127]. User experiment results are, therefore, only reliable if instruments used in gathering data are consistent. In this work, we used the After Scenarios Questionnaire

(ASQ) because of its reliability [66]. The Integrated Security and Usability Testing (ISUT) tool (See Appendix C) used for logging performance data was thoroughly tested to verify that it consistently produced the same results each time it was used under similar conditions.

3.2.2 Internal validity

Internal validity refers to an instrument measuring what it says it measures [115]. In user experiments, researchers measure specific aspects of usability including efficiency (completion times), effectiveness (completion rates), and subjective scores such as satisfaction. To measure these parameters, instruments such as questionnaires, hardware and software tools are employed. There are a number of standard usability questionnaires whose validity has been verified including After Scenarios Questionnaire (ASQ) [66], Post Study System Usability Questionnaire (PSSUQ) [67], System Usability Scale (SUS) [11], and Software Usability Measurement Inventory (SUMI) [58]. ASQ was used in this research particularly because it consists of only three questions that focus on three aspects of usability: efficiency, effectiveness, and satisfaction. The length of the ASQ was important in this research because it was crucial to ensure that the questionnaire did not become a distraction to participants.

3.2.3 External validity: context

In a laboratory experiment, a researcher has a large degree of experimental control over variable manipulation [43] and the approach may be cost effective compared to field (real-world) studies. Despite conducting studies in a laboratory setting, researchers aim to achieve external validity; to draw conclusions that extend beyond the laboratory environment. To achieve this, participants must be representative of target users and scenarios and tasks must be closely matched to real world experiences. In addition, procedures of how instructions are given and how data is collected and analysed have an effect on the results and conclusions drawn from them. These tight controls on participants and their recruitment, scenarios, tasks, instructions, and procedures are aimed at ensuring that an experiment closely represents a real world scenario (and its context) and that results obtained remain valid in the real world as much as they are in the laboratory. In this research, we closely followed these controls to ensure external validity of the studies.

3.2.4 Statistical analysis

Statistical analyses enable researchers to generalise their findings with a degree of certainty. For example, if participants are able to successfully complete task A four times out of five tries, it is incorrect to claim that the 'target population will be able to successfully complete task A 80% of the time'. However, with statistical analysis we can make statements such as 'the target population will be able to successfully complete task A 80% of the time at 95% confidence level'. The second statement, not only generalises the findings but also states a margin of error within which the generalisation is expected to be correct. There are many other areas where the use of statistical analysis is crucial such as finding differences between two systems, two populations or correlations between variables. Moreover, statistical analysis is the only way to nullify or validate a hypothesis. In this work, statistical analyses were employed to nullify or prove a hypothesis and to determine the differences between different methods and between different groups of participants.

3.2.5 Sample sizes

In controlled experiments, obtaining results with external validity requires statistically significant sample sizes. HCISec studies, however, have largely resorted to discounted [83] or informal [95] usability studies — a method meant for iterative usability evaluation and recommended for industry practitioners — where small sample sizes are acceptable. In our research, where the motivation for conducting a usability experiment is to prove or nullify a hypothesis and build on the theory of usability, we determined sample sizes based on the confidence level and statistical power (the probability of correctly rejecting the null hypothesis) [122].

3.3 Data collection methods

3.3.1 Questionnaires

Questionnaires are a common methodology for gathering qualitative and quantitative data. They are popular because they are inexpensive to administer, and can be analysed rapidly [49]. Despite being popular, designing a good questionnaire is a difficult and complicated process [28, 49, 111]. A good

questionnaire is one that is both reliable and valid. While standard questionnaires whose reliability and validity has been verified exist, they are inappropriate in certain scenarios, forcing researchers to 'improvise'. A criticism of these improvised questionnaires is that no one knows whether they are reliable and valid. More so, they are unavailable to researchers other than those using them. In this research, we used standard ASQ questionnaires to capture data during an experiment and an After Experiment Questionnaire where participants indicated their preferences for the different methods.

3.3.2 Interviews

Interviews are qualitative and descriptive in nature [49, 111] and may take one of three forms; unstructured, structured, or semi-structured. Semi-structured interviews are particularly useful because they give room to an interviewer to follow on any interesting leads that a participant may bring out and they also allow one to exercise control and direct the interview. Unlike questionnaires that may be collected and analysed later, interviews provide immediate feedback which may influence future interviews or design of same study. For example, first few participants may complain about a particular feature of an interface (which may not be a variable in the study but may be distracting if they focus on it). An experimenter may decide to deal with that issue so that future participants are not distracted by it. In our studies, interview were used for two reasons. First, they were used to gather data in addition to that captured through questionnaires. Second, they were used as a validation tool for the data gathered through other means.

3.3.3 Observation

Interviews have greatest value when conducted in conjunction with some form of observation [49]. Observation may reveal unexpected behaviour such as an unanticipated sequence of actions, for example. Such unexpected behaviours may enrich the data on users' interactions with a system. Recording observations may take the form of note taking or video. Even though a user may be aware that interactions are being watched, note taking may be less obtrusive than video recording [111]. In addition, it may be quicker to analyse notes compared to video. Note taking, however, is not convenient when a study involves a group of participants acting simultaneously or where an observer is likely to envade a participant's personal space. In such cases, video recording may be more appropriate. Moreover, there may be actions that may not be regarded as important at the time

participants are being observed but may later be found important — only video can capture such actions. In this research, a mix of video and note taking was used. We used note taking in single-user scenarios and video recording in group scenarios because of the difficulty in observing a group of participants.

3.4 Research approach

3.4.1 Laboratory experiments

The aim of this research was to investigate the security and usability of HISPs, specifically empirical channels. To measure performance of empirical channels requires quantitative data such as completion times and completion rates. To identify and understand factors that affect security and usability of a system requires a qualitative approach where participants can provide feedback on their interactions with a system. Laboratory experiments were, therefore, employed (to capture performance data) together with interviews, questionnaires, and observation. To ensure that studies were conducted according to initial objectives and that data was correctly captured checked for inconsistencies, the following were followed:

3.4.1.1 Pilot studies

A pilot study is a small trial run of the main study to evaluate the experiment itself. It involves recruiting a small number of participants, with similar characteristics as target population, and conducting the test [91, 95, 111, 112]. We used pilot studies to identify and clarify ambiguities in test instructions and questionnaires, ensure hardware/software worked as expected, and data was captured correctly.

3.4.1.2 Test plan

When conducting a user experiment, it is crucial to outline the main tasks that a researcher needs to undertake before and during a study [95]. It helps a researcher to articulate planning decisions [27] and adhere to initial objectives of the study and also to keep track of the tasks that have been com-

pleted and those that are still pending. Before designing and conducting a user experiment, a detailed plan must be developed stipulating required equipment and tools (such as software, questionnaires, instructions, devices *etc.*), participant recruitment procedures, test design and administration, and data collection and analysis procedures. An example of a test plan used for the group association scenario study is given in Appendix D.

3.4.1.3 Triangulation

Triangulation entails using more than one method (method triangulation) to collect data or more than one analysis technique on the same data to arrive at a conclusion [69, 111]. In conducting user experiments of HISPs, different methods of collecting same data were used. Dependency on a single method may be unreliable in some cases. For example, a participant may struggle in performing a task but when interviewed claims the task to be 'simple and straightforward'. Sometimes it was essential to follow up on participants when conflicting information was discovered after the study.

3.5 Validity of research

Usability experiments have been criticised for being artificial [95] because they are conducted in a laboratory environment, possibly unfamiliar to participants, using representative (as opposed to real) tasks. In addition, the laboratory environment may be different from the environment in which the studied system may work. For example, use of mobile devices may occur in any place imaginable and it is not possible to model each of these environments in a laboratory setting.

Studies to evaluate the effectiveness of user experiments compared to field studies in mobile device applications have been conducted. The results of these studies, however, are mixed. For example, [9, 51, 59] concluded that the benefits of conducting field studies (compared to user experiments) are not worth the cost. They argued, based on their results, that the differences between results from a user experiment and those from a field study are not significant. On the other hand, [26, 81] argued that there are significant differences between usability experiments and field studies. Despite this finding, Duh *et al.* [26] could not conclude whether a field study is superior to a laboratory one while Nielsen *et al.* [81] concluded that field studies are worthy the effort because they reveal significantly more usability problems, interaction styles, and cognitive load than laboratory studies.

Laboratory experiments, however, produce meaningful and valid results when strict guidelines for designing and conducting them are adhered to. A good laboratory study pays attention to the design and conducting of an experiment, creating realistic and representative tasks, and using participants who are a true representative of the target population [89,95,111,115]. In addition, attention must be paid to recruitment of participants, test orders (to counter learning effects), instructions, and tools (such as questionnaires) for recording data. These measures are crucial to ensuring that results are not biased, they represent actual performance of participants, and are reliable and valid. Moreover, laboratory experiments have been applied in other studies and have contributed to the scholarly literature (e.g. [129]). This research closely followed the rigour of good experimental design, as discussed above, using questionnaires whose validity and reliability have been proved. Attention was paid to:

1. Participant recruitment: participants were recruited through an online advertisement to ensure that samples were not biased to university/college students or a particular age groups.

2. Experimental design: when a between-subject was used, participants were randomly assigned to different groups to ensure that no group had advantage over another. In a within-subject design, tasks were either randomised or counterbalanced to minimise learning effect.

3. Consistency: participants were treated equally to ensure that results were not influenced by differing instructions.

4. Triangulation: inconsistencies in the data can be detected before a participant leaves the laboratory. For example, if one mentions that a particular method is the best but have indicated a different one on the questionnaire we can have it corrected before it gets to analysis stage.

3.6 Conclusion

In order to evaluate the effectiveness of HISPs in different scenarios and contexts, we employed laboratory experiments to gather quantitative and qualitative data. Quantitative data was valuable in assessing the performance of HISPs while qualitative data was crucial to understanding users' views and expectations. Despite criticisms of laboratory studies, careful design and adherence to good experimental design increases both internal and external validity.

Chapter 4

HISPs: Security and Usability in Single-User Scenarios

4.1 Introduction

In this chapter, we discuss device association in single-user scenarios and present results of a user experiment evaluating the effectiness of empirical channels in these scenarios. The chapter is organised as follows: in Section 4.2 we discuss device associations in single-user scenarios and follow on with presentation of the user experiment in Section 4.3. We present the results of the study in Section 4.4 and discuss the effectiveness of empirical channels in Section 4.5. In Section 4.6, we draw attention to the factors affecting effectiveness of empirical channels in single-user scenarios. We summarise and conclude this chapter in Section 4.7.

4.2 Device association in single-user scenarios

4.2.1 Association scenarios

Secure single-user device associations may be characterised into 4 scenarios as summarised in Figure 4.1. The arrows in the figure indicate the direction of authentication.

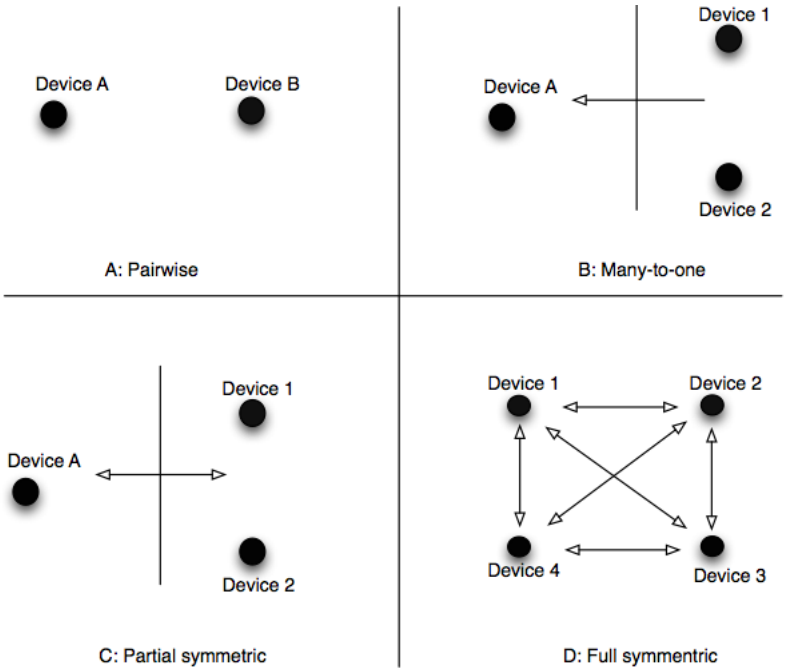

Figure 4.1: Single-user scenarios

1. **Pairwise**

   In a pairwise scenario, devices may act as peers or one may act as a more trustworthy device. For example, a user pairing two personal devices such as mobile phones does not care which device authenticates the other (peers) whereas pairing between a mobile device and a WiFi access point may dictate that the former authenticates the latter. When only one device carries out authentication, the process is known as asymmetric (or one-way) as opposed to symmetric (or mutual) where both devices authenticate each other.

2. **Many-to-one**

   The second single-user device association scenario is many-to-one — an example of asymmetric authentication. In this scenario, one device authenticates several. This is appropriate in scenarios where one wishes to create a local area network among devices. One example of such a scenario is joining two or more devices to a WiFi Access Point (AP) simultaneously. The access point needs to authenticate all devices being introduced to the network. In this example, the benefit of authenticating devices simultaneously compared to authenticating one at a time is obvious: a user goes through the association process once.

3. **Partial symmetric authentication**

   Partial symmetric involves a single device authenticating multiple devices while each of the

other devices authenticates the former. Devices authenticated by a single device authenticate the latter by checking that the digest they have matches that of the single device. The single device authenticates other devices by checking that each of the other devices has the correct digest. To illustrate partial symmetric, consider a WiFi example above. After an initial exchange of information using HISPs, devices joining a network authenticate the AP after the AP authenticates each of the devices. In this example, devices authenticate the AP to ensure that they are connecting to the desired network. On the other hand, the AP has to ensure that only devices that the owner/legitimate user intends to join are allowed access to the network.

4. **Symmetric authentication**

   The previous scenarios require authentication in one direction or from a group of devices to one and vice versa. There are scenarios, however, where each device participating in the association needs to authenticate every other device. This may be viewed as a many-to-one scenario repeated for each device involved. Each device authenticates other devices by verifying that its digest matches every other device's. An example of full symmetric is a ticketing system where a personal device authenticates a vending machine by verifying that digests match. The vending machine releases a ticket only when a user of the personal device indicates that the device accepted the connection.

4.2.2 Definitions

The following definitions will be used throughout this chapter.

- *Method*: a method is a way of comparing or transferring digests independently computed by two or more devices.
- *Representation*: a format in which a digest is displayed to a user. This includes numeric and alphanumeric strings, words, barcodes, images, etc.
- *Method-representation*: Some methods can use more than one representation and, hence, method-representation will be used to associate a particular representation with a particular method. For example, *manual comparison*-alphanumeric refers to *manual comparison* method using alphanumeric strings.

4.3 Empirical evaluation of single-user association scenarios

Given the range of proposed empirical channels, an evaluation of their effectiveness — in terms of both security and usability — is crucial to gaining an understanding of how these methods can be improved, adapted or modified to achieve optimum effective security. Moreover, understanding how users interact and identifying fundamental factors that affect usability and security of HISPs is crucial to designing empirical channels that fit human and contextual needs.

4.3.1 Participants

Participants in the study were respondents to an online advertisement. This mode of recruitment was employed to provide a diverse sample. A total of 30 paid participants were recruited. Table 4.1 summarises participants' demographics.

Gender	Male: 47% Female: 53%
Age	18 - 25 40% 26 - 35 27% 36 - 45 13% 46 - 55 3% 56 - 65 13% 66 - 75 4%
Education	High School: 27% College: 27% Graduate: 26% Postgraduate: 20%

Table 4.1: Participant demographics

4.3.2 Material and apparatus

In real world device associations, devices are associated in order to achieve a specific goal such as exchanging files. Therefore, the process of association is merely a means to an end. As earlier argued in this thesis, user experiments must present realistic and representative tasks to participants so as to evaluate tasks of interest in near real world conditions to produce results that are useful beyond laboratory settings. Taking this into account a simulated *Peer-to-Peer (P2P) payment system*, in which one uses a personal device to make an electronic payment to another, was used as a primary

task. The participants' goal was thus to carry out a successful payment transaction rather than merely associating devices.

The P2P payment system was developed using Java 2 Micro Edition (J2ME) [119] for portability and Bluetooth support on mobile devices. The study was conducted using two mobile phones; Nokia N95 with 2.6 inch screen, 240x320 pixels resolution, 332MHz CPU, 160MB memory capacity, and running Symbian OS v9.2 and Nokia N73 with a 2.4 inch screen, 240x320 pixels resolution, 220 MHz CPU, 42MB memory capacity, and running S60 operating system. Both devices support Mobile Information Device Profile version 2.0 (MIDP 2.0), a specification for use of Java on embedded devices [120], and both had cameras.

To handle connection, data transmission, and logging of participant activities, the ISUT tool (see Appendix C) was used. One device was designated as controller device to generate tests and keep a log. The following participants' events were logged:

- time to complete the association process: this is the time between the display of a digest on one or both devices and when a participant completes comparison or copying. Since performing the payment was not of interest to the experiment, no participants' actions were logged for it.
- number of security failures: security failures are those failures that may result in a user associating a personal device with an unintended one. The controller device generated tests that specifically tested participants on whether they could detect any discrepancies in the digest displayed and, after doing so, if they could take a correct action.
- number of non-security failures: non-security failures are those failures that result in a user or device aborting a connection in the absence of a digest mismatch. For example in *manual copying and entering*, both devices may have the same digest but a user mistyping it into one device leads to a failed device association.

In addition to quantitative data logged by the controller device, three types of questionnaires were used (provided in Appendix E) together with a brief semi-structured interview. An Enrolment Questionnaire (EQ) provided information on participants' demographic data, while After Scenario Questionnaires (ASQ) [66] provided subjective data on three main components for each method:

- satisfaction with the ease with which a method-representation was used,

- satisfaction with the amount of time spent on a method-representation,
- whether participants felt they could effectively carry out a transaction using a particular method-representation.

ASQ is a Likert-type or summative rating scale consisting of 3 questions with answers based on a scale of 1 to 7, with 1 corresponding to *strongly agree* and 7 to *strongly disagree*. Many rating scales use a scale of 5 intervals rather than 7. However, it has been found that reliability of rating scales increases with the number of items and also the number of interval points for each item, and levels off at about 7 intervals with no significant increase after 11 intervals [68], hence the use of a 7 interval scale.

An End of Experiment (EoE) questionnaire gave participants an opportunity to identify methods they felt were easy, difficult, and which ones they preferred or they would avoid. Interviews gathered participants' views and comments on what they felt about method-representations and what their experience in general was. Moreover, interviews were also used to double check participants' responses to questionnaire questions. In order to maintain consistency across participants but also be flexible enough to discuss issues that were to be raised, the interview was semi-structured.

4.3.3 Methods tested

Given the number of proposed types of empirical channel, we eliminated some based on a number of criteria. First, all methods that require specialised hardware such as accelerometers and laser light beamers and readers were eliminated. This is because, at the time of the study, such hardware was neither default nor common on mobile devices. Second, methods that provide no guarantees to users or are subject to MiTM attacks were also eliminated. These include Integrity regions [15] and Infrared [7, 29]. Methods based on integrity regions rely on devices measuring the distance between them and also on users ensuring that there are no other devices, other than those expected, within that distance. This method may work if users have an effective way of detecting presence of malicious devices and if such devices cannot present themselves as being closer than they physically are. Infrared is legacy technology and not a default feature in state-of-the-art devices. It is also subject to MiTM attacks [114]. Having carefully eliminated some methods, the following were used in the study: manual comparison, compare and select, manual copying and entering, and barcode.

4.3.3.1 Manual comparison (MC)

With this method, a user compares strings, sounds, or images displayed on both devices and presses a button to indicate a match or disparity. In the study, participants were required to press a button on their personal device only. Several digest representations were used with this method: numeric, alphanumeric, numeric & sound, alphanumeric & sound, words, sentences, melodies, names of countries/cities, and images. In presenting these to participants, three scenarios were used where possible:

1. Matching one where the two values matched: In this scenario, both devices displayed the same digest and a participant was expected to indicate a match on a personal device.
2. Non-matching where the two digests were significantly different: A participant will be presented with digest that has no similarities between them and it was expected that the association would be rejected.
3. Near-matching where the two digests were different but nearly matching: Near matching means that the strings differ by a single digit for numeric, a single character for alphanumeric, a single word for words and sentences, and a single country/city name for countries/cities.

This was done to draw attention to the potentially problematic near match case, since it was suspected that these might cause more security failures compared to other scenarios. Each of the representations used for this method is discussed below.

- **Numeric**: Each device displayed a 6 digit value and a participant compared the values on both devices with the instruction: *"Compare the two numbers. Are they DIFFERENT?"*. The participant then pressed *'SAME'* or *'DIFFERENT'* depending on whether the valued were perceived to be same or different respectively. This wording is the same as that recommended by Uzun *et al.* [123], which they found improved the usability of the method. The values were displayed in two blocks of three digits. This separation was used when displaying numeric and alphanumeric values with a view that it might help users to split the comparison into two rather than the full string at once.
- **Alphanumeric**: With alphanumeric characters, a 32 character set was used. This includes all the numeric characters (0-9) and all characters in the English alphabet with the exception

of ‘I’, ‘O’, ‘Q’, and ‘U’, since these could cause confusion. For example, the letter ‘O’ could be confused with the number 0 or the letter ‘Q’, ‘I’ with 1, and ‘U’ with ‘V’. Thus each character in the set could represent 5 bits of a digest, and thus the complete string represented 30 bits. Despite alphanumeric representing more bits than numeric, it was felt necessary to display both types of strings in equal length to the user for comparative analysis.

- **Words**: Words were constructed from a dictionary of 1024 English verbs. Each word represented 10 bits and a set of four words was used in the study. For the experiment, a digest was calculated from a randomly generated string and each segment of 10 bits was used to look up a word in the dictionary. With 10 bits per word, it would have been sufficient to compare two words. However, some mobile phones may use dictionaries that are much smaller than this for reasons of memory, and two or three words may not be sufficient for a 20 bit digest in such cases.

- **Sentences**: It has been suggested that users find it easier to deal with meaningful strings such as words and sentences than meaningless ones like alphanumeric [39]. Sentences were generated from the digest based on MadLib [39] puzzles. A total of 32 sentences were stored which had at most 7 words of which 3 were missing. During the test, a sentence was selected and the missing words were queried from the dictionary using values from the digest.

- **Images**: People have been found to be better at dealing with images than dealing with strings [76] and proposals for users comparing images in HISPs have been made based on this finding. In the study, images were stored locally on the devices and only two scenarios were tested: matching and non-matching images. It was difficult to simulate near matching images since this is subjective as opposed to other representations such as numeric ones. A participant compared images displayed on both devices and pressed ‘SAME’ or ‘DIFFERENT’ on a personal device. Figure 4.2 shows the images used in the study.

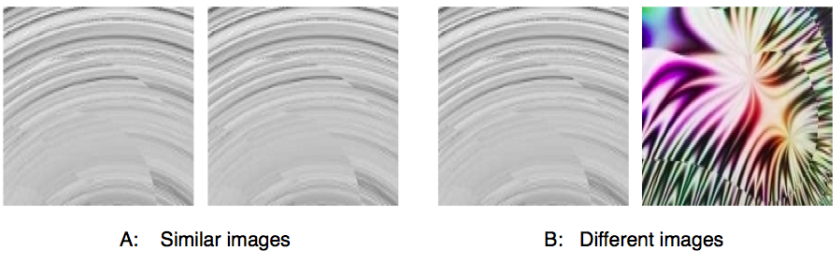

Figure 4.2: Images used in the study

- **Melodies**: While all the above representations try to utilise human visual abilities, melodies try to utilise the ability to distinguish two audio sequences. In the study, melodies were generated by playing a note based on each digit of the 6 digit digest. In this test, only two scenarios were tested: matching and non-matching melodies. Participants played one melody after another, but it was also possible to play melodies on both devices simultaneously by pressing buttons on both devices.

- **Sound**: Another variation is to utilise both the visual and audio abilities of users by having a digest displayed on one device while the other device reads out its value. A user listens to the string read on one device and compares it with the one displayed on the other device. Only numeric and alphanumeric strings were tested in this manner for three variations: matching, near-matching and non-matching. Words and sentences were not used for this method as this would require text to speech capabilities which most mobile phones do not possess.

4.3.3.2 Compare and select (CS)

Unlike *manual comparison*, a string was displayed on one device and four strings of the same format were displayed on the personal device. Due to limitations on the size of the displays, only numeric and alphanumeric values were used. For each of these, 4 scenarios were constructed:

- *Match* where one of the strings among the 4 displayed on the personal device matched the string displayed on the other device,

- *Match and near-match* where one string matched and at least one other is near-matching the string displayed on the non-personal device,

- *Different* where none of the 4 strings displayed on the personal device matches one on the other device,

- *Different and near-matching* where there was no string among the 4 matching the one displayed on the non-personal device but at least one is near-matching.

Participants were asked to choose a string from the personal device that matched a string on the other device and press “MATCH” otherwise press “NOT FOUND” if no matching string was present.

4.3.3.3 Manual copying and entering (MCE)

With this method, one device displayed a string while the other asked a participant to enter the same string. The device then compared the entered string with a locally generated one. If the two strings matched, the device accepted otherwise rejected the association. The major difference between *manual copying and entering* and the other methods above is that a user does not do the comparison but rather only enters what is displayed on the other device. This reduces security failures to a single random guess, and makes it hard to simulate them. It was, therefore, only possible to capture non-security failures on this method. Due to limitations of the keypads on mobile phones, only numeric and alphanumeric strings were used for this method.

4.3.3.4 Barcode

This method differs completely from all the other methods in that, first, it requires a mobile phone with a camera and, second, a digest is not displayed in a human readable format. One device encodes a digest into a 2-dimensional barcode and the other is required to read it off the screen. An open source barcode reader software (ZXing [87]) was used to decode a barcode displayed on another mobile device. In the study, one device displayed a qrcode barcode [117] while the other automatically activated the camera function and asked a participant to point the camera at the other device and take a snapshot of the displayed image. The image used was 162x162 pixels and contained all letters of the English alphabet.

This method was included in the study because of its ability to accommodate more bits than the other methods and it also presented a complete diversion from all the other methods in that a user only needs to point a personal device at an image displayed by another device and press the "CAPTURE" button. The image size used in the study was chosen to be big enough for easy focusing while the size of digest encoded in it was consistent with the proposal that a barcode may be required to contain a digest of a device's public key and other information such as name and address of device [72].

4.3.4 Participant tasks

The study was conducted in a laboratory environment. Upon arrival, a participant was taken to a room where the study was conducted. A summary of what was to be done was given verbally, and, where this had not been received in advance, participants were asked to fill in an EQ. The participant then moved to a desk where s/he was provided with an instruction sheet, ASQ questionnaires, and two mobile phones. The instructions were provided in written form to achieve consistency across all participants.

Part of the instructions included informing a participant about which of the two devices was to be assumed a personal device (in this case Nokia N73) and which one was the payee's (for the P2P system). For most of the methods, participants interacted only with the personal device while only observing the payee's, except for a few cases where a participant was required to press a button on the payee's device. However, participants had the freedom of holding in their hands the payee's device for their convenience.

Each participant carried out 33 tests, aimed at testing 14 different methods and method-representations. Since certain methods and method-representations require a user to decide whether digests match or not, additional scenarios were used to test a user's ability to correctly identify a match or lack of it (where appropriate, details of these additional scenarios are described in detail with each method).

The study used a repeated measure design and the system presented these scenarios in a random order to increase internal validity [74] by minimising learning effects. Each participant completed 14 After Scenario Questionnaires (ASQ), for each of the 14 different methods and method-representations. After completing all 33 tests, a participant filled in an EoE questionnaire and was interviewed. Interviews were recorded. Participants required between 35 and 60 minutes to complete the study.

4.3.5 Hypotheses tested

Two hypotheses were formulated:

- Hypothesis 1: there is no statistically significant difference in the dependent variable completion times for the within-subject factor method-representation.

- Hypothesis 2: there is no statistically significant difference in the dependent variable completion time for the between-subject factor age (18-35 and 36+) across method-representations.

4.4 Results

Each participant's actions generated a log file for completion times and failures for all the method-representations and their variants. This constituted the main source of objective data. In addition, each participant completed 14 ASQs, an EoE, and an EQ. This data was later compiled into Microsoft Excel worksheet in readiness for analysis using statistical tools provided by various packages. An audio recording of the interview for each participant was transcribed for later analysis.

4.4.1 Objective results

This data revealed failures that various method-representations are prone to. The study was a repeated measure in which each participant was tested on all the scenarios. For 30 participants with 33 scenarios, a total of 990 data items were available for analysis.

4.4.1.1 Manual comparison

Table 4.2 shows a summary of results for *manual comparison*. For each representation, the percentage of security and non-security failures according to three categories simulated in the study is shown. For images and melodies, no simulation was done for near-matching as explained above and are indicated by (-) in the table. For each scenario, we only show the type of failures that are possible. For example, for a matching scenario, only non-security failures are possible.

In *manual comparison*, security failures are only possible in non-matching and near-matching scenarios while non-security failures are only possible in matching strings. The table shows that non-security failures ranged from 0% for numeric & sound to 36.7% for melodies while security failures ranged from 0% to 13.3%. It is worth noting that security failures in this method are too high for a security application and they are just as likely to happen in a non-matching scenario as in a near-matching one.

	Matching %	Non matching %	Near matching %	Total	
	Non-security(NS)	Security(S)	Security(S)	S	NS
Numeric	3.3	0	0	0	3.3
Alphanumeric	16.7	3.3	10	13.3	16.7
Words	16.7	3.3	0	3.3	16.7
Sentences	16.7	0	0	0	16.7
Images	3.3	0	-	0	3.3
Melodies	36.7	6.7	-	6.7	36.7
Numeric & sound	0	0	3.3	3.3	0
Alphanumeric & sound	20	0	3.3	3.3	20
Country/City names	3.3	0	0	0	3.3

Table 4.2: Manual comparison: Security and non-security failures (in percent).

Table 4.3 shows completion times for each representation. The results show that numeric and alphanumeric had the shortest completion times while melodies had the longest. The table also shows that there were a number of outliers in completion times for each representation. For example, while numeric had a maximum completion time of 62 seconds, the mean was only 6 and the mode 3. These outliers could be explained in terms of participants getting distracted as they carried out a task. Outliers, however, were few in the data and their influence on the calculated means was minimal.

	Time - seconds					
	Mean	Mode	Median	SD	Min	Max
Numeric	6	3	5	7	1	62
Alphanumeric	6	2	5	4	1	25
Words	7	6	6	4	1	20
Sentences	11	6	8	10	2	56
Images	8	2	5	12	1	85
Melodies	24	15	20	16	4	88
Numeric & sound	14	10	11	11	4	76
Alphanumeric & sound	12	10	10	7	4	48
Country/city names	9	4	8	5	2	26

Table 4.3: Manual comparison: Completion times

In order to evaluate the significance of the differences in the dependant variable (time) with within-subjects factor method-representation and between-subjects factor (individual differences), a one-way repeated measure analysis of variance (ANOVA) was performed. The results showed statistical significance in both factors, with $F(8, 472) = 1.776$ and $p = .0000$ for within-subjects factor and $F(59, 472) = 23.393$) with $p = .0007$ for between-subjects factor. The variations in time is apparent from the means in Table 4.3; some methods took longer than others. The variation in the between-subjects factor could be attributed to the observation that younger participants performed better in terms of completion times than older ones.

4.4.1.2 Compare and select

With *compare and select*, four scenarios were simulated as discussed above giving a total of 120 data items(for 30 participants) to analyse. Half of these had matching strings while the other half had non-matching strings. With this method, a user indicating that there is no match when there is (or selecting a value other than the actual digest) results in a non-security failure. However, when a user selects a value and indicates that it is a match when there is none has two possible outcomes; either non-security or security failure. It was, however, decided to take the worst case scenario and regarded all failures resulting from selecting a non-matching value as security failures even though there is a chance that they might not be.

		Numeric	Alphanumeric
Matching	non-security	0	13.3
Non Matching	security	0	6.7
Near matching	security	10	6.7
Matching and near matching	non-security	10	16.7
Total	security	10	13.4
	non-security	10	30

Table 4.4: Compare and select: Security and non-security failures

Table 4.4 shows a summary of failures for *compare and select*. Alphanumeric had a higher rate of both security and non-security failures. Despite the differences in both types of failures, there was no significant difference in terms of completion times as summarised in Figure 4.3. A statistical test using one-way repeated ANOVA on completion times showed that the result was significant for the between-subjects factor with $p = 0.0000$ ($F(119, 119) = 2.207$) while it was not significant for the within-subjects factor with $p = 0.9255$ ($F(1, 119) = 0.009$). The significance of the between-subjects factor could be explained in terms of the differences between younger and older participants and also participants' familiarity with the models of the mobile phones used in the study.

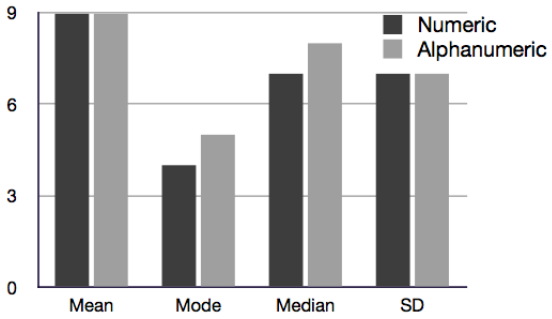

Figure 4.3: Compare and select: Completion times

4.4.1.3 Manual copying and entering

Manual copying and entering had no variants; resulting in 30 data items to analyse. It is apparent from Figure 4.4 that participants took longer to enter alphanumeric compared to numeric values. Entering alphanumeric also had more failures than numeric. Of these failures, however, 75% of numeric were as a result of the confusion between copying and typing the displayed 6 digit digest and typing a four digit PIN for the payment transaction while 43% of alphanumeric were as a result of unfamiliarity with the model of the mobile phone.

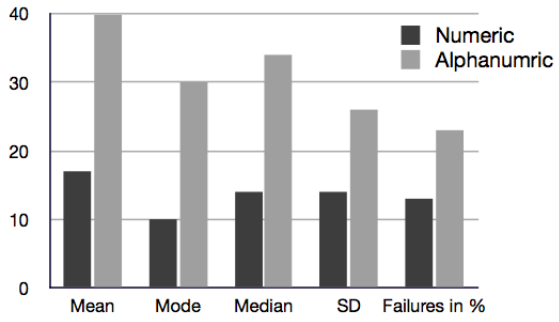

Figure 4.4: Manual copying and entering: Completion times and failures

A one-way repeated measure ANOVA on completion times showed that there was no significance for the between-subjects factor with $F(29,29) = .774$ and $p = .7531$ while the results were significant for the within-subject factor with $F\ (1, 29) = 15.78$ and $p = .0004$.

4.4.1.4 Barcode

The results of the *barcode* method indicate that participants spent a significant amount of time focusing the camera on the displayed image. It also shows a high percentage of non-security failures. These failures were as a result of not taking a clear shot of the image resulting in a failure by the decoding algorithm to reconstruct the image in order to decode it. Part of the problem is failure by participants to get a clear shot of the image, but more so was the implementation itself. The software worked quite robustly when an image was displayed on a laptop screen. This was not the case, however, when the same image was displayed on the phones mainly because of the size and resolution of the mobile phone screen.

A source of concern for this method is not the number of failures (the results could be different with a more robust implementation) but what participants thought of it. Participants seemed to be

Time - seconds						Failures %
Mean	Mode	Median	SD	Min	Max	Non-security
37	33	33	14	15	79	53

Table 4.5: Barcode: Completion times and failures

confused that the method, unlike other methods, was not intuitive. They could not figure out the purpose for taking a snapshot of an image displayed on another mobile phone.

Security failures can only be as a result of taking a snapshot of an unintended barcode. For example, a barcode displayed on a bogus cash machine or a fixed barcode on an access point that has been replaced by another from an intruder. However, in this study, such scenarios were not covered.

4.4.2 Subjective results

Objective results above show each method-representation in terms of two dependant variables; errors and time. There was a need, however, to gather participants' views on each method-representation in view of the possibility that despite a method-representation performing objectively well in terms of the two dependant variables, participants may not necessarily favour such.

4.4.2.1 ASQ (rating scores)

Participants gave their rating scores to each of the three items on the ASQ. These ratings were summed and averaged to calculate each participant's single score for each method-representation. Raw data was inverted before presentation so that a high score represents a high agreement from the participant rather than what was in the questionnaire where a high score indicated a disagreement (low score) from the participant.

The results, summarised in Figure 4.5, show that most methods had a score higher than 5 except melodies, *barcode*, and *manual copying and entering*-alphanumeric. *Manual comparison*-numeric and *compare and select*-numeric had the highest scores of 6.3 followed by *manual copying and entering*-numeric at 6.1.

On a scale of 7 intervals, a method-representation was regarded as usable if it had a score of 5.6 or more. This is based on the results of [84] which indicate that a system is usable if it has a score of 4

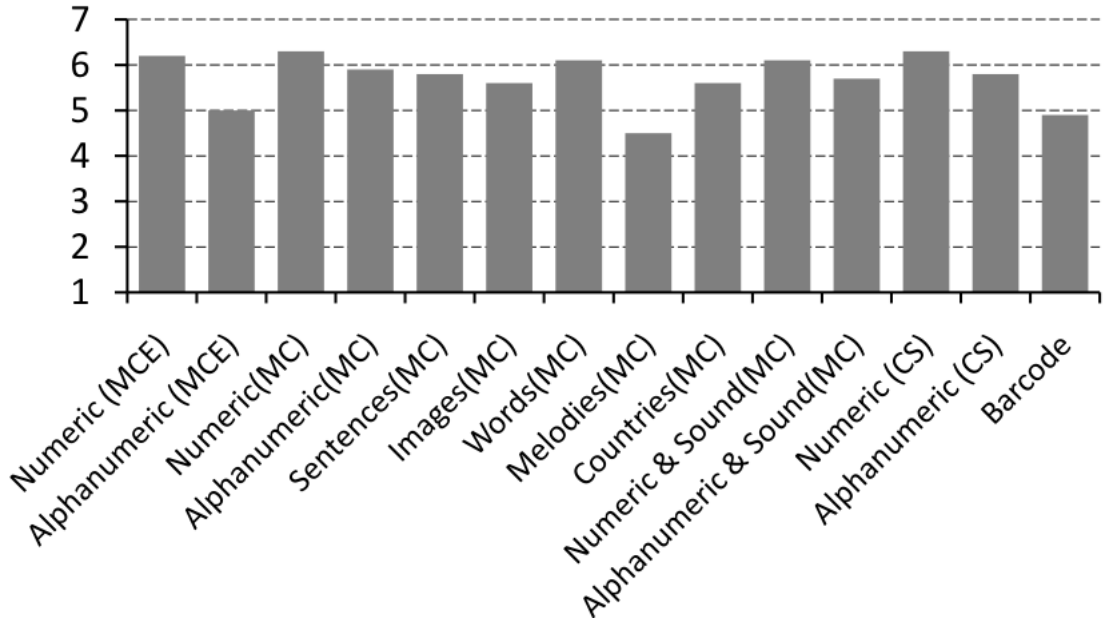

Figure 4.5: Participants ASQ scores

or more on a scale of 1 to 5 or a score of 5.6 on a scale of 1 to 7. Based on this result, then *manual copying and entering*-alphanumeric, melodies, alphanumeric & sound, and barcode are less usable.

4.4.2.2 Preferred methods

In addition to assigning rating scores to each method-representation, participants were asked to indicate all method-representations that they felt were easy to use and also to indicate their preferred one. The results are summarised in Figure 4.6.

4.4.2.3 Unpreferred methods

Despite participants indicating which method-representations they felt were easy to use, it was also necessary for participants to explicitly indicate which method-representations they felt were difficult and which one they would avoid, given a choice. The results, summarised in Figure 4.7, correlate with those in Figure 4.6; melodies had the lowest score in Figure 4.6 but the highest score in Figure 4.7. Generally, method-representations that had high scores in Figure 4.6 had low scores in Figure 4.7 and vice versa indicating that the results correlate.

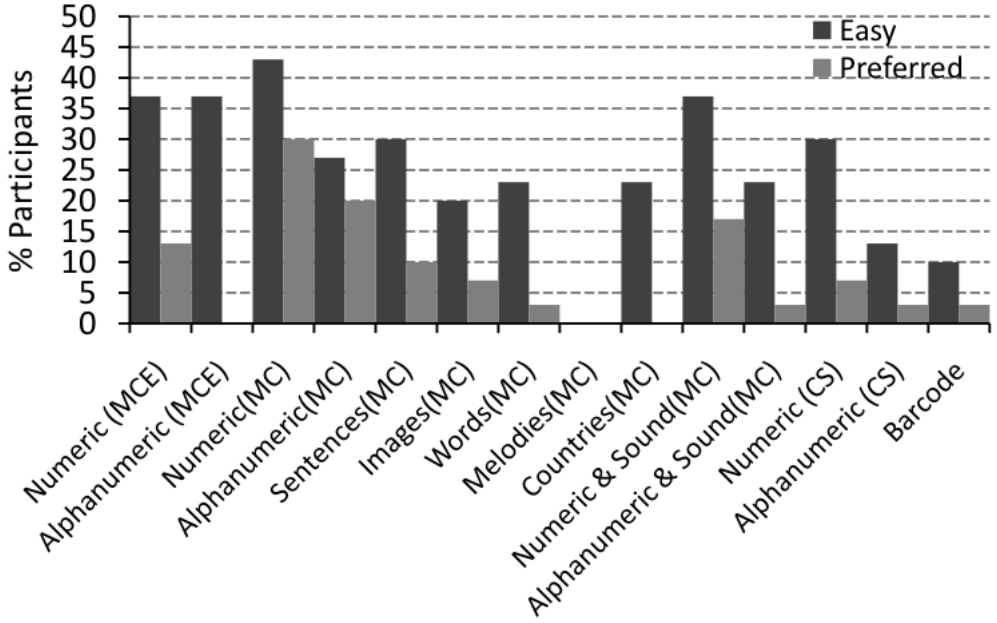

Figure 4.6: Participants' choices: Easy and preferred methods

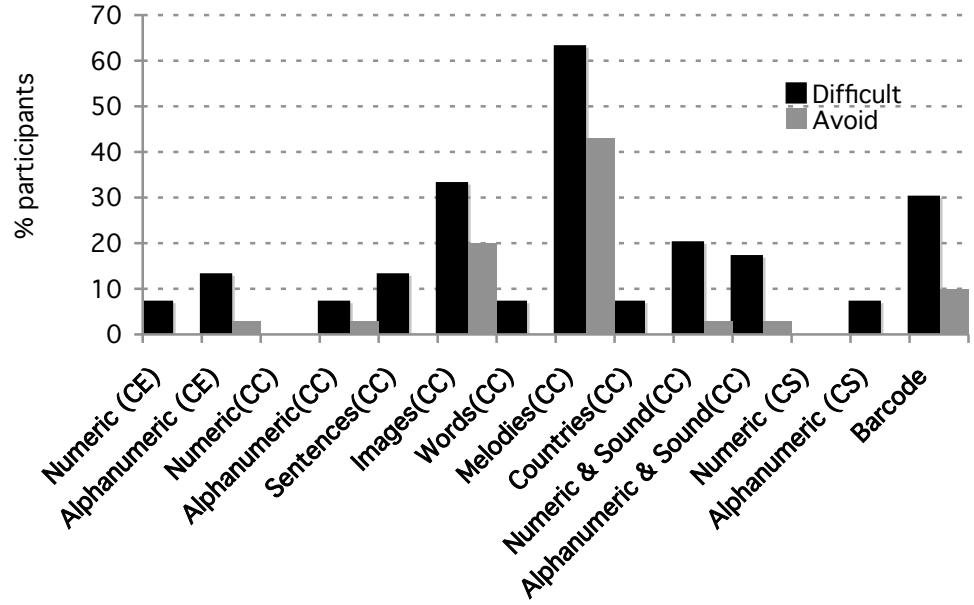

Figure 4.7: Participants' choices: Difficult and unpreferred.

4.4.3 Hypotheses validation

To understand the differences between method-representations, hypotheses were validated according to each method. There was no validation across methods because of the differences in the number

of data points. Moreover, because the barcode has only one representation, no hypothesis validation was performed for this method.

- Hypothesis 1

    1. Manual comparison — the ANOVA test shows that there is a statistically significant difference in completion times for the within-subject factor method-representation (F(8,472) = 1.776, p = .0000).
    2. Compare and select — the ANOVA test shows that there is a statistically significant difference in completion times between the within-subject factor method-representation with F(119,119)= 2.207 and p = .0000.
    3. Manual copying and entering — the ANOVA test shows that there is no statistically significant difference in completion times between the within-subject factor method-representation with F(29,29) = .774 and p = .7531.

- Hypothesis 2

    1. Manual comparison — the ANOVA test shows that there is a statically significant difference in completion time for the between-subject factor age across method-representations (F(59,472) = 23.393, p = .0007).
    2. Compare and select — the ANOVA test shows that there is no statistically significant difference in completion times for the between-subject factor age across method-representations with F(1,119)= .009 and p = .9255.
    3. Manual copying and entering — the ANOVA test shows that there is a statistically significant difference in completion times for the between-subject factor method across method-representations with F(1,29) = 15.78 and p = .0004.

4.5 Effectiveness of empirical channels in single-user scenarios

To analyse the relative performance of each method-representation in terms of all the parameters measured in the study, it was necessary to have a single overall score for each method-representation.

While trying to find a method to quantify the measurements and calculate a single score, only Single Usability Metric (SUM) [99] was found to suit this purpose. This is a method of calculating a single usability score by standardising raw data for each of the parameters under the study. Standardising scores allows comparison of values of different variables regardless of their original unit.

To calculate standardised scores, specification (limit) scores [99] must be set for some of the parameters measured. Specification scores are the acceptable values below which a method is regarded as less usable. There are various ways by which specification scores may be determined including using an existing system, a prototype, a user's earlier performance, or an absolute scale [47]. However, because of lack of previous data on task completion times for the methods tested, twice the modal score was used as the specification score. Unlike mean, mode is not affected by outliers. In addition, it represents a value that most participants achieved as opposed to median which represents a middle number in an ordered list even if it is not the most frequent. On a rating scale of 1 to 7 (7 indicating best), an average score of 5.6 is the minimum for a product that is deemed usable [84], and as such this value was used as a specification score for the ASQ scores.

With specification scores for completion times and rating scores determined, SUM scores were calculated for each method-representation. Table 4.6 shows the ranking of the method-representations in order of their single usability scores. The scores shown under each column are quality [99] scores rather than defects. For example, for *manual comparison*-melodies the table shows that despite a potential for non-security failures, 63.4% of cases resulted in a successful accurate comparison.

	NS	Time	ASQ	PRD	Easy	SUM
Numeric(MC)	98.4	50.6	74.8	30	100	73.7
Alphanumeric(MC)	91.7	83.1	58.7	20	93	72.5
Words(MC)	91.7	88.1	63.7	3	93	70.6
Numeric & sound(MC)	100	63	68.9	17	80	69.2
Numeric(MCE)	97.5	59	68.1	13	93	69
Numeric(CS)	95.8	44.6	80.1	7	100	68.3
Alphanumeric & sound(MC)	90	84.8	52.7	3	83	65.8
Alphanumeric(CS)	98	56.4	55.4	3	93	64.2
Sentences(MC)	91.7	54.9	55.5	10	87	62.9
Alphanumeric(MCE)	87	78	34.5	0	87	60.4
Countries(MC)	96.7	41	50	0	93	59.1
Images(MC)	96.7	37.1	50	7	67	54.3
Barcode	47	97.2	33.6	3	70	53
Melodies(MC)	63.4	63.8	27.2	0	37	40.7

Table 4.6: Ranking based on Single Usability Metric (SUM) Scores (in %): NS = Non-security failures, PRD = Preferred

Based on the SUM scores, Table 4.6 shows that *manual comparison* (numeric, alphanumeric and words, numeric & sound) ranked top ranging from 69.2% to 73.7%, followed by *manual copying and entering* (numeric) at 69%, *compare and select* (numeric) at 68.3% and finally *barcode* at 53%.

Among the methods tested for *manual comparison*, the following had low ratings; melodies (40.7%), images (54.3%), country/city names (59.1%), and alphanumeric & sound (65.8%). Sixty percent of participants indicated that *'not being a musical person'* made comparing melodies hard for them. Country/city names and sentences were lowly rated because they were *'too long'* for most participants while 10% of participants felt that it was *'strange'* for a mobile phone to be *'talking'* to them (alphanumeric & sound, numeric & sound). Participants also found comparing images challenging especially those that were meant to be similar. This may be attributed to the way the questions on the devices were phrased "are they DIFFERENT?" — a key recommendation from Uzun *et al.* [123]. For similar images, participants spent considerable amount of time looking for differences in the images but this was not the case with non-similar images since the differences were quite apparent.

In *manual copying and entering*, it was expected that alphanumeric would receive a low rating because of the difficulty in entering text on a mobile phone keypad especially where one is required to switch between numeric and text. *Barcode* was the lowest ranked method. This was because, first, participants did not understand how the method fitted into the simulated payment system. Second, they were not sure to what level of detail or how clear the image should be. Third, 67% of older participants (>45 years) had difficulties because they are *'not used to taking pictures using a mobile phone'*.

While the rankings in Table 4.6 provide the relative usability of the methods tested, they do not provide a complete means by which one may make an informed decision on a method suitable for the empirical channel. This is because security also needs to be considered. Security failures are not included in Table 4.6 because these are considered critical and cannot carry the same weight as other factors analysed. To this regard, it was necessary to re-rank the methods in a manner that gave security failures more weight compared to other factors. Table 4.7 shows this ranking. The method-representations are ranked according to their susceptibility to security failures, followed by the number of actual security failures observed in the study and finally their SUM scores.

	Subject to security failures	Security failures	SUM score
Numeric(MCE)	No	0	69
Alphanumeric(MCE)	No	0	60.4
Barcode	No	0	53
Numeric(MC)	Yes	0	73.7
Sentences(MC)	Yes	0	62.9
Countries(MC)	Yes	0	59.1
Images(MC)	Yes	0	54.3
Words(MC)	Yes	3.3	70.6
Numeric & sound	Yes	3.3	69.2
Alphanumeric & sound	Yes	3.3	65.8
Melodies(MC)	Yes	6.7	40.7
Numeric(CS)	Yes	10	68.3
Alphanumeric(MC)	Yes	13.3	72.5
Alphanumeric(CS)	Yes	20	64.2

Table 4.7: Ranking based on security failures

Manual copying and entering was ranked first because it is not susceptible to security failures and it had a relatively high SUM score. *Barcode* was ranked second despite a relatively very low SUM score, followed by *Manual comparison* and *compare and select.*

An analysis of interview data revealed two distinct groups of participants. One group was only concerned with the ease-of-use of the method-representations tested. For this group, speed, mental and physical workload of completing the association was of utmost importance. This is supported by scores for methods such as *manual comparison* numeric and alphanumeric that took least amount of time and participants pressed only a button once. The other group wanted a method that was efficient but at the same time reassuring them that correct devices were being paired through the accurate comparison of digests. These participants were found to favour *manual copying and entering*, first, because they felt it was secure since they *'can double check the string entered'* and hence were more likely to copy correctly. Second, they indicated that they were used to typing short strings such as PINs and short text messages on mobile phones and cash machines. Third, they were afraid that it was *'easy for one to be distracted in manual comparison'* or *'only compare the first few digits and think that the rest are matching'* resulting in pairing with a wrong device.

While *manual comparison* was ranked first in terms of usability, it is subject to security failures. It does not compel users to compare strings accurately. Users may cause security failures deliberately (by choosing not to compare), or because they are distracted, stressed, or conditioned to seeing matching values. Though these situations cannot easily be captured in a laboratory environment,

they do exist in the real world and cannot be ignored. In fact, method-representations are likely to perform better in a laboratory environment than in the real world. The study was designed according to the interface design recommendations in [123], however, the number of security failures observed in the various representations of the *manual comparison* method indicate that a user interface design alone is not sufficient to reduce user mistakes.

Despite *manual copying and entering*-numeric being ranked below *manual comparison* in terms of usability, it is not subject to security failures as users are compelled to copy accurately, otherwise the association will fail. While it is recognised that it may be difficult for some users to enter text on mobile devices such as a phone, the popularity of Short Message Service (SMS) means that a growing number of users are familiar with this means of interaction. The results do not indicate that this method is unusable, but its relative usability compared to other methods, combined with its inherently stronger security, makes it the best candidate among the methods tested.

Barcode's resistance to security failures, together with its ability to accommodate more bits than other methods in this study, makes it a very interesting candidate for an empirical channel. However, this method is limited to devices with cameras, which most laptops or PDAs generally lack. Moreover, most participants did not understand the process intuitively, and a substantial number of them felt it was an *'added complexity'*. It may be possible to overcome this problem through education and exposure to the technology. However, given the necessity of using a camera, this method is somewhat limited in its scope of application.

4.6 Factors affecting security and usability of empirical channels in single-user scenarios

4.6.1 User conditioning

Repetitive security tasks to which users can predict an outcome should be avoided. For example, an empirical channel using *manual comparison* used in an everyday application such as mobile payments may result in users anticipating matching digests and getting used to pressing "SAME". For example, one participant admitted, "*I pressed the wrong button once but that was my own mistake*" after accepting a digest that was not matching.

4.6.2 User motivation

Whitten and Tygar [129] concluded that users display 'unmotivated user property' when security is orthogonal to the task at hand and Ajzen [44] found that an individual's intention to show a particular behaviour is affected by that person's motivation to comply to subjective norms. Motivation, therefore, plays an important role in how users decide what action they can engage with. In addition to motivation, attitudes about a behaviour based on beliefs about and evaluations of that behaviour affect the intention to conduct the behaviour [44]. For example, Weirich and Sasse [126] found that users' lack of compliance with security policies is because of beliefs and attitudes that the risk is not real and that their behaviour is insignificant even when the risk is real. Moreover, security-critical tasks must be aligned with user goals [25]. They should help and not deviate or hinder users from accomplishing the goals at hand.

Users have different levels of motivation to perform security tasks in different circumstances. In the study, a good number of participants indicated that they would prefer typing digits longer than 6 digits for financial transactions exceeding a certain monetary value. In essence, motivation is contextual and depends on how one views the criticality of a security task. For example, some users may see their personal details, such as contact information, as crucial to protect hence may be motivated to engage in tasks that promise protection of such information. On the other hand, those who view such information as not crucial may be unwilling to actively apply protective measures.

In HISPs, motivation is crucial especially for tasks that do not compel users to perform them accurately. Unmotivated users may choose to not compare digests or only compare part of it. Given that different users may have different levels of motivation in similar contexts, empirical channels must ensure that critical tasks are enforced without putting undue burden on users.

4.6.3 Attentiveness

Users can easily be distracted causing them to shift their attention from the pairing process. Empirical channels must not demand undivided attention throughout the pairing process as this is likely to cause frustrations in scenarios where the user is distracted. For example, one participant commented, "*It's quite easy to compare things but listening to the melodies was the hardest because you have to, kind of like, once it finished you kind of forgotten the beginning of the melody. And the*

pictures [comparing images] *as well was a bit harder because I was really studying it to see if every single line was in the right place...*".

In the study, participants' inattention was revealed by the number of failures in *manual copying and entering* that were a result of users confusing copying of a digest with typing a PIN for the payment. Despite instructions on the digest screen being different from those of PIN screen, the presence of the text box made participants confuse the latter for the former. These failures also reveal that users will of course focus more on the primary task (in this case making a payment) rather than security tasks.

Concerns about attention were a common theme among participants. For example, one commented, "*Now there were sometimes the digits were very similar and there was only one digit that was different. Now if you were doing that standing on the Waterloo or Victoria stations with so much noises, you listen up for the train and you are trying to do these transactions so I think that's were the problem would arise because there is so much noise around, you have got other thoughts, you have a train to catch, and then you are trying to do this quickly and that's where the one digit that you can miss very easily, huh, will probably cause room for error. In a relaxed environment, if I was doing that* [comparing sounds] *from home then that would be fine, but if I was at home with kids screaming around me my attention is diverted and I am trying to do it very quickly, its very easy to slip and make mistakes*".

4.6.4 Device affordances

Affordances are the means through which a user interacts with a device [16]. They provide a means by which users can input information or instruction, and how they receive feedback. For example, a mobile phone may have a keypad and camera — as a means through which a user can pass information or commands to it — a display and speaker — through which the user gets feedback from it.

After the study, differences in the quality of affordances offered by the devices used had an impact on the usability and security of empirical channels. For instance, participants had difficulties taking a picture of a barcode due to screen size, comparing two images due to different screen resolutions, and comparing melodies due to differences in speaker quality. In order to be applicable to the different

association scenarios and contexts, empirical channels should be adaptable to the differences in affordances among devices.

4.6.5 Social contexts

People are governed by social norms and tend to conform to socially acceptable (informally) set of behaviours [32]. These norms and acceptable behaviour virtually govern how humans interact with various artefacts in different environments. The presence or absence of users not participating in a secure device association may be considered a social variable as users may behave differently in either case [35, 45]. For example, a number of participants thought that by having a mobile phone read out a digest, someone overhearing it can attack the association and hence would not want to use such a method in public.

4.6.6 Personal variables

Cranor [18] calls for considering users, and their characteristics, of a secure system. This is the main basis of design methods such user centred design [95], AEGIS [34], and participatory design [109]. They emphasise on putting users at the centre of the design and understanding their needs and characteristics. Users may broadly be grouped demographically in terms of age, gender, education, culture, occupation, and disability [18].

A user's security decision process may be influenced by their knowledge and experience of the system and the context in which it is made. Users usually misconceive the risks they are exposed to [126]; they either underestimate the risk — in which case security decisions expose them to the risk — or overestimate the risk — in which case they feel there is nothing they can do to protect themselves.

In the study of HISPs, personal variables such as age, experience with using a mobile phone, and health conditions of participants were identified as having an effect on the usability and security of empirical channels. For example, one participant commented, "*What I found interesting is that I am dyslexic* [health condition] *and if I was asked, to compare the numbers was a bit difficult. Selecting a number was fine but when I kind of had to take my eyes back and forth I transpose the numbers so and one of the comments I put down was I think, not everyone has dyslexic... I can see that*

[comparing numbers] *possibly presenting a problem for a small cross section of people and that was the biggest thing that was flagged up to me*".

In addition, personal experience with particular technology has an effect on how usable users perceive a method to be. For example, 67% of older (>45 years old) participants of this study complained about the barcode method for a single reason that they are 'not used to taking pictures using a mobile phone'. Understanding the target population and its characteristics is crucial to designing empirical channels that are both secure and usable.

4.7 Summary and conclusion

In this chapter, we presented the results of a user experiment of empirical channels for single-user scenarios. Our results show that technical security requirements of HISPs conflict usability factors. *Manual comparison*-numeric is the easiest method but subject to security failures while methods that are not subject to security failures (such as barcode and *manual copying and entering*) are difficult-to-use. Taking the security and usability of studied methods into account, we recommended *manual copying and entering* as the best method. *Manual comparison* and *compare and select* allows for complacency hence is unsuitable for HISPs where security may depend on users' attention. Moreover, the barcode method requires significant improvements in the decoding mechanism of the software to improve its efficiency and accuracy before users can accept it.

We also discussed factors that affect usability and security of HISPs. First, users get conditioned by repetitive tasks to which they can predict an outcome. Second, users must be motivated to effectively engage in a security task. Lack of motivation by users may increase the chance of security or usability failures or both. Third, HISPs are secondary tasks but require users' attention. The more attention an empirical channel demands on users, the less usable it is and, possibly, insecure too. In addition, different environments put different demands on users attention. A user carrying out device association in a quiet or well lit place may experience different attentional requirements compared to one carrying out the same task in a noisy or dark environment. Fourth, affordances that a particular device offers determine possible user actions. In addition, differences in affordances among devices has an effect on the usability and security of HISPs. Finally, personal variables such

as age, experience with mobile devices, knowledge and skills, and health conditions affect usability and security of empirical channels.

To design empirical channels that are both usable and secure for different contexts of use requires taking the above factors into account. These factors are broad, within the context of HISPs, and it is unreasonable to assume that an empirical channel can be developed that suits all contexts of use and all users. The factors, however, should be reasoned about within the context of a specific application scenario and target users.

Chapter 5

HISPs: Security and Usability in Group Scenarios

5.1 Introduction

This chapter discusses the theoretical usability and security challenges of group associations and presents results of a user study that evaluates effective security of empirical channel in these scenarios. Based on the results of this study, we discuss factors affecting usability and security of HISPs that are specific to these scenarios. The chapter is organised as follows; in Section 5.2 we discuss device association for groups while in Section 5.3 we discuss security and usability challenges of HISPs in group association scenarios. An empirical evaluation of empirical channels is presented in Section 5.4 and results of the study presented in Section 5.5. We discuss the results of the user experiment in Section 5.6 and highlight factors affecting security and usability of empirical channels in Section 5.7. We summarise and conclude the chapter in Section 5.8.

5.2 Device association in groups

Bootstrapping security in *ad hoc* networks for groups differs in many respects with single-user scenarios. The increased number of devices also increases the chance of these devices being significantly

different in terms of affordances, computation ability, and other features. Parallel to these differences among devices are the differences among humans using them.

Chapter 4 of this thesis identifies and discusses security and usability challenges of HISPs in single-user scenarios. Given the differences between single-user and group device association scenarios, the challenges identified, and recommendations made in the previous chapter may not apply to groups.

In this thesis, the term group is loosely used to mean two or more users interacting using mobile devices. Secure device association for groups may be categorised into 5 scenarios.

1. **Pairwise**

   One of the scenarios of group device association involves having only two participants (users) interacting using their mobile devices. In this scenario, either one device authenticates another or both authenticate each other. Practical examples of this scenario include two users wanting to exchange digital artefacts such as music, photos, or e-cash.

2. **One-to-many**

   The second group association scenario is one-to-many. In this scenario, one device is authenticated by two or more devices. An authenticated device may be manned such as another mobile device or unmanned such as a vending machine. In either case, HISPs specifically require that there is an empirical/unforgable channel from the party that is to be authenticated to the one that needs to be sure of it

   A practical example of this scenario is a medical emergency. In a medical emergency, say an earthquake with several victims, first responders will attend to survivors before taking them to a nearest available medical facility. In order to provide efficient and effective service at the hospital, first responders may need to transmit information to the medical facility so that medical stuff at the hospital are prepared for coming victims. However, first responders may not have devices powerful enough for long range transmission of data and may want to create a local area network among their devices with only one powerful device through which other devices transmit information to the medical facility. The crucial factor is for first responders to ensure that their devices are connected to the right transmitter — in which they have to authenticate it before any information is sent. In this life and death situation, the authentication process must be efficient as well as secure.

Another example scenario is a game, poker for example, in which players play the game using mobile devices while it is centrally managed by a single device. Individuals who want to play together form a group and authenticate the control device to ensure their devices are connected to the right device. The control device could be handling several groups, hence a number of participants participating in a single session of device association form a group that is managed independent of other groups.

3. **Many-to-one**

   The third group association scenario is many-to-one. In this scenario, one device authenticates two or more other devices. This may be appropriate in scenarios where group membership is controlled by one individual.

   One example of such a scenario is a meeting where a group controller wants to share sensitive information with other participants. For example, a CFO wanting to share sensitive information with members of a marketing team would want to have control on the attendees and would also endeavour to ensure that the information is only shared with known participants and none else.

4. **Partial symmetric**

   Partial symmetric is a congruence of many-to-one and one-to-many. In this scenario, a group of devices authenticate one device and vice versa. Group members authenticate a single device by ensuring that digests displayed on the devices match with one displayed on the single device. The single device owner authenticates group members by checking that devices of group members display *success*.

   To illustrate partial symmetric, consider a vending machine that issues cinema tickets. The machine can issue multiple tickets at a single instance to facilitate group orders. Each member of a group will receive a digital ticket on their device because the gate to the cinema allows for a single entrant hence one person cannot receive tickets on behalf of others in the group.

   In this example, group members will authenticate the vending machine to ensure that they do not receive fake tickets from a rogue device. On the other hand, the vending machine does not immediately issue tickets until a person manning the machine or one of the group members indicates on the vending machine that the association was successful. If the vending machine does not wait for instructions to distribute tickets, it may send tickets to users who are not members of the group ordering them.

5. **Full symmetric**

   The previous two scenarios require authentication in one direction or from a group of devices to one and vice versa. There are scenarios, however, where each device participating in the association needs to authenticate every other device. This may be viewed as a many-to-one scenario repeated for each device in the group. Each participant in the group authenticates other members by ensuring that a digest on a personal device matches every other device's. An example of full symmetric is a multi-player game where there is no central device to which participants' devices can connect. In this scenario, each participant is keen to ensure that the game is played with only the users within the vicinity and no one else.

5.3 Security and Usability challenges of group scenarios

The number of users involved in bootstrapping security between mobile devices may have serious implications on the security and usability of empirical channels. In group association scenarios, a well designed empirical channel may consider distributing human work among participating users and, as such, it may give an opportunity for using digests of sufficient size (for theoretical security) as opposed to where a single user is expected to do all the work.

As security is a process rather than a product [107], the number of nodes where security can fail may increase with each additional device or device/user pair since the correct behaviour of all participants is necessary to achieve desired security [23]. In secure device associations, participants achieve global security — by sharing a common cryptographic key, for example — among them only when they all behave correctly and are diligent in detecting anomalies. Empirical channels, therefore, can only achieve acceptable effective security when they make desired user actions easier to perform than undesirable ones within the context in which they are used.

The concerns to be addressed here are: how can we design (or how do we propose) empirical channels that allow for distribution of human effort among participants? How can a single user establish a secure association of multiple devices with acceptable mental and physical effort? How does increasing the number of devices or device/user pair affect the usability and security of a particular empirical channel?

Authentication is categorised as either one-way (asymmetric) or mutual (symmetric). In one-way authentication, one device authenticates one or more participating devices. For example, an Access Point (AP) authenticating mobile devices wanting to access the Internet through it (assuming the AP is configured to authenticate devices). In this scenario, a user may be happy to identify the AP by name (if they know it) or by other means. In short, the user conducts a weaker authentication of the AP. The AP on the other hand requires a stronger authentication in which it may prompt the user to transfer some information, using an empirical channel, to verify that the owner of the device is within the vicinity and hence (presumably) has access rights to it.

In mutual authentication, however, each of the participating devices authenticates all the other devices. In the AP example, the user or the personal device may require more than just a name of the AP. The device may require the AP to compute something which the user can verify.

Either of these scenarios poses different usability challenges. In one-way authentication, an authenticating device's acceptance of an association request is good enough for the authenticated device. For example, once a connection to a named AP is established, that is good enough for the device. In practice, the AP may require the user to transfer some information from the AP to the device and no further action from the user.

In mutual authentication, users may need to take extra steps. An AP may be required to indicate to the user acceptance or refusal of the association request and require the user to indicate to the device appropriately. The amount of effort expended in mutual authentication may be double that expended in one-way authentication. For example, using a 2D barcode [72] (as discussed in Chapter 4) to encode a digest of exchanged information, the barcode may have to be captured *n-1* times for one-way authentication and *n(n-1)* times for mutual authentication where n is the number of devices participating in the association. Understanding this difference in human effort between the two scenarios is essential to designing usable empirical channels.

The extra step in mutual authentication is not only an increase in human effort but also a step where security may fail. For example, a user misinterpreting a refusal by the AP as an acceptance of the association may result in associating a personal device to an unintended AP or the user may interpret a message on the AP correctly as a refusal but fail to indicate accordingly on the device.

Another challenge to security and usability of bootstrapping security in group scenarios concerns group size. A hidden node may participate in device association without revealing itself to legitimate

participants thereby exposing all shared secrets. To prevent this attack, participants need to ensure that the number of devices participating in the association is what is expected. This may be a job of an initiator, for groups with a leader, or each group member must ensure that only the expected number of devices are communicating with the personal device. It is a usability challenge because it will require participants to verify the number of devices involved. In applications where the number of devices can be predetermined, it may be reasonable to set this at the application level without needing users to verify.

In group device association scenarios, two channels of communication are important. First, an initiator (person controlling the group) must be able to communicate to each group member. This communication may carry information about digests, status of association, or other group members. An initiator may benefit from a broadcast channel where one transmission gets to all group members. For example, to announce the value of a digest, the initiator may read it loudly for everyone else in the room rather than passing around a personal device to each participant.

The second communication channel important in group device associations is from each group member to initiator. Group members need to communicate the result of the security task performed, whether it succeeded or not, and any relevant communication that may help in the association process. Without this channel, initiator will be in no position to know the status of the association once the digest has been read out to others.

With this background, sources of security and usability problems in group scenarios are discussed below:

- **Failure of communication from initiator to group members:** when initiator fails to communicate correctly, group members may take wrong information which may result in devices rejecting legitimate associations. This may cause frustration as the process has to be restarted after a failure.
- **Failure of communication from group members to initiator:** group members must communicate results of an association to initiator for the latter to make the correct decision of either accepting or rejecting an association. A failure in communication may result in usability problems because initiator may reject perfectly valid associations and also in security problems because initiator may accept invalid associations.

- **Inattentiveness by initiator:** initiator should be attentive and interpret messages from group members correctly. Failure to do so will result in similar problems as discussed in the previous bullet point.
- **Inattentiveness by group members:** this is similar to the problem in the first bullet point except that in this instance, group members do not pay attention to initiator's messages.

5.4 Empirical evaluation of group association scenarios

In order to evaluate security and usability of empirical channels in group scenarios, possible sources of both usability and security problems were identified by applying the model for security and usability analysis of secure system (details of this model are in [55] and Chapter 8 of this thesis). For usability, the following were identified: effectiveness, efficiency, satisfaction, and accuracy. These elements are commonly used as metrics in usability studies. For security, the following were identified: attention to the association process, conditioning, social context (group), vigilance (can participants be actively attentive to the association process throughout), and motivation. During the design and conduct of the study, particular attention was paid to these elements.

Upon identification of elements that may pose challenges to security or usability or both, the process for evaluating usability and security of secure systems (proposed in [55]) was used in designing the study scenarios. This process was used because, rather than just paying attention to usability during the designing and conducting of the study, it compels one to focus on security issues as well. Using this process, usage scenarios were identified — scenarios that represent real world applications of a secure device association rather than just security tasks as these are a secondary goal to users. The following usage scenarios were used: exchanging contacts, digital cash transfer, group messaging, and group quiz. Threat scenarios [55] were then identified. These are events that should never happen in a secure system. In secure device association of groups, threat scenarios are: accepting a non-matching digest, initiator interpreting failure of device association from one or more devices as success, intruder joining network without knowledge of initiator or other group members. These threat scenarios were incorporated as part of the experimental design of the study so as to determine how likely users may detect and defeat them.

The study used the partial symmetric association scenario for a number of reasons:

1. It covers both one-to-many and many-to-one scenario. One can, therefore, use a single study to evaluate performance of empirical channels for both scenarios covered in partial symmetric.

2. Choosing only one-to-many or many-to-one limits the generalisation that one can draw from data. For example, data on one-to-many association scenario may not be extended to any other scenarios.

3. Full symmetric is a special form of, and can be achieved using, partial symmetric. For example, in partial symmetric, rather than having participants report the status of the association to initiator, they may report it to other group members as well.

5.4.1 Participants

To increase power and reduce variability, a repeated measure with counterbalancing (to minimise learning effects) was used. Forty nine participants (24 male, 25 female) were recruited through mailing lists and online advertisements. Participants were randomly grouped into 13 groups with group sizes of 2 (2 groups), 3 (4 groups), 4 (3 groups), 5 (3 groups) and 6 (1 group). Each group performed the same test conditions (counterbalanced) and primary tasks. In each group, one member was randomly assigned to be an initiator. One group of 2 participants was later excluded due to errors in data collected. Table 5.1 summarises 47 participants' (excluding 2 males as above) demographics.

Gender	Male: 46.7% Female: 53.3%
Age	18 - 25 51.1% 26 - 35 21.3% 36 - 45 17% 46 - 55 8.5% 56+ 2.1%
Education	High School: 19.1% College: 31.9% Graduate: 27.7% Postgraduate: 21.3%

Table 5.1: Participant demographics

5.4.2 Materials and apparatus

In conducting the study, a tool that incorporated logging of user actions and user interfaces for interacting with mobile devices was used. The tool was implemented using Java Micro Edition (J2ME)

and runs on mobile devices that support Mobile Information Device Profile (MIDP) framework implementations. It supports test configurations (such as number of tests to run), event logging (i.e. completion time, number of buttons pressed, security and non security failures), different experimental designs (e.g. randomised, counterbalanced), and error simulation (See details of the tool in Appendix C). Usage and threat scenarios were implemented as a top layer of the tool. The study was conducted on Nokia N95 and Blackberry Bold 9000 devices.

After Scenario Questionnaires (ASQ) [68] were used to capture user ratings for each method immediately after encountering the method. For each method, ASQ captured data on three main components of usability (satisfaction, efficiency, and effectiveness):

- satisfaction with the ease with which a method was used,
- satisfaction with the amount of time spent on a method,
- whether participants felt they could effectively carry out primary tasks using a particular method.

ASQ is a rating scale type questionnaire consisting of 3 questions with answers based on a scale of 1 to 7, with 1 corresponding to *strongly agree* and 7 to *strongly disagree*. Many rating scales use a scale of 5 intervals rather than 7. However, it has been found that reliability of rating scales increases with the number of items and also the number of interval points for each item, and levels off at about 7 intervals with no significant increase after 11 intervals [68], hence the use of a 7 point interval scale.

An End of Experiment (EoE) questionnaire gave participants an opportunity to identify methods they felt were easy, difficult and which ones they preferred or would avoid. It also asked participants to rank each method on a 7 point scale with 1 corresponding to *very difficult* and 7 being *very easy*. Interviews gathered participants' views and comments on what they felt about the methods and group interactions.

Each test session lasted for about an hour, including a discussion. The sessions were video taped so as to analyse and understand how participants interacted and help identify elements that may help or hinder secure device association in these scenarios.

5.4.3 Methods tested

Methods tested in this study were chosen based on results of previous studies of single-user device associations presented in Chapter 4 (and those of Kobsa *et al.* [60] and Uzun *et al.* [123]) and on their practicality to group scenarios. It was also assumed that in group scenarios, devices will have reasonable input/output interfaces to allow for interactions such as messaging, gaming, and digital object transfer. Based on this, the following methods were tested: *manual comparison-numeric, manual comparison-images, manual copying and entering-numeric, repeated numeric comparison*, and *word-matching and number-typing.*

5.4.3.1 *Manual comparison-numeric*

Among the proposed forms of presenting digests to users in *manual comparison*, numeric is the most basic and participants in previous studies (such as [123] and the study in Chapter 4) rated the method as the most usable. This method was included to assess whether users would still find it usable in group settings. In addition, it was of interest to assess the security of the method considering that it is subject to security failures. Two cases were used in the study:

1. Usage case: every device in a group displayed a digest that matched one shown on initiator's device. This resulted in a successful association if participants executed their tasks correctly. A mistake by a participant may result in initiator rejecting an association, that is, non-security failure.

2. Threat case: one device displayed a digest that did not match with initiator's and an accurate comparison should result in the initiator rejecting an association similar to the previous case. If a participant whose digest did not match paid no attention to the tasks or failed to communicate the mismatch to the initiator, it resulted in a security failure.

5.4.3.2 *Manual comparison-images*

As earlier discussed in Chapter 4, participants in the study of single-user device association scenarios had difficulties comparing images on two mobile phones, especially when the same image was displayed on both devices. It was realised that this was the case because participants where looking

for differences between the two images. It was later on suspected that this could have been due to instructions given to participants — asking "ARE THE IMAGES DIFFERENT?". For group scenarios, it was decided to change instructions to "ARE THE IMAGES SAME?" to see whether this would improve the performance of image comparison. We used the same images as shown in Figure 4.2. During pilot studies, nonetheless, we decided to leave out this method from the actual study for two reasons. First, changing instructions had no positive impact on participants' behaviour towards the method and, second, it was taking much longer compared to other methods.

5.4.3.3 *Manual copying and entering-numeric*

This method is not susceptible to security failures but it was of interest to assess its usability in a group context given that previous studies of single-user device association scenarios have found that users have difficulties in using it. Though there have been proposals to use other formats, such as alphanumeric, only numeric was used because it is easier to type on a multi press keypad such as one found on standard mobile phones as compared to alphanumeric entry. Two cases were tested with this method:

1. Usage case: every device in a group had the same digest as initiator and upon typing it in correctly, an association should be successful. An incorrect entry will result in the initiator rejecting the association causing a non-security failure.

2. Threat case: to simulate a threat case for this method, one device was randomly selected to reject any string of numbers entered. For example, a participant may enter a number correctly as read by initiator but the device will alert the user that the association failed. While in practice a successful attack may be impossible to carry out for this method, the aim of the threat case was to assess whether users could respond correctly in such an event. In addition, failure to communicate to the initiator that an association failed could result in initiators accepting an association with a wrong device(s) resulting in a security failure.

5.4.3.4 *Repeated numeric comparison*

To compel users to carry out manual comparison securely without undue effort, *repeated numeric comparison* was proposed in [54] (see Chapter 6) as a two step process. In addition to a digest, an

authenticating device generates a random string of similar format to the digest. The authenticating device then randomly chooses to display either its digest or the random value. The user compares and indicates whether the string displayed on the authenticating device matches that on the other device. An authenticating device then displays the remaining string and the user does the comparison again. An authenticating device accepts a connection only when a user indicates a match for a digest and a mismatch for the random value. The argument for this method is that, unlike manual comparison, it is difficult for users to ignore the crucial task of comparing digests without causing one device to refuse connection. Like with previous methods, repeated numeric comparison had two test cases:

1. Usage case: every member device in a group had an actual digest and correct comparison by participants would result in a successful association.
2. Threat case: one device in a group was randomly selected by initiator device and assigned values of which none is a match to an actual digest. A participant with this device saw (after correct comparison) a "connection failed" message which should be communicated to initiator. Failure to communicate this message, initiator may accept association — when in fact one device in the group has rejected it — resulting in a security failure.

5.4.3.5 *Word-matching and number-typing*

This method is based on the fact that *manual copying and entering* is not subject to security failures but is regarded as difficult to use. While earlier work has argued that typing short strings on devices with limited input interfaces is hard for most users, the popularity of Short Message Service (SMS) is an indication that users are comfortable with such a task. One should, however, be cognisant of the lack of motivation from users to type strings for the sake of security. *Word-matching and number-typing*, proposed in [54], is aimed at offering the same level of security as *manual copying and entering* but only requiring users to type a smaller number of digits. Details of this method are in Section 6.3.1.

Figure 5.1 shows *word-matching and number-typing.* Initiator's device displays 3 words, two of which represent an actual digest. Group members' devices randomly display one word from a computed digest and prompt users to enter the position of the word shown as displayed on initiator's device. In Figure 5.1, for example, a user will type '3' for SON and press confirm. The device will display a second word and user enters the position of that as well. Two cases were tested:

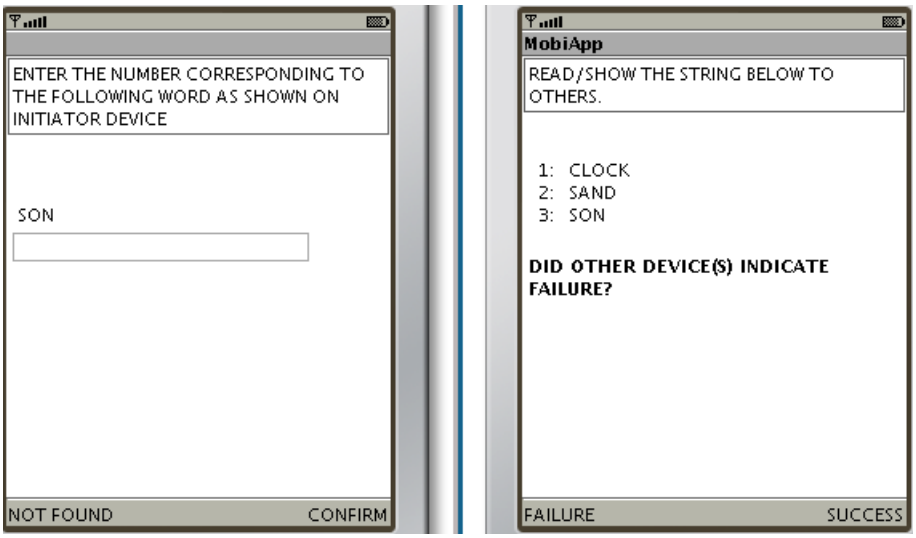

Figure 5.1: Word-matching and number-typing method: group member device on left and initiator's on right

1. Usage case: all devices display two words, one after another, that are among a list of 3 displayed on initiator's device. Correct entry of positions of these words on member devices results in successful association. Mistakes in entry results in a safe failure.

2. Threat case: one randomly selected device displayed a word that was not among those displayed on initiator's device. This resulted in a security failure if the mismatch was not communicated to the initiator so that the association could be rejected.

5.4.4 Participant tasks

As earlier stated, usage scenarios (representative tasks) were developed to represent real world applications in which security tasks could be applied. Upon arrival at the laboratory, participants were taken to a room where the study was conducted. They sat around a square table and were asked to sign consent and enrolment forms. Mobile devices were then distributed randomly to participants, except Blackberry devices that were only given to participants who had used one before. Participants were then given an overview of what the study was about and what the tasks were, outlining the roles of initiator and other group members.

During the study, participants were allowed to ask the test observer or discuss amongst themselves any issues they were not clear about. Mobile devices prompted participants to complete ASQ as they

encountered each method. This was deliberately done so that these questionnaires were completed while users still had a vivid picture of a method a particular questionnaire was about.

After a successful connection among devices, the initiator then started an application (primary task). Upon completion of the primary task, another association process was initiated. Figure 5.2 shows screen shots of steps participants were required to take in order to achieve their primary tasks — in this case, exchanging contacts. As earlier discussed in this chapter, the tests were designed such that there were two usage cases and one threat case for each method. Two usage cases were used for each method so as to check if there would be an improvement the second time a method was encountered. In addition, 3 of the 8 usage cases were meant to result in a failure due to a wrong number of devices being displayed on initiator's device.

5.4.5 Hypotheses tested

Three hypotheses were formulated:

- Hypothesis 1: there is no difference in the dependent variable time for the between-subject factor age across methods.
- Hypothesis 2: there is no difference in the dependent variable completion time for the within-subject factor method.
- Hypothesis 3: there is no difference in the dependent variable rating scores for the within-subject factor method.

5.5 Results

To test the above hypotheses, results are analysed by between-subject factor age and within-subject factor method. For both factors, 4 dependent variables are analysed: time to complete association, rating scores, preferences, and failures.

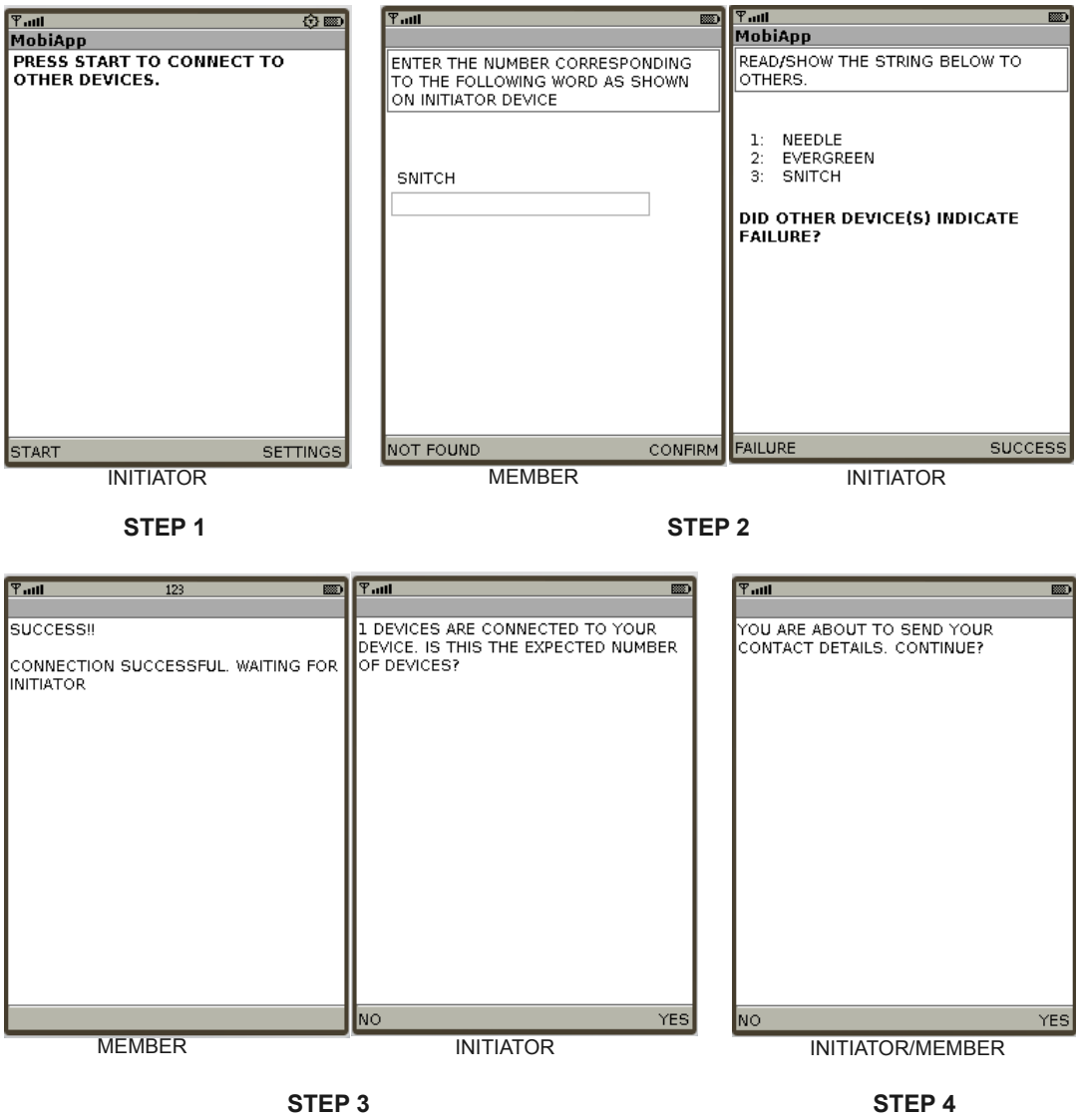

Figure 5.2: Task sequence: Initiator starts a connection and carries out a security task (in this example *word-matching and number-typing*) with group members. After successful association, initiator confirms number of devices and activates a primary task, in this case exchanging contacts

5.5.1 Analysis by age

Similar to Chapter 4, we analysed results by age to determine whether the differences between the age groups are significant. Participants were divided into two age categories: 35 years old and below (n=34) and over 36 (n=13). Of interest are time to complete security tasks, failures, and ASQ scores. Initiators' completion times were excluded from this analysis because these were significantly higher than other group members. Moreover, there were more initiators in the younger group compared to the older. Failures were calculated as percentage of maximum possible failures.

Rating scores were treated as ordinal data and are, therefore, summarised using mode rather than mean. Modal scores are presented with frequencies (converted to percent due to difference in number of participants between the two groups) to give an indication of the number of participants with that score. Preferences were calculated as percentage of age group who preferred a particular method. The results are summarised in Table 5.2.

	Rating (Mode)		Failures (%)		Time (Seconds)	
	Y	O	Y	O	Y	O
Manual comparison	7(30)	6(56)	0	0	7	8
Repeated numeric comparison	6(33)	6(33)	9	14	14	19
Manual copying and entering	7(33)	7(56)	5	0	12	13
Word-matching and number-typing	6(19)	6(33)	3	11	18	26

Table 5.2: Performance by age: Y=younger group (<36 years, n=27), O=older group (>35 years, n=9). Rating scores (x(y)): x=mode, y=frequency as percentage of n.

The results show that for either group, the mode for rating score was at least 6 with relatively higher frequencies for the older group. The results, however, show no evidence of a significant difference in rating scores for either group across methods.

Completion times for the older group are consistently higher than the younger participants across methods. A statistical test (using *t-test*) showed that there is no statistically significant difference in completion times for *manual comparison* (p (2-tailed)= .666), *Repeated numeric comparison* (p (2-tailed) = .185), and *manual copying and entering* (p (2-tailed) = .414). There was significance, however, for *word-matching and number-typing* with p (2-tailed) = .024.

5.5.2 Analysis by method

Performance of each method is analysed according to security and non-security failures, completion times, participants' rating scores, and preferences.

5.5.2.1 Security and non-security failures

Non-security failures are events where device association is terminated by initiator either because one member made a mistake on the security task or because of miscommunication between initiator and members. Security failures are events that make participants in a group to believe they have

established a secure communication channel among themselves when, in fact, an unwanted device may have joined the group or a subset of the group is duped into connecting to someone else other than those intended. Security failures may occur at two levels: when one device indicates unsuccessful association but initiator indicates success on the personal device or when device association is successful but number of devices connected to initiator device is not what is expected and is unnoticed. Table 5.3 summarises the results on security and non-security failures.

	Security %	Non-security %
Manual comparison	1.2	1.2
Repeated numeric comparison	2.4	17.9
Manual copying and entering	0	3.6
Word-matching and number-typing	0	8.3

Table 5.3: Security and non-security failures

None of the security failures observed were due to accepting the wrong number of devices connecting to initiator but rather due to failure in communication between group members and initiator. Failure in communication was also partly the problem with non-security failures. With *repeated numeric comparison*, the high percentage of non-security failures was due to group members misunderstanding the method; they would compare the first number displayed on the device and, rather than comparing the second number, they reported this directly to initiator. For example, one participants reported, *"oh, I had a number which is same as yours* [initiator] *but now I have a different one"*, with initiator responding *"OK, that is a failure then"*. *Manual copying and entering*'s failures were mostly due to typos while there were some confusions with *word-matching and number-typing* which caused some participants to type same digit for both words.

5.5.2.2 Completion times

Completion times were analysed at two levels: a group member's time to compare or type digests and initiator's time to confirm success or otherwise. Initiator's completion time represents a group's time to complete an association — the time from when a digest is displayed on the personal device to when every member has completed the security task and communicated the result to initiator. Table 5.4 summarises completion times for group members.

A repeated measure analysis of variance (ANOVA) was used since this is a within-subject design with more than two dependant variables. Mauchly's test indicated that the assumption of sphericity

	Min	Max	mean
Manual comparison	2	42	7.89
Repeated numeric comparison	3	64	17.63
Manual copying and entering	5	46	12.97
Word-matching and number-typing	6	94	22.89

Table 5.4: Group members' completion times (in seconds)

had been violated ($\chi^2(5) = 74.36$, $p < 0.05$) hence a corrected value (Greenhouse-Geisser correction) of F was used. The test showed that there are significant differences in completion times among methods $F(2.084, 216.77) = 36.6$ and $p = .000$. Pairwise comparisons of completion times between methods showed that each method's completion times were significantly different from each of the other methods with p-values ranging from .000 to .017.

Table 5.5 summarises initiators' completion times. A repeated measure ANOVA test on completion times for initiators showed that there are no significant differences between methods with $F(2.04, 71.3) = 1.22$ and $p = .277$ (Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5) = 31.24$, $p < 0.05$) hence a corrected value, Greenhouse-Geisser correction, of F was used).

	Min	Max	mean
Manual comparison	7	278	40.97
Repeated numeric comparison	11	105	33.94
Manual copying and entering	8	107	36.27
Word-matching and number-typing	11	147	48.27

Table 5.5: Initiators' completion times

This seems contradictory with earlier analysis on group members where differences in completion rates are significant. The video evidence, however, reveals that initiators allowed some time to elapse before asking group members if their devices displayed failure or success. In some cases, this time was way after members had completed their tasks while in others they were still doing the tasks. To some extent, no matter how fast group members completed their tasks, initiators allowed for some time to pass before they thought it was time to move to the next one hence the lack of statistically significant difference between methods.

5.5.2.3 Rating scores

Initiators' and group members' rating scores are analysed separately. This is due to differences in tasks carried out by each group on each method. For example, while group members are required to type 6 digit numbers on their devices in *manual copying and entering*, initiators only read a number displayed from a device. Tables 5.6 and 5.7 summarise the results for group members and initiators respectively. The tables show the minimum and maximum scores (and their frequencies) for both ASQ score and overall (O) rating scores. Unlike the ASQ score that was used to rate a method on three different criteria, the overall score is a single score that participants assigned to each method at the end of the study.

For group members, the table indicates that participants changed their rating scores between initial encounter with a method and completion of study. For example for *manual comparison*, initially only 10 participants rated the method with a score of 7 while 19 gave the same score for overall rating.

	Min	Max	Min(O)	Max(O)
Manual comparison	5(3)	7(10)	4(1)	7(19)
Repeated numeric comparison	4(2)	7(6)	2(1)	7(13)
Manual copying and entering	4.3(1)	7(13)	4(4)	7(19)
Word-matching and number-typing	4.3(1)	7(6)	4(2)	7(18)

Table 5.6: Group members' rating scores. X(Y): X = score, Y = frequency. Min(O) = min for overall score

Rating scores were analysed for statistical significance using Friedman test. The test showed that there are significant differences in rating scores (ASQ) between methods with $\chi^2(3) = 11.655$ and p = .009. The test ranked *manual copying and entering* first, followed by *manual comparison*, *word-matching and number-typing*, and finally *repeated numeric comparison*. A pairwise Friedman's test was also carried out to find which methods had significant differences between them. The tests showed that there is statistically significant difference between *manual copying and entering* and *repeated numeric comparison* with $\chi^2(1) = 8.91$, p = .003 and between *manual copying and entering* and *word-matching and number-typing* with $\chi^2(1) = 4.84$, p = .028.

A Friedman test on overall scores, however, showed no statistically significant difference between methods with $\chi^2(3) = 5.526$ and p = .137. Again, this statistic just shows that participants changed

their ratings — by the end of the study they had a better understanding of the tasks required of them — more participants gave favourable scores thereby normalising the initial differences.

Initiators were expected to give higher ratings compared to group members considering the difference in the tasks they performed. It was also expected that for similar tasks, e.g. reading a number, initiators will give similar ratings. Results, however, show that this was not the case. First, a Friedman test shows that there is no statistically significant difference in ASQ scores ($\chi^2(3) = 4.558$, p = .207) while there is statistically significant difference in overall scores ($\chi^2(3) = 11.082$, p = .011). For both scores, *manual copying and entering* was ranked first, followed by *repeated numeric comparison*, *word-matching and number-typing*, and lastly *manual comparison*.

	Min	Max	Min(O)	Max(O)
Manual comparison	3.7(1)	7(3)	3(1)	7(2)
Repeated numeric comparison	4.3(2)	7(4)	2(1)	7(5)
Manual copying and entering	2(1)	7(4)	5(1)	7(8)
Word-matching and number-typing	3(1)	7(4)	3(1)	7(4)

Table 5.7: Initiators' rating scores. X(Y): X = score, Y = frequency. Min(O) = min for overall score

A pairwise Friedman test shows that there is statistically significant difference between *manual copying and entering* and *repeated numeric comparison* with $\chi^2(1) = 4$, p = .046, *manual copying and entering* and *word-matching and number-typing* with $\chi^2(1) = 5$, p = .025, and between *manual copying and entering* and *manual comparison* with $\chi^2(1) = 6$, p = .014.

5.5.2.4 Preferences

Similar to rating scores, preferences were analysed in terms of initiators and group members. Participants were asked to indicate all methods that they felt were easy as well as those they felt were difficult to use. They were also asked to indicate which method they felt was the easiest, the most difficult, their personal choice, and which one they would avoid, given a choice. Figure 5.3 summarises the results for group members and initiators.

It is interesting to note that even though the tasks for initiators and non-initiators were different, both graphs generally follow similar trends; methods that are preferred by group members are also preferred by initiators. It is surprising to note that initiators gave different ratings to methods that had similar tasks. For example, for CC, CE, and RC initiators only read out a number displayed on

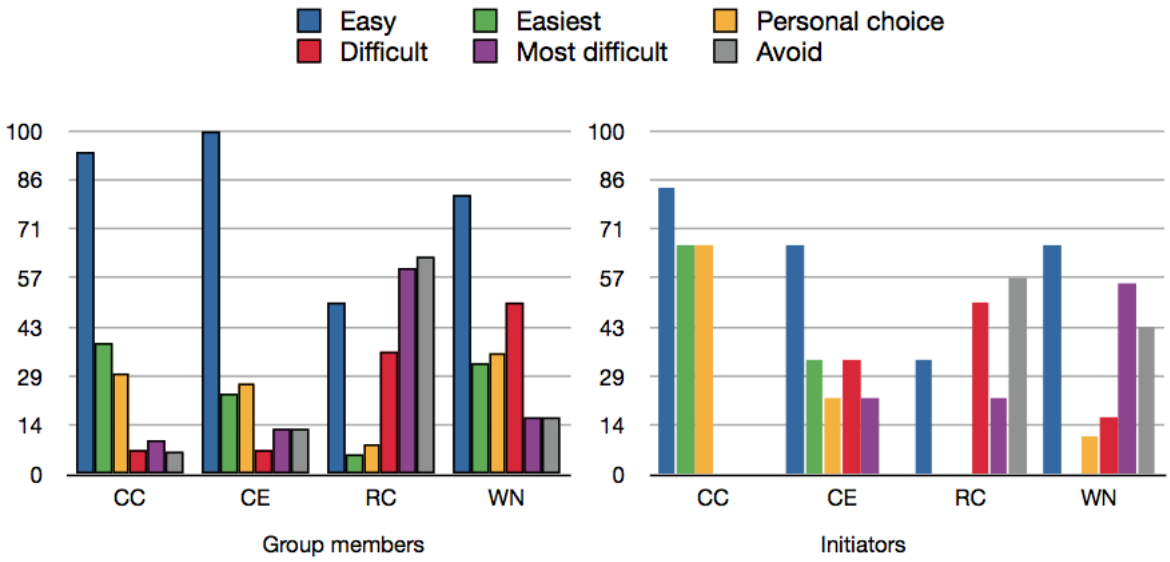

Figure 5.3: Preferences: Group members (left) and Initiators (right). CC = manual comparison, CE = manual copying and entering, RC =repeated numeric comparison, WN = word-matching and number-typing

their device. The data, however, shows that initiators gave 'holistic' ratings; they took into account the tasks they performed but also those performed by group members.

5.5.3 Hypotheses validation

Having analysed the data, initial hypotheses are now revisited.

- Hypothesis 1 — there is no statistically significant difference in the dependent variable completion time for the between-subject factor age across methods: results of the statistical analysis presented above prove that this hypothesis is true except for *word-matching and number-typing* where *t-test* statistic p = .024.

- Hypothesis 2 — there is no statistically significant difference in the dependent variable completion time for the within-subject factor method: for group members completion times, this hypothesis is nullified since analysis of variance shows that the differences in completion times for all the four methods are significant. On the other hand, the hypothesis is true for initiators completion times.

- Hypothesis 3 — there is no statistically significant difference in the dependent variable rating scores for the within-subject factor method: this hypothesis is not true for *manual copying*

and entering and *repeated numeric comparison* with $\chi^2(1) = 8.91$, p = .003 and for *manual copying and entering* and *word-matching and number-typing* with $\chi^2(1) = 4.84$, p = .028.

5.6 Discussion

The previous section has presented an empirical study on the usability and security of four methods for secure device associations in group scenarios. Unlike single-user associations, completion times in group scenarios are affected by activities of the slowest group member since a task is only completed when all individuals have performed their role. These activities tend to normalise group completion times and minimise the differences between methods. In some cases, group completion time is dependant on the slowest member while in others it is dependant on the initiator.

The main source of failures (both security and non-security) is lack of communication between group members and initiators. A failure on a group member's device that is not communicated to an initiator may be interpreted as success by the latter. Initiators may also fail to communicate effectively to group members. For example, in one group during the study, the initiator read a wrong digest only to realise it after group members had typed it in.

Even though, in this study, initiators correctly observed the number of devices and rejected device association when an unexpected number of devices was displayed, it is possible that this may be problematic in day-to-day interactions. A possible solution is having initiators commit to the number of devices they expect. Initiators may be asked to enter the number of participants before device association which initiators' devices can verify once association is complete.

Preferences are subjective and difficult to quantify but provide an insight into how (potential) users feel about a particular interaction. During interviews, participants revealed that they preferred certain methods to others because they felt those methods were either easy to use or secure. This may be a possible explanation to why Figure 5.3 shows that group members rated *manual comparison* and *manual copying and entering* highest in terms of ease-of-use while *word-matching and number-typing* was the most preferred. There are, however, a number of questions that needed answers:

1. In the study, what is the distance between ease-to-use and difficult-to-use? To answer this question requires referring back to the study interviews and rating scores. From interviews, it

was evident that because questionnaires asked which methods were difficult, participants felt obliged to nominate at least one method. Looking at ASQ scores and overall ratings, methods labelled difficult in preferences have high ASQ scores. Both scores and interviews show that the distance between easy and difficult is not compelling enough that it would force a user to choose one method over another.

2. What did participants mean when they indicated that they would avoid a particular method? The same analysis as to the previous question was applied here. The same result was found; participants nominated methods to avoid because they felt compelled to do so. A surprise, however, is the consistency — some methods were consistently nominated for being difficult or avoidable. For example, most participants nominated *repeated numeric comparison* for both categories. It was discovered that this was the case because participants had a reference point; *manual comparison*. Participants felt *manual comparison* was easy and sufficient hence no need for a similar method that required them to compare twice.

5.7 Factors affecting security and usability of empirical channels in group scenarios

Given the initial theoretical analysis of security and usability challenges and an empirical study of device associations in group scenarios, factors that may affect security and usability of HISPs in these scenarios are presented below:

5.7.1 Trial and error

One of the five elements of usability is learnability — emphasising design of user manuals that are accessible to users. The basic assumption is that users will take time to read through manuals and understand how a system works. There is sufficient evidence, especially in secure systems, that users will attempt to use a system first and consult a manual only when it is absolutely necessary. For example, during the study, basic background information about device association was given and participants were asked whether they had any questions or understood what was required of them. The response was usually *'We will give it a go and ask when we get stuck'*.

A secure system must not depend on correct execution of instructions inside a user manual but must be designed to accommodate a 'trial' phase. This is a learning period that users attempt to 'check' how a system works. To accommodate this phase, possibility of security failures must be minimised to an acceptable level if not eliminated. In HISPs, an empirical channel must be designed with a 'trial and error phase' in mind, that is, must accommodate failures and allow users to recover without compromising technical security. In other words, security of HISPs must not depend on human attentiveness but rather human effort should only compliment security requirements. Considering that a single security failure may result in substantial loss or damage, mistakes committed during and after 'trial and error' phase must result only in non-security failures.

One approach to mitigate the risks that trial and error may bring to empirical channels is to employ commitment rather than confirmation. An empirical channel should only reveal partial information that can be used to commit to a final outcome of an association. With incomplete information, a user is only limited to a commitment rather than confirmation. Earlier in this chapter, we discussed an example of commitment to group size. In this example, initiator knows before hand the number of devices (or participants) expected. Using this information, initiator can enter the expected number of devices before association is initiated so that a final outcome, acceptance or rejection of number of participants, is determined by initiator's device. This way, a user cannot change the outcome after entering the number of devices expected and reduces the chance of a successful attack assuming an attacker is not able to block messages transmitted by one or more member devices.

5.7.2 Context

During the study, it was realised how crucial a context of operation is to understanding and analysing issues surrounding a system under investigation. Participants had a clear idea what their 'primary' tasks were at every instance. For example, on using a messaging application one participant commented, *"I am a social worker and hold highly confidential discussions about child welfare and I have reservations in using this system in that environment. I prefer face to face and paper based communication which limits where that information can go. This may be just an age issue but that's how I feel about it"*. Context in laboratory studies not only prevents participants from focusing on security tasks, as though they are primary tasks, but also helps in soliciting data that goes beyond laboratory settings. For example, when asked which method was most user friendly a participant,

rather than focusing on the laboratory setting of six participants, commented, *"All these methods are straight forward but I can imagine where there are 50 of you and want to play a game..."* .

Contextualising the study ensured collection of data that has external validity. Moreover, it also revealed how contextual security requirements are. For HISPs, mobility and pervasiveness of devices enable users to carry out different tasks in different environmental, social, and technological contexts. Empirical channels must be designed with an understanding of the different contexts in which they will be used. Failing this may result in significant security and usability issues.

To give further examples of how context may have significant effects on usability and security of HISPs, the following are considered. On using a camera phone to take a picture of a barcode, most participants pointed out that they gave the method low ranking because it was not clear to them how it fitted into a P2P payment process. This example shows that context may cause a method to be ranked as less usable (low satisfaction scores). On the other hand, context may also help a method to be regarded as usable. During our user studies, participants felt that if they had to use HISPs in conducting high value transactions, they would be happier to type (*manual copying and entering*) PINs of longer than 6 digits.

The above examples of context relate to specific applications. Context may also be social or physical environment. Noisy environments, for example, may have a significant impact on the usability and security of methods that rely on sound. For the same methods, users may feel uncomfortable using them in public places — as it was found during our single-user scenario study.

Humans are constantly making security decisions [25]; conscious or unconscious. Different users may make different security decisions under similar circumstances. Naturally, risk averse users take fewer risks in the digital world compared to those who are not. In device association, changing context means that users are likely to make different decisions in different environments. Ion *et al.* [45], for example, mention that participants in their study indicated that they would change a choice of an empirical channel depending on whether they were interacting with friends or a business colleague, in a private space (such as an office) or in public. A choice of method based on context is also discussed in [53].

5.7.3 Sum of efforts

It is widely acknowledged that system security is equivalent to the weakest link in the chain [105]. However, Anderson and Moore [5] have argued that system's security may also depend on the best effort or sum of efforts. Device association for groups is an example of where security depends on a sum of efforts. While initiators were in 'charge' of their groups, group members were observed to be helping each other. For example, group members recited digests for other members or took the effort to look at each other's device and helped taking a correct action. Compared to previous studies of single user device association, group effort reduced the number of failures.

Design of secure systems where users work as a group, rather than independently, to achieve a common security goal should exploit the principle of sum of efforts. A well designed secure system for groups should ensure that the security or insecurity of a system depends on multiple users rather than a single user. In group device association for example, a large group may be split into smaller groups that compare digests and report success or failure as a group rather than as individuals.

5.7.4 User conformity

During the study, some participants felt it was their fault that their devices displayed 'connection failed' (and refused association). Consequently, they hesitated to communicate association failure to other members of the group. Initiators, however, consistently asked each group member what message was displayed on a device (except in two cases, in which it resulted in security failures). The issue was that members whose devices displayed 'success' quickly communicated to other members and members whose messages were different found it difficult to communicate otherwise. Group members wanted to conform to what they viewed as an acceptable behaviour, that is, getting a successful association.

Users' attitudes and behaviour are influenced more powerfully by what they see than what they are told [65]. Whilst conformity is undesirable for security as discussed above, it is good in certain scenarios considering that it is not only applicable to insecure actions but to secure ones as well. For example, in the study reported in this chapter, it was observed that some members of a group may not want to carry out a security task or report the results of it but once they realised that other members were doing so they followed suit.

In group interactions of HISPs, empirical channels must be designed such that it is difficult for individual group members not to report the result of an association. This may be implemented at device level, for example, where if a device rejects an association it beeps so that other group members can hear it. Such solutions not only make it easier to detect if there is a failure but also relieves users of the requirement to report success.

5.7.5 Interpretation of security

Users will almost always come up with a plausible explanation of how a system's security mechanism works. Such an explanation may be based on a number of factors including experience with other systems, observation of the system's user interface, and the context in which they experience the system. During the study, participants came up with various explanations on how the association worked and what should be done to make it more secure. For example, in answering to why he felt *word-matching and number-typing* was more secure than *manual comparison* a participant responded, *'...it is harder to attack three random words from a big dictionary of 9000 words. That's not possible to brute force'*. While this explanation is plausible, it is not correct in this instance.

Problems arise when users come up with explanations that cause them to behave insecurely. For example, users of peer-to-peer systems end up sharing files they do not wish to share because they do not realise that by default the system creates shares [38]. Secure systems should be designed such that their user interfaces closely represent how it works and prevents users from engaging in insecure actions regardless of how they think the system works. For example, participants in the study thought digests should be kept secret among group members. While this is not correct for HISPs, it does not result in insecure behaviour.

5.7.6 Perception of security

Secure systems that are complex and difficult to use are so pervasive that users encounter them in their day to day interactions. As a result, users have come to believe that a system that is complex and difficult to use is more secure than a simple and easy to use system. In this and previous studies, participants felt *manual copying and entering* was more secure than *manual comparison* simply because the former is more difficult to use than the latter. In addition, some participants

were prepared to type numbers longer than 8 digits for applications they considered sensitive even though this does not change the security of HISPs in a linear sense.

Users' perception of security reveals how pervasive complex and difficult to use secure systems are. It shows that insufficient work is being done in this direction even after at least 10 years of Human-Computer Interaction Security (HCISec). This view can only be changed by consistently designing secure systems with usability in mind from the onset as well as evaluating and improving existing systems. In HISPs, users' perception of security influences subjective usability of empirical channels.

5.8 Summary and conclusion

The challenges of security and usability of HISPs for groups differ in many respects from those of single-user scenarios. In group scenarios, coordination among participating members is crucial to achieving required security. In addition, the number of participants in a group affects usability and security depending on the empirical channel used.

In this chapter we presented an analysis, evaluation, and comparison of empirical channels in group scenarios and highlighted factors that affect security and usability. While it has been believed that group settings may be more subject to failures during an association process compared to single-user associations, our findings show the converse to be true. Group members feel the need to help each other and cover up the weaknesses of struggling members. Comparatively, there is no statistical significance in the differences among methods evaluated in terms of group completions times, ASQ scores (initiators), and overall rating scores (group members). There is, however, statistical significance in individual group member completion times and initiators overall rating scores. Analysis of ASQ and overall scores, preferences, and completion times indicates that all studied methods are acceptable in group settings.

Security and usability of empirical channels are affected by a number of factors. In group device association scenarios, effective security is a function of sum of efforts rather than of the weakest link. Users tend to work as a team and weaknesses of a subset of members are covered by more active members. In addition, users designated as initiators take the responsibility of asking other members the result of the association.

In addition to sum of efforts, security and usability of empirical channels are affected by context and users' perception of what security is. The context of operation dictates possible user actions while perception affects users' subjective usability and acceptance of empirical channels. Furthermore, users learn by trial and error only the parts of a system they consider important to their primary task.

In conclusion, the design of empirical channels should utilise the coordination effect of group scenarios to effectively defend against attacks that may target potentially weak participants. Moreover, attention must be paid to context of operation, users' interpretation of security, conformity, perceptions, and learning behaviours. These factors have an impact on usability or security or both.

Chapter 6

Principles for designing empirical channels

6.1 Introduction

In this chapter, we present principles for designing effectively secure empirical channels. The HISPs framework presented in Chapter 7 is designed for researchers and designers to identify and reason about factors that affect effective security of empirical channels in a given context. The principles presented in this chapter mitigate factors identified in the framework.

The design principles were developed using task analysis and understanding the different types of user interactions required in HISPs. In addition, there is a body of literature in HCI and social sciences discussing human limitation on various tasks (see discussion in Chapter 7). These limitations helped in developing principles for designing empirical channels by contextualising them within required HISPs interactions. The principles are demonstrated using two proposed empirical channels. A usability experiment was conducted to compare results of these methods with previously proposed techniques.

The chapter is organised as follows; in Section 6.2 we present the design principles. We present empirical channels proposed on the basis of the principles in Section 6.3. We present a user study

of the empirical channels and discuss its results in Section 6.4 and we summarise and conclude in Section 6.5.

6.2 Principles for designing empirical channels

6.2.1 Principle of commitment

Users tend to focus on primary tasks rather than secondary ones resulting in security tasks receiving minimal attention whenever possible. While some users pay particular attention to security tasks even though they are not their primary goals, it is crucial that security is guaranteed even for complacent users. In HISPs, changing contexts, and users' personal characteristics, exert strain on users as they perform device association tasks.

In order to mitigate security failures that legitimate non-malicious users may cause in HISPs, the *principle of commitment* must be employed. This principle can simply be stated as 'a user is committed to a particular value/action without knowing what the outcome of such a value or action will be'. The outcome of a user's value/action is only revealed after the user is committed to it.

The essence of this principle is that it ensures that users fail to achieve their desired outcome at the cost of security. In HISPs, users' desired outcome is one that leads to accomplishing a primary goal. If users can determine what action (or set of actions) lead to an intended goal, they have no incentive to explore alternative actions. For example, a user who knows that acceptance of a displayed digest value always results in completing a payment transaction has less motivation to explore a rejection of the value.

Empirical channels, proposed in other literature, have violated this principle by focussing on improving usability at the cost of security. *Manual comparison*, for example, requires users to confirm whether two digests match or otherwise, thereby relying solely on users' motivation to pay attention to the comparison. Relying on users' motivation to compare digests impacts security negatively as has been shown by other studies of security applications including those reported in previous chapters of this thesis.

The *principle of commitment* requires that users are provided only with partial information that allows them to commit to a final outcome. For example, *manual copying and entering* reveals partial information that users utilise to make a commitment by entering it into other devices. Users at this stage do not know if this information will be accepted by the devices. It is up to these devices to determine whether received information is correct or otherwise. By doing so, users cannot force a device to accept a value that does not match its own digest.

In other cases, users may have partial information well in advance before devices can calculate it. In group associations, for example, users are required to confirm whether the number of participants (devices) displayed is as expected. In this scenario, users have this information even before an association is initiated. The *principle of commitment* requires that, rather than confirming what users know, users are committed to such information. In the group size scenario, users can be prompted to enter the number of devices expected in the association and that information can later be verified by one or more devices. The *principle of commitment* mitigates issues of complacency or lack of motivation.

6.2.2 Principle of unpredictability

When users interact with a system over time, they tend to learn and master how a system can be controlled or used with least effort. The learning process may be deliberate but over time users end up achieving their primary goals without consciously engaging with the subtasks involved. This phenomenon is known as *habituation* or *user conditioning* [41]. Habituation occurs in activities that have non-changing task sequences and following a specific sequence results in the same outcome every time.

In HISPs, habituation is undesirable for security because users are required to carry out security tasks consciously and accurately. The principle of *unpredictability* simply states that 'a user should not be able to predict the sequence of actions that lead to a particular outcome'. This principle differs from the *principle of commitment* in that it focusses on making a sequence of actions unpredictable while the latter makes it difficult for a user to determine an outcome based on a particular action. For example, in web browser SSL certificate warnings, users know that they have to take one of the two actions (*principle of unpredictability* violated), either accepting or rejecting a certificate and

they also know that accepting results in continuing to the intended website (*principle of commitment* violated).

The *principle of unpredictability* seems to go against usability principles since it appears to force users to always resolve what needs to be done. However, rather than forcing users to resolve what needs to be done, it only ensures that whatever actions a user takes, they are taken consciously rather than by blind selection. Using *manual comparison* as an example, a user knows that there are two actions, one for accepting and another for rejecting a digest (possibly with a reject button always on one side of the screen and accept button on the other). With this in mind a user can accept a digest without even looking at the screen. This principle mitigates habituation.

6.2.3 Principle of single interaction path

Device association is theoretically a simple and straightforward task. In practice, however, users have significant problems that must be acknowledged. A critical issue of concern that users have problems with is the number of interaction paths available to achieve a particular goal. One secure design principle is to ensure that the path of least resistance is the most secure [133]. For example, Firefox 3.x web browser's implementation of allowing users to add exceptions of invalid, expired, or untrusted SSL certificates requires a user to single-click a rejection of the certificate and at least 4 clicks to accept the same certificate [118]. The problem with this approach, however, is that in most instances users' desire to achieve their primary goals (check bank balance, for example) outweighs the amount of effort required to accept a suspicious SSL certificate.

In HISPs, path of least resistance may mean an insecure route — for example, accepting a digest without comparing. Moreover, multiple paths to a single goal are likely to cause users difficulties in understanding an empirical channel. The *principle of single interaction path* demands an implementation that provides a single path from start to end of a device association process.

In Figure 6.1 the arrows indicate the direction of interactions. For example in (A), a user interacts with personal device, then other device, personal device again and so on. Each interaction must be clear to the user in terms of where it starts and ends and there must be clear instructions to inform the user when the next required task occurs at a different device other than the current. In

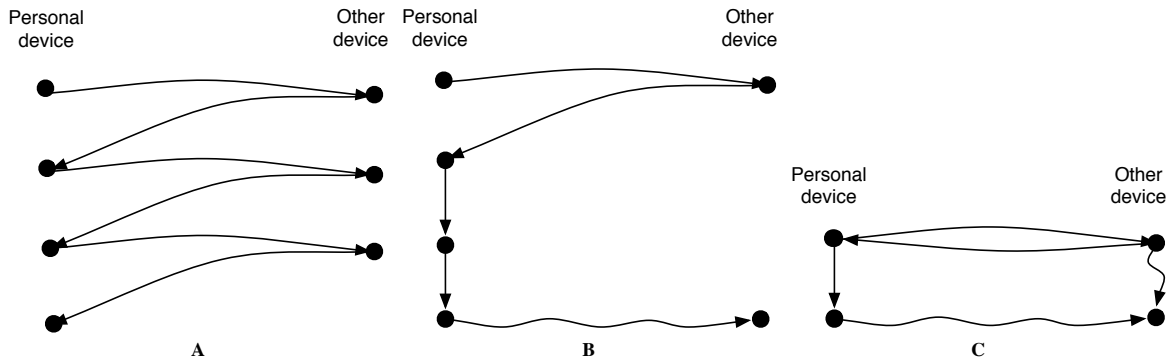

Figure 6.1: Principle of single interaction path: (A) - provide single path. (B) - minimise inter-device interactions. (C) - violation of single interaction path principal

Figure 6.1, (A) shows the general principle, (B) shows a sub principle of single interaction path i.e. minimise inter-device interactions, and (C) represents a violation of the main principle.

Users feel to be more in control when most of the interactions are carried out on their personal devices than otherwise. They also treat personal devices as more trusted than one belonging to somebody else. In HISPs, users complain about interactions that spread across devices and prefer having most interactions (at least those that take time to complete) confined to one device. For example, a participant in the single-user scenario study reported in Chapter 4 commented, *'even though I am quite comfortable using my mobile device, I find it difficult and confusing when I have to switch between devices'*. The sub principle in (B) is intended to reduce the number of times that a user is required to switch attention from one device to another.

Figure 6.1 can better be explained with an example. In *manual comparison*, a user can choose to either compare digests or skip, either act on a personal device or another. In Figure 6.1 (C), the arrows on top point in both directions indicating that a user can start from either device. The arrows pointing downwards on both ends illustrate a possibility of a user ignoring interactions on one end and continuing with the other. It is, however, crucial that if there are multiple paths to an intended outcome, users should find it difficult to change the prescribed path, as also argued in [123].

The *principle of single interaction path* encourages implementing a single path from start to finish of an association, employing inter-device interactions to a minimum, and shifting as much work as possible to one device, preferably a personal device. The principle reduces the chance of both security

and non-security failures (users learn through trial and error) by simplifying user interactions and improving learnability, and puts users in control.

6.2.4 Principle of design by context

It is a well known fact in general software engineering that systems must be designed to work within the context of operation. Designing for specific contexts is crucial because different environments and scenarios have different challenges on a system. In HISPs, for example, single-user and multi-user scenarios have different demands on empirical channels while device affordances limit the type of interactions that users can perform. Moreover, the physical environment such as lighting and noise affects both usability and security of empirical channels.

The *principle of design by context* refers to designing empirical channels within the context of an application in which they may operate and considering what factors may affect their usability and security. This requires thinking about specific user interactions, within a specific application context, that may hinder or help usability and security of empirical channels. For example, during the single-user study reported in Chapter 4, a mock payment system in which participants were asked to type a PIN number for a bank card was used as a primary task for associating devices. During log analysis, we found that there were non-security failures on *manual copying and entering* that seemed odd because participants entered strings that were never close to the digest. A closer look, however, revealed that participants were confused between the screen for PIN entry during payment and one for digest entry.

The above example demonstrates how an interaction within an application can affect usability and security of empirical channels. Following the principle of *design by context* would compel one to consider such an interaction and find ways of avoiding confusing users. In the example above, one may design a screen for digest entry in such a way that it is distinct from the PIN screen or use an empirical channel that does not require typing digests.

Understanding the context in which an empirical channel will operate is crucial to meeting human and contextual needs. There is a general consensus among researchers that device association process must be 'fast' to complete. However, fast usually means user must spend as less time as they consider appropriate. This definition of fast neither provides useful information nor does it define the term

itself. It may, however, be reasonable to think that users are likely to measure the appropriateness of time spent on the association process in relation to the time spent on the primary task. An association, for example, that takes 120 seconds for a primary task that takes 20 seconds is likely to be criticised by users. A participant in our study, for example, commented, *'comparing images just took forever because...'*. 'Forever' for this particular user meant about 60 seconds but this was in relation to a primary task that took 20 seconds.

The principle of *design by context* may be more useful to system designers than researchers. System designers have specific application requirements that must be met. Contextual information can be derived from these requirements and an appropriate empirical channel that meets the needs within that context can be selected. Researchers, on the other hand, may have generic rather than specific contextual information allowing them to develop methods that work across contexts.

6.3 Demonstration of design principles

The effectiveness of the design principles is demonstrated by two proposed empirical channels. This section discusses the proposed methods and how they demonstrate the above principles.

6.3.1 Example 1: Word-matching and number-typing

Word-matching and number-typing is based on the fact that *manual copying and entering* is not subject to security failures but is regarded as difficult to use. While earlier work has argued that typing short strings on devices with limited input interfaces is hard for most users, the popularity of Short Message Service (SMS) is an indication that users are comfortable with such a task. We are, however, cognisant of the lack of motivation from users to type strings for the sake of security. *Word-matching and number-typing* is, therefore, aimed at offering the same level of security as *manual copying and entering* while requiring users to type a smaller number of digits without negating security.

Word-matching and number-typing uses two locally stored dictionaries of words that are phonetically distant [50]. Associating devices generate words representing a digest using a method similar to the S/Key one time password [42] method. One device displays a set of words with numbers assigned

to them in increasing order starting from 1. Other devices randomly display one of the locally generated words and prompt a user to enter the number assigned to that particular word on the other device. The user types the digit and the device displays the next word until all the words have been displayed. The device then checks if the assigned numbers are correct and informs the user of the result. Figure 6.2 shows screen shots of an implementation of the method.

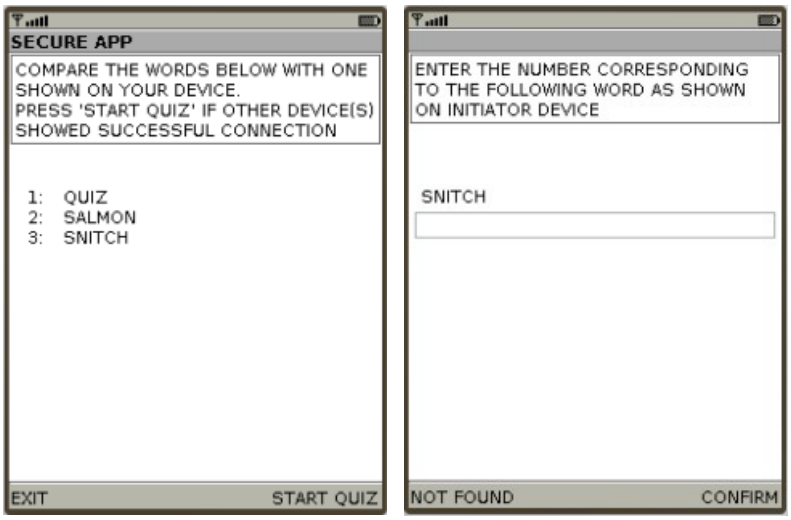

Figure 6.2: Screen shot of Word-matching and number-typing

In comparison with *manual copying and entering*, this method requires a user to type 2 digits for a digest of 20 bits (for 1024 word-dictionaries) while the former requires about 7 digits to achieve the same level of security. This method makes it easier to transfer reasonable amount of information with limited human effort to achieve high levels of security. To transfer 40 bits, for example, *word-matching and number-typing* requires typing 4 digits compared to 12 digits for *manual copying and entering*.

Potential problems to *word-matching and number-typing* include display of duplicate words and users predicting the next digit to be typed (specifically for the last word on the list). The first problem is countered by using two dictionaries, similar to PGPfone proposal [50]. This ensures that when two consecutive bit sequences are similar, two different words will be produced: the first bit sequence from the first dictionary and the second bit sequence from the second dictionary. In current protocols, two words are sufficient. The second problem is due to the fact that once users have entered the first word, they know that only one other word remains and hence will have no

motivation to check if that word actually exits. To counter this threat, an extra word is displayed together with a digest but devices ask users to enter positions of words that correspond to the digest only.

6.3.1.1 Design principles used

- Principle of commitment: users commit to an outcome by entering digits without knowing the values that a device expects. Once values have been entered, the outcome is dependant on the device rather than on users' action. Users are compelled to enter values accurately since they cannot predict the expected values and any complacency will result in failure of association.
- Principle of unpredictability: *word matching and number typing* counters predictability by adding an extra word to the digest, that is, for n number of words displayed users only enter $n-1$ digits. This way, users have no way of predicting the next input digit. In addition, each word is randomly chosen for display on the screen; ensuring that the chance of having words displayed in the same sequence on two consecutive associations is minimised.
- Principle of single interaction path: other than observing values displayed by one device, a user will always type digits on one device (preferably personal device). A message alerting a user for success or failure of association is also displayed on the device where digits are typed. This ensures that interactions are concentrated on one device, rather than being spread across devices, and maintain a single interaction path of execution.

6.3.2 Example 2: Repeated numeric comparison

Manual comparison lacks the ability to compel users to compare digests accurately. Similar to pop-up boxes prompting users to either accept an 'invalid' SSL certificate or not, users will discover that, by pressing a button that indicates that digests match, they will be able to accomplish their primary task.

To compel users to perform comparison without undue effort, *repeated numeric comparison* is a two step process. In addition to a digest, an authenticating device generates a random string of a similar format as the digest. The authenticating device then randomly chooses to display either its digest or the random value. Users compare and indicate whether the string displayed on the authenticating

device matches that on the other device. An authenticating device then displays the remaining string and users compare again.

There are 4 possible outcomes from the user's input: indicating MM (Matching-Matching) for both strings, DD (Different-Different) for both strings, DM, and MD. The authenticating device will only accept one input; M for a digest value and D for random string. A user, however, does not know which one is the digest and which one is the random string, hence, is forced to pay attention when comparing to avoid safe failures.

6.3.2.1 Design principles used

- Principle of commitment: users commit to a final outcome, without knowing the digest that a device has, by choosing what they think is the digest. The outcome depends on the device's value rather than users' input. Users are motivated to compare accurately to ensure that the primary goal is achieved.
- Principle of unpredictability: comparing two strings consecutively without knowing in advance which one is the digest forces a user to pay attention to the task. In addition, a device displays the two strings in a random order which eliminates the possibility of users determining in advance when a digest is to be displayed.
- Principle of single interaction path: similar to *word-matching and number-typing*, *repeated numeric comparison* allows users to carry out all the tasks, other than reading off a digest value, on one device.

6.4 Usability evaluation of the empirical channels

6.4.1 Experimental design

A repeated measure design, using counterbalancing, with 28 participants (12 male and 16 female) of varying age, education and professional background was used. Participants were recruited via web advertisement and mailing lists. They were asked to participate in a quiz, using two mobile phones, where they were first required to establish a secure connection. The quiz application and

device association interfaces were developed using Java 2 Micro Edition (J2ME). The application also logged participants' actions. Other data were collected through questionnaires and interviews. Participants filled in After Scenario, for each method, and End of Experiment Questionnaires. Using the End of Experiment Questionnaire, participants indicated which of the two methods they thought was easy and which one they felt was difficult, were used. In line with previous studies, we also asked participants which method they preferred between the two.

6.4.2 Results

Word-matching and number-typing had mean completion time of 12.7 seconds and 3.6% non-security failures while *repeated numeric comparison* had mean completion time of 13.4 seconds and 7% non-security failures. Similar to *manual copying and entering*, security failures are difficult to simulate for both methods.

A statistical test using paired t-test showed no statistical significant difference between the methods in completion times at 95% confidence interval with $t(55) = .53$, $p = .598$. In terms of user ratings, 93% of participants indicated that *word-matching and number-typing* is ease to use compared to 89% for *repeated numeric comparison*. In addition, 57% of participants preferred *word-matching and number-typing* compared to 25% who preferred *repeated numeric comparison*.

Participants also rated each method on the three main elements of usability — efficiency, effectiveness, and satisfaction — on a 7 point scale. For each participant, scores were summed and averaged for each method. As this data is ordinal, only median, mode, and a Wilcoxon test statistic are reported. *Word-matching and number-typing* had a mode of 7 and median of 6.5 (min = 2, max = 7) while *repeated numeric comparison* had a score of 7 (min = 4, max = 7) for both the mode and median. A Wilcoxon test showed that there is no statistical significant difference in ratings between the methods ($Z = -0.275$ and $p(2\text{-tailed}) = .78$).

In the post study discussion, participants indicated that the methods were similar in terms of ease-of-use. Preferences were just for one's 'taste' and not to do with how the method is used. There was, however, concern over *repeated numeric comparison* from two participants who are dyslexic — a condition that affects one's ability to deal with digits.

6.4.3 Comparison with earlier methods

The above results were compared with those of Uzun *et al.* [123] and our single-user scenario study. A common result in these two studies is that *manual comparison* had the lowest completion time and ranked as the most usable. *Manual copying and entering* was found as the most preferred (personal choice) in Uzun's work while *manual comparison* was found to be the most preferred in our work. The results were compared on three crucial issues: efficiency (completion times), satisfaction, and security.

Efficiency: With an additional step in *repeated numeric comparison*, it was not expected that participants would complete an association process using the method in a time comparable to previous studies. On the contrary, participants completed the process with an average time of 13.4 seconds compared to 16.4 seconds reported in Uzun's work. In addition, while participants had to look up a word on the list before typing its position (in *word-matching and number-typing*), average completion time was 12.7 seconds compared to 13 seconds in Uzun's work.

User rating: We found that the proposed methods were rated higher in terms of ease-of-use, 93% for *word-matching and number-typing* and 83% for *repeated numeric comparison*, compared to *manual comparison's* 40-45% and *manual copying and entering's* 20-37% in previous studies. On a rating scale of 1 (worst) to 7 (best), the above methods were rated above 6.5 compared to earlier results where *manual comparison* and *manual copying and entering* were rated below 6.3.

Security: Unlike *manual comparison*, the proposed methods are resistant to security failures. While *manual copying and entering* is resistant to security failures as well, it does not offer room for larger digests — typing 6 digits is already deemed difficult to use — hence limited in its application. *Word-matching and number-typing* offers room for larger digests than currently possible by both *manual comparison* and *manual copying and entering*. For example, for the same amount of human effort (typing 6 digits), *word-matching and number-typing* can offer security of 2^{60} entropy while *manual copying and entering* can only offer 2^{20}.

6.4.4 Applications scenarios of proposed methods

Previously proposed methods are either insecure, difficult to use, or limited to a single scenario and application context. This section discusses the scenarios to which the above proposed methods can apply and to demonstrate that they can work across contexts.

- **Close/distant devices:** One of the contextual factors discussed in the framework presented in Chapter 7 is proximity of devices — devices could be close or distant. With devices close to each other, users can directly read digests off the screen of a device displaying it. In a scenario where devices are distant, a number of channels including audio/video and Short Messaging Service (SMS) can be used to exchange digests among participants.

- **Input/output constrained devices:** Common device association scenarios involve a device with rich input/output interfaces and an input/output constrained device. For both methods, a display and a single button is sufficient for interface constrained devices. In *word-matching and number-typing*, a device with rich interfaces will display 3 words and a user will be required to press the button 3 times consecutively to enter the numeric value 3, for example. Since 2 words provide sufficient entropy for security, users will not be required to enter numbers higher than 3. Similarly, in *repeated numeric comparison* users may be required to press the button in a particular way, a 3 seconds press for example, to accept that digests match and a short press to reject.

- **Group associations:** In a group association scenario, e.g. meeting, a user with a device displaying a digest can read it out to other users in the group. In *word-matching and number-typing*, words in the dictionary should be phonetically distant to avoid confusion between similar sounding words. For both methods, strings must be meaningful and human readable so that one user can easily read a string for others.

6.5 Summary and conclusion

Human-Interactive Security Protocols require users to perform security critical tasks to establish secure device association. While device association is not usually a primary task for users, security critical tasks must be carried out correctly. Given users' lack of motivation and the demands of

competing tasks, empirical channels must be designed such that users do not compromise the desired security.

In this chapter, we proposed principles for designing secure and usable empirical channels. The *principle of commitment* ensures that users do not compromise security by disclosing only partial information that can be employed to commit to a final outcome. The *principle of unpredictability* achieves the same goal by preventing users from predicting required action. To improve usability, thereby security, the *principle of single interaction path* requires implementations that provide a single ordered sequence of actions that users follow every time to achieve the intended goal. In addition, tasks must be confined to one device to minimise the number of inter-device interactions. In contrast to other principles that focus on empirical channels, *design by context* focusses on the context and application in which a method is applied.

We demonstrated the effectiveness of the design principles by proposing two empirical channels. The methods were designed to satisfy the first three design principles. The last principle, *deign by context*, requires contextual information that is dependent on a specific application. The proposed methods were evaluated for usability and security using a laboratory experiment and results compared to those of previously proposed methods. The results showed the, by following the principles, it is possible to design methods that are effectively secure.

We also discussed scenarios to which the proposed methods can be applied. The discussion highlighted the fact that it is possible to design methods that are secure, usable, and applicable to a wide range of scenarios. Unlike previously proposed methods that are either insecure, difficult to use, or limited to a single-user scenarios, proposed methods ensure that users carry out security critical tasks accurately without demanding undue effort.

Chapter 7

HISPs framework for reasoning about empirical channels

7.1 Introduction

In this chapter we present a framework for developing or choosing empirical channels that suit security, human, and contextual needs. We discuss how the HISPs framework can be applied and give examples, through a case study, of how it can be used.

The HISPs framework is designed to help both researchers and system designers in understanding and reasoning about different factors that are crucial to developing or choosing effectively secure empirical channels. It puts these factors into different contexts in which HISPs may operate and, hence, provide a better understanding of the elements that pose challenges to security and usability of empirical channels. System designers can use the HISPs framework to reason about a specific system within the context in which it operates and help them choose, among existing methods, an empirical channel that suits their application environment. On the other hand, researchers can use the framework to understand and reason about different factors to develop empirical channels that apply to a wide range of scenarios.

The HISPs framework builds on previous work including Cranor's [18] — *The human-in-the-loop security framework*. The human-in-the-loop framework was designed (based on the simple communication-processing model) to help understand human behaviour in performing security-critical functions. Whilst the *human-in-the-loop* framework provides insights into human behaviour in the wide context of security, the HISPs framework is a tool for reasoning about the different factors specific to the mobile and *ad hoc* device association environment. It is aimed at helping in the designing or choosing of empirical channels that suit both human and contextual needs while fulfilling security requirements.

The HISPs framework also builds on the findings of various Human-Computer Interactions Security (HCISec) user studies and a review of mobile device *ad hoc* interactions together with currently proposed empirical channels. HCISec research has revealed users' experience of varying levels of difficulty when using secure systems [97, 129]; specifically, studies on passwords show that users experience difficulties with systems that demand their attention, memory, and accuracy. While these difficulties are better understood in the context of secure systems in fixed computing environments such as desktop computers, significant other factors put further demands on users in the mobile environment. These factors were explored, by examining HISPs task requirements together their demands on users, and incorporated in the HISPs framework.

The HISPs framework was validated and refined using expert knowledge. Expert validation has been used as a methodology to validate research results. For example, Brostoff [12] used the methodology to validate a model for password security and usability. It entails engaging experts in a specific domain to critique a proposal. This critique is then used to refine the proposal. In the absence of a real world case study, due to non deployment of HISPs, an initial design of the framework was circulated to experts in HCISec to provide feedback on the proposal. This feedback was then used to refine the proposal.

This chapter is organised as follows; in Section 7.2 we present the HISPs framework. We follow on with a discussion of how the framework can be applied in Section 7.3, and present an example of its application in Section 7.4. We discuss the expert validation of the framework in Section 7.5 and summarise and conclude the chapter in Section 7.6.

7.2 HISPs framework

The HISPs framework (Figure 7.1) consists of three elements: technical and contextual factors, human factors, and empirical channels. Technical and contextual factors relate to technical security requirements, association scenarios, modes of authentication, environment, number of users, device proximity, and device affordances. These factors directly affect empirical channels. For example, technical security (size of digest) imposes constraints on empirical channels — the size of the digest determines what empirical channels are practical in a given scenario.

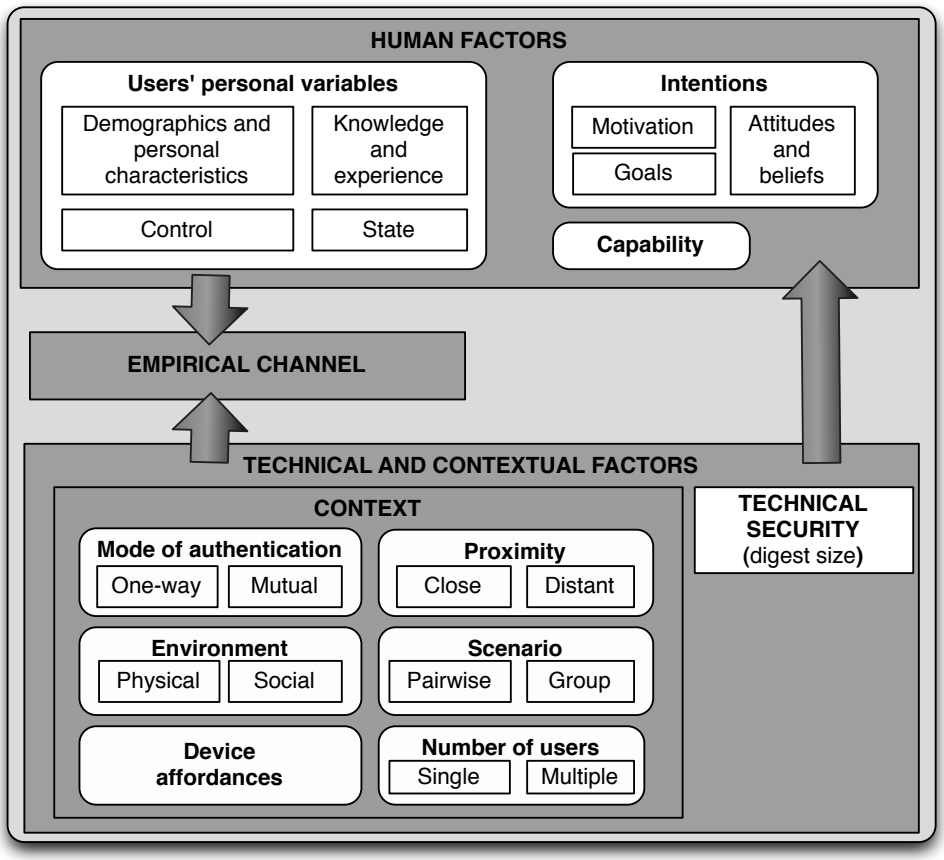

Figure 7.1: HISPs framework

In addition, they have direct effects on users in relation to specific human factors. Device affordances, for example, affect human performance when performing security tasks in device association. Human factors also impose constraints on the choice of an empirical channel or how it is designed. An empirical channel that requires touching devices simultaneously, for example, may be inappropriate if users are incapable of doing so because devices are distant. To develop or choose empirical

channels that are effectively secure, researchers and designers must consider all of human, technical, and contextual factors.

7.2.1 Technical and contextual factors

Secure systems are socio-technical — they operate in concert with other systems and are used by humans. Such systems, therefore, must be secure at the technical as well as social level. While computers have the ability to deal with complexity and repetitive tasks producing the same result, humans face significant challenges in dealing with such tasks. Consequently, secure systems must be designed with simplicity in mind otherwise complexity can introduce vulnerabilities. These vulnerabilities are introduced because of either unintentional insecure user actions or difficulty-of-use which causes users to abandon a secure system altogether and resort to insecure methods.

In addition to technical elements, secure systems operate within specific social and physical environmental contexts. Context has positive and negative effects on both security and usability of a system. For example, poor lighting conditions may impair usability. Understanding how specific contexts affect usability and security of a system is crucial to designing systems that work in a usably secure manner in different scenarios.

7.2.1.1 Technical security

Technical solutions to security problems are the centre of many applications. This is due to emphasis on sophisticated attackers who have access to state-of-the-art tools to attack a system. Examples of technical solutions to security problems include firewalls, encryption, and machine generated passwords assigned to users.

Security requirements are also situational — they depend on circumstances. Technical solutions are usually rigid and require someone with technical expertise to be reconfigured. Unlike in organisational settings, such expertise might be unavailable in both home and mobile *ad hoc* interactions. In HISPs, technical security requirements specify the theoretical security of a protocol based on mathematical proofs. They determine the size of the digest that users should transfer or compare using an empirical channel. All HISPs specify a minimum size of a digest while the maximum is "as large as is necessary". In these protocols, technical security can, therefore, be set to fit security

requirements at hand. Care must be taken, however, to avoid increasing technical security at the expense of usability.

The challenge on technical security is ensuring that technical solutions are used correctly within the context in which they are deployed. The design and implementation of an empirical channel that facilitates secure human interaction with the system is crucial. An empirical channel must be able to adapt to changes in technical security with little or no effect on the usability of a system.

7.2.1.2 Context

Systems, together with their users, operate within context. Context is the extension of a technical system to consider factors outside it. For example, an authentication system may operate in context with external (outside the system) objects such as Closed Circuit Television (CCTV) cameras, humans, and computing devices. These external artefacts directly or indirectly affect how one evaluates a system's security or how users interact with the system. Given the different views on what *context* is, as discussed in Section 2.1, an enumeration approach is taken with respect to the HISPs framework. Enumerating entities that affect user interactions in mobile environments provides necessary information to model and design a secure and usable system.

In *ad hoc* mobile associations, context can be characterised in terms of both the physical and social environments, association scenarios, device proximity, number of users, and mode of authentication. The usability and security of empirical channels may vary significantly between different contexts.

Environment: Mobile device associations take place in varying social (e.g. private verses public crowded spaces) and physical (e.g. light intensity, background noise) environmental contexts. As mobile devices become more pervasive, so do *ad hoc* interactions. For example, for a payment and ticketing application running on a mobile device, it is reasonable to expect *ad hoc* interactions for such a system to range from vending machines on streets, crowded environments such as bars, train and bus stations to online transactions. Each of these example environments has characteristics that hinder or improve the usability and security of empirical channels.

In addition to the physical environment, it has been found that users are governed by social norms and tend to conform to socially acceptable (informally) set of behaviours [32]. These norms and acceptable behaviours virtually govern how humans interact with various artefacts in different en-

vironments. In HISPs, for example, the presence or absence of users not participating in a secure device association is a social variable because users behave differently in either case.

In choosing and designing empirical channels, there is need to consider how different physical and social environmental contexts affect their usability and security. Specific questions at this stage must address concerns of user acceptance of a method, as they interact with a system within the physical and social contexts, and also a method's resistance to security failures or adaptability to different contexts.

Association scenario: *Ad hoc* device associations are either pairwise such as a PDA and a printer or group wise (3 or more devices) such as multiple devices joining a network simultaneously. The distinction between pairwise and group association scenarios warrants one to reason about the scalability of empirical channels with increasing number of devices.

Moreover, security is a mutual achievement of multiple parties. In HISPs, each device must "behave securely" to achieve desired security among all participants. It is, therefore, crucial to think about how an empirical channel can adapt to group scenarios in terms of both security and usability.

Group size dictates what actions are acceptable for users (See Section 5.7). A major concern at this stage is whether a proposed empirical channel can securely and efficiently be used in group scenarios. Specific concerns on group size must focus on whether an empirical channel is usable with a single human user associating multiple devices simultaneously or multiple human users are able to sufficiently share tasks to avoid burdening a single user with all the work.

Device proximity: Secure device associations can be between devices that are in close proximity, using Bluetooth for example. In this scenario, users have access to all devices involved or at least each member is able to see all the devices or their human owners. Secure device associations can also occur between devices that are physically far apart and communicate, say, using the Internet. In this scenario, users have no physical direct access to one another or devices; access is only through an empirical channel which the intruder is unable (or at least finds difficult) to forge.

In both scenarios, an empirical channel must be human-verifiable by providing visual cues or otherwise, to give assurances to users that the devices they wish to associate are the only ones participating in the authentication process. This is consistent with recent calls to make security relevant actions visible to users rather than hiding them. Specific concerns such as the effect of close as well as

distant devices on empirical channels, and how cues are presented to users for verifying that their devices have achieved the required security must be addressed.

Number of human users: The number of users involved in bootstrapping security between mobile devices has implications on the security and usability of empirical channels. User scenarios can be categorised as either single-user, where an individual controls all devices involved, or multi-user, where each device has its own user. In a multi-user scenario, a well designed empirical channel distributes required human effort among participating users and, as such, provides an opportunity for using larger digests (for theoretical security) contrary to scenarios where a single user is expected to do all the work.

As security is a process rather than a product, the number of nodes where security can fail increases with each additional device or device/user pair since the correct behaviour of all participants is necessary to maintain desired security. In HISPs, non malicious participants achieve global security — by sharing a common cryptographic key, for example — among them once they all behave correctly and are diligent in detecting anomalous actions. In order to achieve this, desired user actions must be easier to perform than undesirable ones within the context in which empirical channels are used.

The concerns to be addressed here are: how can we design (or how do we propose) empirical channels that allow for distribution of human effort among participants? How can a single user efficiently establish a secure association among multiple devices with acceptable mental and physical effort? How does increasing the number of devices or device/user pair affect usability and security of a particular empirical channel?

Mode of authentication: Authentication can be categorised as either one-way (asymmetric) or mutual (symmetric). In one-way authentication, one device authenticates one or more participating devices. For example, an Access Point (AP) authenticating mobile devices wanting to access the Internet through it (assuming the AP is configured to authenticate devices). In this example, a user may be happy to identify an AP by name (if they know it) or by other means. In short, a user conducts a *weaker* authentication of the AP. The AP, on the other hand, enforces a stronger authentication in which it may prompt the user to transfer some information, using an empirical channel, to verify that the user of the device knows some secret that the AP has and hence (presumably) has access rights to it. In mutual authentication, however, each of the participating devices authenticates all other devices. In the AP example, a user or a personal device may require more

than just a name of the AP. The device may demand that the AP generates a string that the user can verify.

Both of these modes of authentication pose different usability challenges. In one-way authentication, an authenticating device's acceptance of an association request is good enough for the authenticated device. In practice, the AP may require the user to transfer some information to the device and no further action from the user is needed. In mutual authentication, a user may be obligated to take extra steps. The user may be required to read the AP's response (refusal or acceptance of the connection) and indicate appropriately on the connecting device. The amount of human effort expended in mutual authentication may be double that expended in one-way authentication. For example, using a 2D barcode to encode a digest, the barcode must be captured $n - 1$ times for one-way authentication and $n(n - 1)$ times for mutual authentication where n is the number of devices participating in the association. Understanding this difference in human effort between the two scenarios is invaluable to developing or choosing secure and usable empirical channels.

In addition to increasing human effort, the extra step in mutual authentication is a user action that can result in security failures. For example, a user misinterpreting a refusal by the AP as an acceptance of the association can result in associating a personal device with an unintended AP or the user may interpret the message on the AP correctly as a refusal but fail to indicate accordingly on the device. These usability and security challenges that one-way and mutual authentication pose to empirical channels must be analysed and addressed.

Device affordances: Affordances are the means through which a user can interact with a device [16]. They provide a means by which users can input information or instructions, and how they can receive feedback. For example, a mobile phone may have a keypad and a camera — as a means through which a user can pass information or commands to it — a display and a speaker — through which a user can get feedback from it.

In the literature on empirical channels, assumptions on device affordances have been pronounced but they are too restrictive because they consider only devices of similar affordances. Mobile devices may not always connect to other devices of similar type or affordances. Examples where this is the case include connecting a mobile device to a printer, vending machine, or AP. This does not imply looking at all possible devices (and their affordances) as this is impractical for a simple reason that devices (and affordances) are continually being invented. Devices, however, can be categorised and

grouped as mobile, stationary, keypad and display, keypad only, and so on, to help assess and analyse the impact that associations of devices from different groups may have on usability and security of empirical channels.

7.2.2 Human factors

Users are stakeholders, and form the social component of a socio-technical system. Nevertheless, they have personal requirements such as privacy, usefulness of system and easy-of-use on the systems they use. Though they may be aware of their security needs, they are usually unmotivated because security is not their primary task in most applications. These personal requirements form the *human factors* of the HISPs framework.

Humans are constantly making security decisions; consciously or unconsciously. Different users may make different security decisions under similar circumstances. Those who are naturally risk averse take less risks in the digital world compared to those who are not. Similarly, those who have been exposed to certain risks tend to be risk-averse towards such risks [108]. In addition to risk aversion and exposure to a particular risk, humans make security decisions based on a number of other factors as discussed below.

7.2.2.1 Personal variables

HCISec focusses on designing systems that are secure and usable by considering users and their characteristics. This is the main basis of design approaches such as User Centred Design [95], AEGIS [34], and participatory design [109]; they emphasise on putting users at the centre of the design and understanding their needs and characteristics.

Users' security decision making process can be influenced by their knowledge and experience of a particular system including the context in which it is made. They may misconceive the risks they are exposed to by either underestimating the risk — in which case security decisions expose them to the risk — or overestimating the risk — in which case they feel there is nothing they can do to protect themselves.

Despite misconceiving the risks they face, users are a social countermeasure for every dimension of security: prevention, detection, deterrence, and reaction [32]. To be an effective social countermeasure, they need to attain a level of control in the system they use and protect. This can only be achieved if systems are designed with target users in mind and if these users can understand and use the system as intended.

In addition to making security decisions based on knowledge and experience, users operate under different emotional states at different times and in different situations. Though the terms *emotion, feeling*, and *mood* are a source of argument in social science literature, the meaning here is not restricted to one or the other. In this context, examples of emotional states may be excitement, stress, anxiety, concentration or lack of it, and tiredness. Emotional states can change due to peer pressure, time constraints, loss or gain of something. Users, even though presented with all the information they need and have the knowledge and experience to make the right security decision, may make an incorrect one because of the emotional state in which they are. In usability and security studies, however, quantifying how emotional states may impact HISPs is difficult but what is crucial is reasoning about these factors for different scenarios and contexts.

Personal variables influence how one makes security decisions. A number of questions need to be addressed; What is the target population? Does it have the knowledge and experience to make the correct security decision in the security application? Are users empowered and feel in control to make the right decision? How do emotional states affect their security decision making? Can we design empirical channels that are secure and usable across personal variables?

7.2.2.2 Intentions

The effect of intentions on one's behaviour has long been studied in social science studies. Intentions define an individual's willingness to carry out a particular behaviour but they are influenced by one's attitude towards that behaviour as well as the subjective norms (opinions of others and motivation to comply with those opinions) [44]. An individual's attitude towards security has an effect on how one interacts with a secure system.

Users generally display the unmotivated user property when security is orthogonal to the task at hand and, because an individual's intention to show a particular behaviour is affected by the motivation to comply to subjective norms, motivation plays an important role in how users decide what action

they engage into. In addition to motivation, attitudes about a behaviour based on beliefs about and evaluations of that behaviour affect the intention to conduct the behaviour [44]. For example, Weirich and Sasse [126] found that users' lack of compliance with security policies was because of beliefs and attitudes that the risk was not real and that their behaviour was insignificant even when the risk was real. Moreover, security-critical tasks must be aligned with user goals in order to help and not deviate or hinder users from accomplishing the tasks at hand.

7.2.2.3 Capability

Before users adopt new technology, they must perceive it to have ease-of-use [20]; they must believe that they are capable of achieving a required behaviour in order to engage in that behaviour. Preserving security is no exception to this — users must have the capability to use a system in a secure manner without undue effort. Capability may be physical, mental, or technological. Assessing whether a target population is capable of carrying out security-critical tasks is an essential consideration in developing or choosing empirical channels. For example, proposals based on stimulus-response (timing methods, see Section 2.3.5) in order to establish a secure association are only feasible if users can conduct the required action. However, users may not have the physical capability to synchronise multiple devices simultaneously. These factors need to be put into perspective when developing or choosing empirical channels.

7.2.3 Empirical channels

An empirical channel must be evaluated against requirements derived from the technical, contextual, and human factors. Typical questions must focus on security (is a method secure against human mistakes?), scalability (can the size of the digest be varied without significantly affecting usability or security or both?), adaptation (is a method capable of adapting to different physical and social contexts in which it will be applied?), and fit for purpose (does a method fit the tasks within the contexts in which they are carried out?).

An evaluation of methods against relevant requirements is crucial to ensuring that the broad aspects of context, technical security, and human factors are considered. System designers may focus on existing empirical channels while researchers may be interested in developing new ones. In either

case, an empirical channel can be successful only upon critical consideration of all factors that may affect its usability and security, and asking questions that are relevant to both (usability and security) is important.

7.3 Application of HISPs framework

The application of the HISPs framework may differ depending on whether one is a researcher or a system designer. A system designer will have a specific application from which context and technical security are derived. By using the HISPs framework, a system designer will identify human factors that are crucial to the specific use context and identify candidate empirical channels that are likely to support those factors. Researchers may want to reason about empirical channels and the factors that affect them from a wider perspective in order to develop channels that are scalable, usable, and secure across multiple scenarios and contexts. Nevertheless, to develop such methods, it is crucial for researchers to examine specific application scenarios and analyse how a proposed method can be affected by such scenarios.

The goal and approach in which the HISPs framework may be used by researchers and system designers may differ, as previously mentioned. However, the process is similar in both cases; both researchers and designers need to have a target application in mind. For the designer, this may be very specific while researchers are likely to have wider concerns (or focus on specific application domains) than a single application unless it is a specialist application. Having considered a specific application or domain, the HISPs framework may be used to develop or choose an empirical channel.

In order to evaluate a proposed empirical channel, a usability study may be conducted, possibly repeatedly cover all application scenarios and use contexts. If the results of the study are acceptable (i.e. they meet expectations) the method is accepted, otherwise the HISPs framework may be used again. The HISPs framework can be used to propose empirical channels or reason about how a particular channel may be improved. As such, the HISPs framework may be used on methods whose usability study results do not meet expectations to identify areas of improvement rather than proposing a different one.

The HISPs framework fits into the User-Centred Design (UCD) process [19]. UCD (Figure 7.2) is a 3-step process: Analysis, Design, and Evaluation. During the analysis phase, user, task, environmental,

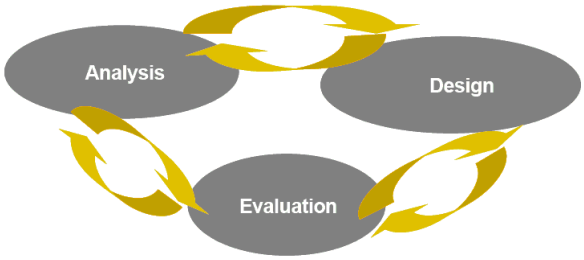

Figure 7.2: User-Centered Design process

and comparative analyses are conducted. It is during the analysis phase that the HISPs framework is proposed to be applied and, based on the outcome, an empirical channel may be proposed. Once an empirical channel is proposed, it must be evaluated by conducting a usability study (or through other means). The outcome of the evaluation stage is used as feedback that may be used to improve the proposed method, propose a different method, or accept the method in its current state.

7.4 Application of HISPs framework: example

To demonstrate how a specific application may be analysed (from the perspective of a systems designer) using the HISPs framework, we consider a meeting application in which participants wish to exchange confidential files. The following initial assumptions about the meeting application are made; a meeting is held by two or more individuals (and may be as large as tens or hundreds) who may or may not be in the same location but wish to share confidential files using mobile devices such as laptops, PDAs, smart-phones etc. When participants are in the same location (i.e. room), they use location limited channels such as Bluetooth to create a local area network for their devices, otherwise they use the Internet and create a secure network. In either case, there is one or more ways in which an empirical channel is implemented. This could be via multimedia channels such as audio and video channels as well as face-to-face conversations. There is a leader who controls membership and manages the meeting, and the target population consists of business executives. Participants may be from the same organisation or from different organisations. In either case, the sensitivity of the information requires that no long term keys are used due to fear that they may be compromised. Under these assumptions, the meeting application is analysed using the HISPs framework and propose an empirical channel that is likely to suit this application. For this example,

only *Manual Comparison* (MC), *Manual Copying and Entering* (MCE), *Auxiliary Device Methods* (ADM), *Timing Methods* (TM), and *Short-range Directed Channel* (SDC) are considered.

7.4.1 Context

Mode of authentication: A leader, who controls membership, authenticates participants resulting in an asymmetric (one-way) authentication. Each participant will present credentials to the leader who verifies the information over the empirical channel. Without considering other factors that are part of the meeting application, all empirical channels are suitable for asymmetric authentication.

Environment: For this specific application, one can assume that meetings take place in quiet and non-crowded environments considering the confidentiality requirements of the information that needs exchanging. Under this assumption, all empirical channels under consideration are eligible.

Device affordances: Devices used in meetings will have basic affordances such as display, keypad, speaker, and microphone. The size of displays may vary from one device to another. For example, a laptop will have a larger display than a smart-phone. The keypad may also vary in size as well as features depending on whether it is a multi-touch or a full qwerty keyboard. Different devices may also have speakers that vary in the quality of sound they produce while recording using microphones may produce varying quality levels of sound. We do not, however, assume that devices have special affordances such as cameras, lasers, accelerometers, Infra-red, or standardised ports. This excludes methods that rely on these affordances.

Proximity: Traditional meetings take the form of face-to-face interactions where participants are in a single room. For such interactions, all empirical channels are feasible when no other factors are considered. Globalisation, however, has changed how business is conducted and how meetings are held. Meetings may be held by participants from different organisations or the same organisation but from different branches that are in different cities, countries, or continents. In this scenario, auxiliary device methods, timing methods, and short-range directed channels are infeasible.

Association scenario: Meetings may be pairwise or group associations. An empirical channel for a meeting application should be one that works across different group sizes without increasing human effort. Timing and short-range directed channels are convenient for pairwise associations because

they require synchronisation and are point-to-point. While auxiliary device methods can work for group associations, human effort increases with the number of devices involved.

Number of participants: The number of participants can determine which empirical channels are feasible and which ones are not. For example, a meeting with more than 10 participants may require an empirical channel that allows the leader to broadcast messages to all participants and also allows participants to respond to the leader. In order to determine an empirical channel that works across contexts for the meeting application, scenarios where the number of participants is greater than 2 are considered. In such cases, short-range directed channels, auxiliary methods, and timing methods are impractical.

7.4.2 Technical security

Technical security is determined by the mathematical assertions of a specific protocol. In the meeting example, digests of between 16 and 32 bits are required — based on the SHCBK protocol [79]. This protocol is chosen on the basis that it is efficient (requires small digest size) and that it supports group associations as required in a meeting application. All the empirical channels under consideration can accommodate digests of up to 32 bits.

7.4.3 Human factors

Personal variables: To understand the context in which empirical channels may operate in a meeting application, one needs to draw general assumptions about personal variables of the target population. In this case, the target population represents a highly educated group, which is generally competent with using mobile devices such as laptops, PDAs, smart-phones, etc. Whilst they are highly motivated individuals, they may lack the time to pay particular attention to tasks such as device association that may be deemed secondary. This requires empirical channels that provide the required security without demanding undue attention and also that they do not take longer than necessary to complete an association. Given that human effort increases with the number of devices involved in device association, SDC and ADM are undesirable methods in this application.

Intentions: The users under consideration here require technical solutions that help them achieve their primary goals without causing unnecessary delays and work load. In short, time is of essence

for these users. They are likely to be aware of the security requirements of the information they want to share but their beliefs of the seriousness of the risk may cause them to pay less attention to the association process especially when it requires unnecessary amounts of time and effort. It is essential, considering these factors, that empirical channels do not introduce complacency by not compelling users to consciously execute the association process.

Capability: Given the context, technical security, and human factors already discussed, are users capable of securely carrying out the demands of a particular empirical channel with acceptable effort? Users' capabilities are limited by technological factors such as device affordances and constraints of an empirical channel. For example, users are incapable of using auxiliary device methods if their devices do not provide mechanisms for such or cannot use SDC when users are not in the same room. Users of the meeting application are, therefore, incapable of using ADM and SDC. They are also incapable of using timing methods because it is unreasonable to assume that synchronisation is possible for tens or hundreds of devices that may even be in different locations.

7.4.4 Empirical channel

	Scenarios	**Meeting application scenario**	**Candidate empirical channel**
Context	Association scenario	Group	MC, MCE, ADM
	Proximity	Distant	MC, MCE
	# of users	3+	MC, MCE
	Authentication	Asymmetric	MC, MCE
	Affordances	Display, Keypad, microphone, speaker	MC, MCE
	Environment	Private, quiet	MC, MCE
Technical security	Digest size	32 bits	MC, MCE
Human factors	Personal variables	Company executives, educated, competent with mobile device, time constrained	MCE
	Intentions	Exchange information	MCE
	Capability	Enter text, compare digests, touch devices	MCE

Table 7.1: Summary of scenarios showing candidate empirical channels

Table 7.1 summarises the meeting application scenarios, together with candidate empirical channels, directly related to the application. Choosing an empirical channel depends on three factors. First, an empirical channel must be applicable to all scenarios of the application at hand. Second, an empirical

channel must be resistant to security failures. The user population in the meeting application is one that has limited or no time for non-primary tasks. Selecting an empirical channel that can easily cause security failures at the press of button is highly undesirable. Third, as mentioned earlier, device association is no primary goal — only an enabling task — hence, it should not take time that may be deemed unnecessary by users. Specifically, an empirical channel must allow for simultaneous exchange of information in a group. Methods such as auxiliary devices and short-range directed channels are more appropriate for pairwise rather than group associations since these channels are directed and unidirectional.

7.4.5 Recommendation of empirical channel

Having presented recommended empirical channels for each scenario and discussed factors necessary for choosing an empirical channel — based on the HISPs framework — for the meeting application, MCE is recommended as the method of choice. First, MCE (as well as MC) is applicable to all scenarios of the meeting application (see Table 7.1). However, the definiteness of the user action in MC makes the method vulnerable to security failures. Moreover, the busy nature of business executives may mean that less attention is paid to comparing while an increase in the group size will mean an increase in the number of nodes where security may potentially fail. Though MC may take less time than other methods, the time difference can be traded for the security guarantees gained in MCE. Based on this analysis, MCE is best candidate for the meeting application.

7.4.6 Discussion

The above analysis and resulting recommendations are based on available empirical data on the performance of existing empirical channels (See Chapters 4 and 5). In cases where performance data is unavailable, a user experiment of candidate empirical channels must be conducted. Moreover, within a specific category of empirical channels, different methods must be evaluated. In MCE, for example, numeric, alphanumeric, or phrases may be evaluated to choose one or determine which method is suitable for different scenarios.

7.5 Expert validation of HISPs framework

The HISPs framework was validated using experts in HCISec from both academia and industry. We define experts as researchers who have worked in the areas of security and usability and have actively participated in the advancement of these research areas either through publishing or by being committee members of conferences or workshops. Table 7.2 lists the experts who participated in the validation together with a summary of research areas they work in.

Name	Profile
Expert A	Professor and Head of Information Security Research at a renowned university.
Expert B	Professor whose main research interests are privacy, security, and usability issues.
Expert C	Professor whose main research interests are context awareness and security
Expert D	Principal scientist in usable security at a mobile device manufacturing firm
Expert E	Works in the area of designing secure user interfaces and enhancing and enabling secure user experience and behaviour.
Expert F	Specialises in human aspects of information security.
Expert G	Leads a number of projects focusing on security and usability.
Expert H	Professor whose main areas of research include privacy, trust, technology use in public places, the impact of age and disability on technology use.

Table 7.2: Experts' areas of specialisation

We sent each expert a paper describing the framework and its application to reasoning about empirical channels. We also sent each expert a proforma (provided in Appendix B) to be completed. Experts were asked to provide feedback on five criteria: strengths of the framework, academic contribution of the proposal, practicality of the framework, weaknesses of the proposal, improvements, and future direction. In addition, we provided an extra row in the proforma for any other comments that experts would want to provide in addition to those on the five criteria.

In analysing expert comments, evaluation criteria of strengths, contribution, and practicality were grouped as benefits of the framework. We provide a snapshot of the benefits and identify key phrases that summarise the comments from the snapshot in Section 7.5.1. We then list the weaknesses (critique of the framework) and discuss how we have responded or addressed them in Section 7.5.2. We have taken experts' proposals for improvements together with pointers for future direction as future work since these will normally require real world studies which we cannot conduct at the moment due to non-deployment of HISPs.

7.5.1 Benefits

Each expert was sent the same document describing the HISPs framework and were asked to critique the proposed framework. To give experts the liberty to express themselves without reservations, they were asked to indicate whether they wanted their comments to be kept anonymous or not. We first summarise in Table 7.3 expert comments on the benefits of the framework and follow on with detailed comments. Each benefit is listed with the number of the quote(s) (provided in Table 7.4) from which it is derived.

Benefit	Quote #
1. Comprehensive	1, 3, 4, 9, 10
2. Places users/human factors into focus	2
3. Adaptive to different scenarios and devices	2, 6
4. Thorough	7
5. Practical	4, 6
6. Applies HCI to a security problem	2, 6, 7, 8
7. Valuable to device manufacturers	6

Table 7.3: Summarised expert feedback - strengths, contribution, and practicality

7.5.1.1 Detailed comments

Table 7.4 is a snapshot of expert comments on the strengths, practicality, and contribution of the framework (a full transcript of expert comments is in Appendix A).

7.5.2 Criticisms

A summary of each of the expert criticisms is provided below. Underneath each criticism is the quote from which the summary is derived followed by a discussion of its merits.

1. Considers all channels/threats to security e.g. technology, user
2. Taking into account the mental load of users performing an authentication procedure. Placing users/human factors in the focus. Talking about social norms and including it in the environment part of the framework - I would like to see this expanded, as it seems highly relevant. Multi-user scenarios might pose novel issues, so having them in the model is good.
3. The framework offers a comprehensive enumeration of contextual factors that may influence the effectiveness of a human interactive security protocol (HISP). It offers designers and researchers a concise checklist of issues that should be considered during the design and/or evaluation of a HISP.
4. Practical for both [researchers and designers] as highlights issues that need to be considered from different groups
5. The work would seem to be useful for both researchers and designers: for researchers, there could be a deeper understanding of the underlying issues and how the current methodologies work; for designers, the work could provide tools and guidelines to implement more usable security. The mobility aspect has been under researched in current body of work, so adding to that corpus of research and hopefully some guidelines will benefit both communities
6. A major strength is application of the UCD method to a security problem. This has not been done before. Frameworks so far have been very abstract, and actual research results have been point solutions for a specific problem in a specific scenario. The work has great potential for being used in practice. Device manufacturers have already problems setting up secure usable pairing between two devices, let alone between many devices. Id like to note that the background work is very thorough
7. The proposal is based on a solid understanding of the current problems in how security is offered to end users. It aims at improvements on both the underlying technical solutions as well as on the user interface level; a combined approach is probably the best way to go. If the underlying technical solutions are crafted with an understanding of the cognitive issues that effect their understandability and easiness of use in short, the human aspects in order to reflect on these issues, the better. The framework aims to bring together and add to the existing body of work on understanding the related issues in secure device pairing and its usability aspects
8. The proposal sets up a few questions on the usability issues of the research problem that if answered in a good way through the work, will greatly benefit the field of usable security, as they represent major challenges. Will users compare the values accurately? Currently they dont in such situations. Will they bother to compare and not skip this step? No. Can they be duped into pairing devices whose displayed information does not match? Yes, easily. How large a value of b can they effectively deal with? Remember the magical number 7 to start with, also understanding of underlying cognitive issues should be well taken into account. Building the technical solution up with solid user research work is likely to lead to a more usable security management, which would be highly desirable from the research point of view
9. Contributes by including more information with regard to user behaviour
10. The framework is an important step in summarizing existing work on empirical authentication protocols, and offers researchers new to the field a concise introduction into the various relevant factors
11. The main strengths are that the framework involves a comprehensive set of factors involved in the success of a system that uses empirical channels for security in ad-hoc networks. In particular, I was pleased to see the human factors identified at some length, including even social and affective issues. To date, the human side of ad-hoc network security has been dealt with only lightly, and work is still being done that does not adequately address important factors that will limit any practical success. This framework may increase awareness of the issues, and help researchers and practitioners better determine the success of new proposals before actual trials. The framework might also be used to construct heuristic evaluation methods like those of Nielsen for methodical but quick and low-cost evaluation of proposals.

Table 7.4: Snapshot of expert comments

Criticism #1: lacks focus

> Lacks focus on motivation e.g. different goals might affect behaviour and outcome, consider how different environments which might be multiple and contexts might impact

- The central motivation for the HISPs is that context, technical security requirements, and human factors affect usability and security of HISPs. It is for researchers and designers to reason about how these factors may affect effective security within the context of an application. This is the central theme of the HISPs framework hence we disagree that there is no focus on the issue.

Criticism #2: human factors not mentioned prominently, *context* is loosely defined, and sections are unconnected

> - Although user intentions and human factors are mentioned, I think that this should be more central. That is, it might need to be mentioned more prominently. Without deep/good integration of security features in the tasks at hand, they will not be used properly.
> - Context is too loosely defined to be usable to the design of authentication system, as the cited works aim at other use cases than security. I think that other aspects of context might be more important here. More examples would help in understanding some of the categories.
> - The main thread of argument through the whole chapter is still unclear. It seems to consist of multiple unconnected sections, so a clear explanation in the introduction on how the "story is told" may help.

- The central theme of the HISPs framework, and this thesis in general, is that security tasks must be designed to fit users' goals within the context of their application. Empirical channels must be chosen or designed after taking human factors into account.
- We acknowledged that *context* was loosely defined to be relevant to HISPs. A precise definition of context that is relevant and useful in the study HISPs is provided in Section 2.1.
- The third comment is partially correct and has been handled within the introduction of this chapter and introduction of this thesis.

Criticism #3: HISPs may not lead to an optimal solution

> The framework might lead to local maxima, e.g., "manual comparison and entering" methods might always be considered superior, as they come with the least usage constraints and thus match the largest number of scenarios

- We disagree with this comment on the basis that, first, MCE does not always come with the least usage constraints. MCE requires devices with at least a numeric keypad and a display sufficient for a given format of digest. Second, human effort in MCE increases linearly with the size of the digest, hence, the method is not the best when digests longer than 32 bits are considered. Finally, associating a group of devices by a single user is impractical with MCE.

Criticism #4: framework is still theoretical, human dimension is too simplistic, and non-consistent terminology

> One weakness (or at least risk) is that the work remains a high-level framework without practical use; this can be avoided easily. My main concern is the model in 7.1. Intuitively it covers all relevant aspects, but especially the human dimension is too simplistic. Table 7.1 requires is jumping the gun. Consistent nomenclature might also help.

- Practical use of any tool is the next level after development and depends to a large extent on the target users and how it helps them achieve their goals. This forms part of future work to explore how the HISPs framework will be applied in practice.
- The human dimension is deliberately simple because it does not target any particular demography of users but tries to highlight generic human characteristics and it is up to a designer or researcher to explore specifics about a target population.
- The analysis in Section 7.4 provides the basis on which Table 7.1 is built.
- We believe that inconsistencies in nomenclature have been addressed.

Criticism #5: focus on specific rather than generic scenarios and lack of detailed analysis of motivations of earlier work

> - The author claims that a universal approach to fit all situations is not a likely one to succeed; instead, several scenarios are the likely outcome of the research. Though this may be true I would suggest to also aim for a generic scenario to start with and only drop the idea if needed as the research has progressed — as the outcome of the research only, not as an assumption to begin with. E.g. Kuo et al (2007) have suggested generic guidelines for building secure associations between devices in a usable way; the work should address this work and this goal, too. The specific scenarios can be based on a universal scenario, at the very least.
> - Though the level of understanding of existing related work seems quite good, there is some oversimplification in the analysis of how and why this research has progressed and been targeted the way it has. In order to avoid easy criticism, the author should take good care in analyzing the motivations and pitfalls of the earlier work properly.

- A system designer uses the framework to address a need in a specific application whereas a researcher may aim to develop an empirical channel that may apply to as many scenarios as possible. In this thesis, we argue that developing an effectively secure empirical channel, whether for a specific scenario or a universal solution, requires taking into account the various factors that may affect its security and usability. How specific and stringent these factors are depends on the the goal of the researcher or system designer. The work the expert refers to (Kuo *et al.* [64]) proposed standardising hardware and a common user interface across devices. While this proposal can undoubtedly address some of the security and usability issues of device association, it is impractical given the cost of standardising hardware and the disparate device affordances.
- The second criticism has been addressed both in the introduction to this thesis and in the literature review.

7.6 Summary and conclusion

In this chapter we presented the HISPs framework, designed for both researchers and system designers, for reasoning about empirical channels. The framework helps researchers to develop empirical channels that are effectively secure and work across contexts. System designers, on the other hand, usually have specific requirements and want to choose an empirical channel that satisfies both technical security and contextual needs. However, designers are faced with empirical channels that only work in specific contexts with a likelihood of not suiting the application at hand. Both designers and researchers, therefore, need to identify and be aware of factors that may affect usability and security of empirical channels. The HISPs framework identifies these elements as human factors, technical security requirements, and context. We discussed how each of these elements is crucial to reasoning about the security and usability of empirical channels.

The HISPs framework can be applied during the analysis stage of a UCD process through which user, task, and context are analysed. We demonstrated the application of the HISPs framework, from a designer's perspective, by considering a specific application scenario and analysing it to arrive at a recommendation for a suitable empirical channel.

The framework was validated using experts from industry and academia. Experts provided feedback on the strengths, contribution, practicality, weaknesses, improvements, and future direction of the HISPs framework. This feedback was used to improve on the initial proposal of the HISPs framework. We have also summarised expert feedback and responded to the criticisms from the experts.

Chapter 8

HISPs: Analysis and Evaluation of Security and Usability

8.1 Introduction

In this chapter, we present a security-usability threat model, detailing factors that are pertinent to the security and usability of empirical channels, together with a process for evaluating the security and usability of a system. We believe the security-usability threat model and evaluation process is not only restricted to HISPs but can be applied to any secure system. We demonstrate the application of the model and process for evaluating security and usability by applying them in conducting user studies of empirical channels.

The chapter is organised as follows: a general background to secure system evaluation is given in Section 8.2, followed by a presentation of the security-usability threat model in Section 8.3. The process for evaluating security and usability of secure systems is presented in Section 8.4. A case study validating the threat model together with a process for analysing security and usability is presented in Section 8.5. Section 8.6 summarises and concludes the chapter.

8.2 Current practice in security and usability evaluation

Laboratory studies of secure systems are popular both in research communities and industry. In a laboratory experiment, a researcher has a large degree of experimental control over variable manipulation [43]. In addition, laboratory studies may be cost effective compared to field (real world) studies. Despite conducting studies in a laboratory setting, researchers' aim is to draw conclusions that extend beyond this environment. Extending results beyond the environment in which a study is conducted is known as external validity [115].

A careful design of laboratory experiments is crucial to achieving external validity. A researcher must carefully consider participants, scenarios, tasks, instructions, and procedures. Participants should be representative of target users, scenarios and tasks must also be closely matched to real world experiences. Previous studies, e.g. [104], have shown how instructions can change participants' behaviour. In addition, procedures of how instructions are given and how data is collected and analysed have an effect on the results and conclusions drawn from them. These tight controls on participants and their recruitment, scenarios, tasks, instructions, and procedures are aimed at ensuring that an experiment closely represents a real world scenario (and its context) and that the results obtained are valid in the real world as much as they are in the laboratory.

HCI analyses user tasks and scenarios while HCISec analyses security tasks that are in many cases secondary goals to the user. Users perform security tasks as a consequence of trying to achieve their primary goals. Being non primary tasks, security suffers from the *unmotivated user property* [129] — users are unmotivated to do security. It is, therefore, unrealistic to ask participants to perform security tasks alone in an experimental setting since this makes security a primary task — a scenario uncommon in the real world.

Security and usability have a common goal of fulfilling user expectations [134]. It is, therefore, reasonable to develop a model that encompasses factors of security and usability — to aid the analysis of a secure system — and a process to aid security and usability evaluation of systems. To develop a security-usability threat model, it is crucial to understand factors that are central to security and usability of a system. Many definitions of usability capture these factors whereas there are no security definitions that capture factors related to security. To discover factors for evaluating security, we reviewed HCISec literature on user studies. We noted that the studies fell into one of six categories:

- *Authentication* — authentication mechanisms have been, and continue to be, extensively studied. Studied mechanisms range from traditional methods (such as text passwords) to novel ones such as grid entry passwords or image based passwords. Studies on authentication mainly focus on measuring factors such as memorability or cognition, and efficiency (the speed with which one can successfully authenticate).

- *Encryption* — these studies focus on secure email. Unlike studies on authentication mechanisms that focus on memorability/cognition, studies on secure email focus on users' understanding of the mechanisms to send email securely. Knowledge of the mechanisms is a crucial factor for correct execution of email encryption as was found in [129].

- *PKI (Public Key Infrastructure)* — the main problem studied here is that of identity; whether a user can correctly identify a website to be secure or otherwise. Studies have been conducted on browser indicators that are supposed to provide users with information about the security and identity of a particular website. Indicators include the traditional padlock symbol at the bottom right corner of a browser, colouring the address bar, as well as symbols and logos on a web page. Like encryption mechanisms, the major focus in these studies is on users' knowledge of particular indicators as well as factors such as vigilance (always looking out for indicators) and attention.

- *Device association* — these studies focus on efficiency, effectiveness (failure/success of association), and security failures. Security failures cannot be classified as part of effectiveness as they do not result in failure of association, but result in users associating a personal device with an unintended one. In this case, the user has accomplished the association only except with the wrong device.

- *Security tools* — these are systems that help users manage their security. They include firewalls, password managers, and privacy managing tools. Studies of these systems focus on users' knowledge in using the tool and whether the tool is used correctly or otherwise. Usability is mainly measured by users' ability to accomplish their goals and whether what users think the system has achieved is what the system has actually accomplished.

- *Secure systems* — systems that do not fall into the above categories fall into secure systems. These systems are aimed at achieving user goals (that are not related to security) but have an element of security. For example, studies of peer-to-peer software have found that while users are able to achieve the goal of sharing files, many users unknowingly end up sharing files

they would not want to sharebecause the system was designed to share certain directories by default [38].

Reviewing studies in each of the above categories highlighted not only the main usability factors but also the security factors that were evaluated. There is, however, a blurred line dividing usability and security factors — some of the usability factors cause users to behave insecurely, and some of the security factors obviously impair performance.

While these studies provided factors that are crucial to evaluating security, they focus entirely on factors of usability. All studies that were reviewed had presentation on *usability analysis* that discussed various factors of usability. A complete evaluation must consider factors that may affect security as well. Evaluating one and not the other introduces similar problems as those discussed by Flechais [31] where a user perceives an action as harmless when it is not; a user has greater incentive to engage in a dangerous interaction; or a user is incapable of desirable interaction. It is, therefore, crucial that a security-usability threat model that helps analyse both security and usability of a system is developed.

8.3 Security-usability threat model

HCISec is centred around the user. The user needs a system that is both secure and usable. Legitimate non-malicious users should not compromise or be duped into compromising a system's security. HCISec, therefore, requires methods and procedures that lessen users' burden and protect the system from the very user. It requires a security threat model that encompasses factors of usability as a difficult-to-use system may force users to resort to insecure behaviour such as circumventing security processes — making protected assets insecure.

The HCISec security threat model should be different from standard security threat models. Standard security threat models focus primarily on malicious attackers who may or may not be legitimate users. HCISec's primary focus is on legitimate users' mistakes that may compromise the system. The security-usability threat model presented in Figure 8.1, therefore, centres around a legitimate user who has no intention of breaking the system.

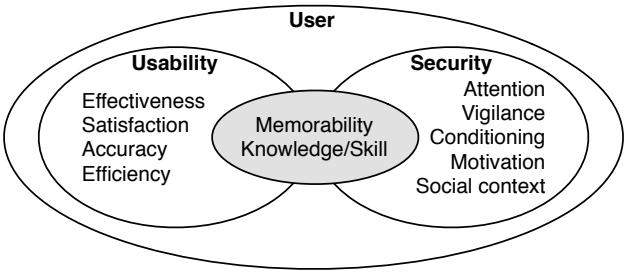

Figure 8.1: Security-usability threat model

The security-usability threat model depicts the critical factors that need investigation during the evaluation of usability and security. It identifies factors that affect usability or security and also factors that are related to both. These factors relate to legitimate users who have no intention of harming the system. We discuss each of these factors below.

8.3.1 Usability

- *Effectiveness* — a system is only useful if its users are able to achieve intended goals. An ineffective system is likely to be abandoned. Effectiveness is usually measured by whether users are able to complete a particular task or not. This approach of measuring effectiveness is appropriate for most studies where a task consists of a single step that can be achieved through a single path. However, complex and multi-step tasks may require a more fine grained definition of success or failure which may include levels such as partial failure/success.

- *Satisfaction* — while objective usability analysis of systems is common, users' subjective assessment is crucial to systems' success. For example, a system may be usable (by usability standards) but users may label it unhygienic [97]. In other words, a system is bound to fail even when it is usable if it is not acceptable to users. User satisfaction can be assessed through interviews and rating scales.

- *Accuracy* — accuracy was identified in authentication and device association studies. In many cases, authentication systems demand that users to enter passwords with 100% accuracy while certain mechanisms in device pairing require 100% accuracy when entering or comparing short strings [52]. Accuracy demands on users are impacted by other factors such as recall of required information, environmental, or personal factors (As earlier discussed in Chapter 7).

- *Efficiency* — using a system and being able to achieve a specific goal is insufficient in itself. The goal must be achieved within an acceptable amount of time and effort. What is the acceptable amount of time or minimum effort in one system or context may not be in another. To this regard, a system is rated as efficient in relation to other similar systems or established benchmarks. Efficiency is captured by measuring the time to complete a task or the number of clicks/buttons pressed to achieve required goals.

- *Memorability* — many authentication systems require users to memorise secrets that they should recall whenever they want to be authenticated by a system. The number of secrets one is required to keep increases with the number of different authentication systems that an individual interacts with. This results in memorability problems where users experience difficulties authenticating themselves to various systems and often ends in requests to reset those secrets [14].

- *Knowledge/skill* — usability definitions often use learnability to refer to how easy it is to learn using a system. This is based on the assumption that users will learn or actually attempt to learn and understand a system. This assumption is flawed particularly in personal secure systems. What we found in our own studies and those that we reviewed is that despite using a system, users only care about those parts that they think are important to specific operations they need — while in many cases security tasks are not seen to be important. The knowledge that users have about a system plays a role in how usable that system is to the users. In our study in Chapter 4, participants who were musical found comparing melodies easier than those who were not musical. This was also the case for the barcode method between participants who frequently used their mobile phones to take pictures and those who did not.

The above factors have a direct effect on the usability of a system. A usability evaluation must determine which factors apply to a specific application and context.

8.3.2 Security

- *Attention* — users can easily be distracted, causing them to shift their attention from a task at hand. Security tasks must not demand undivided attention from users as this is likely to cause frustrations, and possibly security failures. For example, empirical channels such as *manual comparison* of sound require users to be attentive while the sound is played; any distraction

entails restarting from the beginning but it may also result in the user proceeding without cautious comparison. Moreover, users have a view that secure systems are disruptive — they often disrupt one's attention in order to attend to security prompts, for example. It is now, however, common knowledge that both disruptive (such as certificate prompts) and passive (e.g. browser padlock) approaches are usually ignored by users.

- *Vigilance* — secure systems tend to expect users to be alert and proactive in assessing the security state of a system. This has been problematic as studies have shown that even experts (users who understand the working of a secure system) are not always alert. For example, Dhamija *et al.* [22] found that experts on web site security indicators did not even look in places where those indicators were, hence falling for simulated phishing attacks which they would have avoided had they looked and noticed the absence or presence of indicators. Tasks that pose this security risk tend to be those that require users to divert attention from a primary task in order to attend to a security task. Such tasks should be analysed and integrated into users' work flow or eliminated if possible.

- *Motivation* — users have different levels of motivation to perform security tasks in different circumstances. For example, participants in the study in Chapter 4 (where making a payment was used as a primary task) indicated that they would prefer typing passkeys longer than 6 digits for financial transactions exceeding a certain monetary value. In this case, participants saw the risk to be more direct to them (losing money) than in a case where risk is perceived to be low or directed at someone else.

- *Memorability* — authentication systems often require users to memorise secrets that are difficult for someone else to guess or even attack by brute force. As the number of secrets one has to memorise increase, it can become more difficult to recall a particular secret when confronted with a system asking for one — particularly if the system is not used frequently. As a precaution to avoid forgetting and resetting, users write down these secrets. This in itself impacts the security of the system because written down secrets can be found by others who may then use them for malicious purposes.

- *Knowledge/skill* — users' knowledge or skill level plays a major role in the security of a system. For example, users of banking websites cannot distinguish between a padlock at the bottom of a browser window and one displayed as an image on a web page [104]. Previous studies have also found that training users in using secure systems is ineffective [96]. These problems exist

because the tasks — checking presence of padlock and learning about good security practices — are not users' goals in many cases. Consequently, users' knowledge/skill about a system is an essential factor in analysing its security and usability.

- *Social context* — humans are social beings. They help each other and share various artefacts. While sharing is generally good, it is undesirable for security if users share their security secrets. For example, Beckles *et al.* [8] found that users working on a particular project shared one digital certificate rather than each having their own as intended by system designers. Another common example is found in [98] where users shared passwords for various social reasons. Users also disclose secrets because someone is offering to help them if they do so. This has been exploited in many situations and that is why it has been named *social engineering* [75]. Understanding how social context affects security and usability of a system is crucial to designing and system that are usably secure.

- *Conditioning* — repetitive security tasks for which users can predict an outcome can become a threat to the security of a system. A common example are pop-up boxes that ask users whether a particular certificate should be trusted or not. A few encounters with such pop-up boxes make one realise that clicking a particular button will make the pop-up disappear and will allow for continuation of intended task. A security-usability analysis of a system should assess whether security tasks have the potential for condition users and if so, develop ways of mitigating them.

8.3.3 Measurable metrics

For a successful evaluation, both security and usability factors must be measurable. Measurements are crucial for comparative analysis and basic quantification of specific usability or security criteria. Table 8.1 summarises the measurable metrics for each of the factors in the threat model.

8.3.3.1 Usability metrics

Each usability factor can be measured and quantified using one or more metrics that comprise that factor. Effectiveness can be measured by task success rates. Satisfaction is subjective and can be captured using rating scale questionnaires such as ASQ [66] or interviews. Accuracy can

Usability		Security	
Factor	**Measurable metrics**	**Factor**	**Measurable metrics**
Effectiveness	task success	Attention	Attention - failures
Satisfaction	Satisfaction	Vigilance	Vigilance - failures
Accuracy	Success rates	Conditioning	Conditioning - failures
Efficiency	Completion times, number of clicks/ buttons pressed	Motivation	Perceived benefits, susceptibility, barriers, severity
Memorability	Recall	Memorability	Successful recall
Knowledge/ skill	Task success, failures, mental models	Knowledge /skill	Task success, failures, mental models
		Social context	Social behaviour

Table 8.1: Measurable metrics

be quantified as success rate on tasks that require a certain degree of accuracy. For example, the number of users who successfully log on to a system using current text password methods provides a measure of accuracy. A system's efficiency is mainly expressed as the amount of effort users expend to accomplish a task and can be captured as the amount of time it takes to complete a task and number of clicks or buttons pressed. While memorability can be measured as the number of users who successfully recall a previously memorised secret, many usability studies are conducted in laboratory environments where the length of time or usage pattern may be unrealistic. If a system studied is already deployed but a longitudinal study is not possible, users can be asked to report on their experiences in using the system. Users' understanding of a system can be shown through task completion rates, users' mental models, as well as errors committed. A mental model can be measured by comparing a user's perceived security state of a system with actual state. Tasks may be completed successfully but with errors. Thus, it is essential that errors that do not lead to task failures are measured too.

8.3.3.2 Security metrics

Security factors must also be measurable. Attention can be quantified by monitoring and determining whether or not a security failure is due to lack of attention to specific piece of information. For example, eye tracking has been employed in studies of website security indicators to capture whether

users take time to look at indicators or otherwise. We can also capture vigilance by monitoring whether users are consistent in paying attention to security tasks. This information can also be captured through self report questionnaires. In addition, one can capture conditioning by analysing errors that users' committee and determining whether previous events had an effect on the occurrence of those errors. Motivation cannot be directly measured but research shows that motivation to engage in a security task is driven by perceived susceptibility to attacks, benefits of and barrier to engaging in a security behaviour, and perceived severity of a security failure [78]. Measuring these factors gives an indication of users' motivation to use and execute security tasks effectively. Memorability can be captured by counting successful recalls and asking users whether they have memorability problems or not while knowledge/skill can be captured in form of task success, mental models, and errors. To capture effects of social context on the security of a system requires a qualitative approach; studying users behaviours in relation to those around them and elicitation of information on how they interact in relation to a studied system.

8.4 Security-usability evaluation of secure systems

To analyse the security and usability of a system based on the threat model, the concepts of usage scenarios (or simply scenarios) and threat (negative) scenarios [2] are used. In this particular context, we define usage scenarios as actions that are desirable to a system's stakeholders whilst threat scenarios are defined as actions that are undesirable. HCISec, on one hand, is concerned with making usage scenarios accessible to the user with low mental and physical workload. For example, in an email application, usage scenarios could be composing an email, locating a contact, sending email, or creating a new contact.

On the other hand, HCISec is concerned with threat scenarios (undesired actions) that may cause non-malicious users to break security of a system. The focus is on non-malicious users who may break a system's security due to factors discussed in the security-usability threat model. While threat scenarios are usually associated with malicious attackers, in this particular context they are associated with legitimate users whose goal is non-malicious. For example, in a secure email application, as much as one may be concerned about whether users can encrypt emails, it is also crucial that one be concerned with whether they may accidentally encrypt a particular email with a key belonging to an unintended recipient.

Whilst usage and threat scenarios are traditionally used during requirements gathering and design [57], in this context they are applied during system development life cycle as well as after product release. This is particularly so because the focus is on analysing an existing systems, or a design, to identify factors that may affect effective security. Figure 8.2 summarises the steps in the security-usability analysis process.

8.4.1 Identify usage scenarios

HCI researchers identify usage scenarios before conducing a usability evaluation. The scenarios are specific tasks that a typical user of a system would endeavour to accomplish to achieve a specific goal. Usage scenarios are presented to participants (e.g. in usability testing) or evaluated by experts and performance measures are recorded. An evaluation of scenarios provides performance data on factors that may affect a system's usability. A system is usable if performance data proves that it meets pre-agreed criteria and satisfies users' and contextual needs.

8.4.2 Identify threat scenarios

As earlier pointed out in this chapter, HCISec is also concerned with legitimate users making errors that breach a system's security. Events that may result in such behaviour (threat scenarios) must be modelled and evaluated. The goal is to measure how easy legitimate users may unknowingly break security. In HISPs, for example, a threat scenario may be that users may not pay attention to comparing digests — which may result in indicating that compared digests match when the contrary is true. Since security of HISPs relies on users ensuring that the digests displayed match before they accept an association, lack of attentiveness may result in one or more devices associating with an unintended device. To model this threat scenario and determine how likely users are to compromise the system, an attack scenario must be presented to participants (for example by presenting presenting non-matching digests to participants).

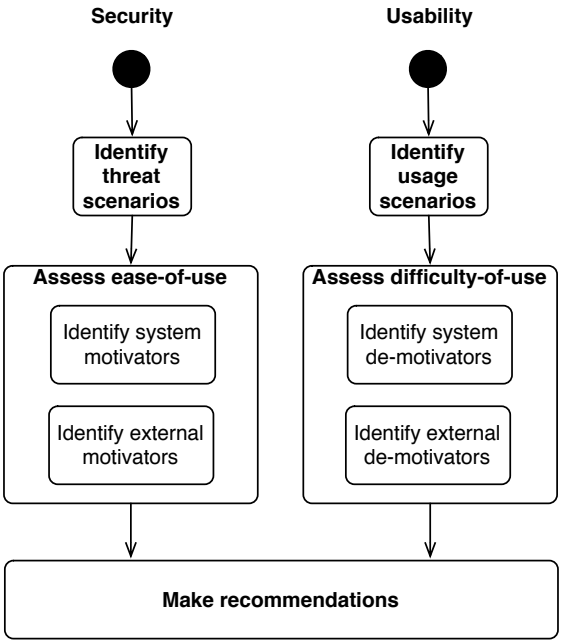

Figure 8.2: Process for security-usability analyses

8.4.3 Assess difficulty-of-use of usage scenarios

The objective of conducting a usability evaluation is to identify factors that may introduce difficulties into a system and to mitigate their impact. An assessment of difficulties of usage scenarios can be in the form of user experiments, cognitive walk throughs, or interviews. Each usage scenario should be evaluated against usability factors in the threat model: effectiveness, satisfaction, accuracy, efficiency, memorability, and knowledge/skill. It is important to note, however, that these factors are system specific. For example, while memorability is crucial in many authentication systems, it is not in most PKI or security tools. A security-usability evaluator must identify factors that affect the usability of a particular system. Once usage scenarios are evaluated for difficult-of-use, the resulting data affords the following:

- *Assess system de-motivators* —the aim is to identify properties that may de-motivate from using a system in a desired and prescribed manner. The performance data collected while assessing difficulties of usage scenarios provides information crucial to identifying system de-motivators. For example, the amount of time it takes to accomplish a particular task may deter users from following prescribed procedure when using a system. Identifying system de-

motivators focusses on factors of a system that deter or make it difficult for users to use it effectively.

- *Identify external de-motivators* — users may also be de-motivated from performing usage scenarios by factors that are external to the system. This is because systems, together with their users, operate in concert with other systems. For example, external de-motivators may include physical environmental variables such as light intensity and background noise, social variables such as presence of people not participating in an association, and personal variables such as age, gender, culture, and education. Users can also be de-motivated if they have access to a competing system that they perceive to be more usable. The competing system may be insecure — sending unencrypted email, for example — but may be seen as effective and efficient by users.

8.4.4 Assess ease-of-use of threat scenarios

Users follow the path of least resistance [133]. The goal for assessing ease-of-use of threat scenarios is to identify factors that may force users to engage in undesirable actions and evaluating how easily they can occur. If threat scenarios are more difficult to accomplish compared to usage scenarios, legitimate users are unlikely to perform the former. Despite having good intentions, users may perform threat scenarios if usage scenarios are harder to carry out. For example, an evaluation of an authentication system may consider assessing how difficult it is for users to memorise secrets (usage scenario) — which may force users to write them down (threat scenario) — while an evaluation of an empirical channel may consider assessing how easy it is for users to bypass tasks such as comparison of digests.

- *Identify system motivators* — to understand why users may perform threat scenarios, factors of a system that may nudge users in doing so must be identified. If a threat scenario is more usable than a usage scenario, this is a system motivator that may force users to perform the former.

- *Identify external motivators* — factors external to a system may motivate users to perform threat scenarios. For example, imperfect lighting conditions may make it harder for users to compare digests in device association and may be a motivator for users to skip digest comparison. Similarly to usage scenario's de-motivators, the aim is to minimise external motivators for

threat scenarios. We have pointed out earlier in this chapter how social context is an external motivator: users may share passwords or security certificates among themselves, or may share passwords with outsiders whom they perceive as trying to help.

8.4.5 Make recommendations

The final stage in the security-usability analysis process is to make recommendations based on the preceding steps. Recommendations focus on areas that need improving to make usage scenarios easily accessible to legitimate users and also areas that need to be *hardened* for threat scenarios.

In addition to users and the system, the analysis process focusses on external factors. A system may be usable or secure in itself but may not be when in actual use because external factors outweigh internal ones. For example, an employee who is aware of password security policies and of the need to avoid sharing passwords may be forced to share them with colleagues in stressful situations such as being late for work and needing to send an urgent report.

It is unrealistic to expect to achieve maximum usability and security in all secure systems. The goal is to minimise as much as possible the possibility of threat scenarios and maximise the accessibility of usage scenarios. For example, allowing users to write passwords down may be acceptable if the threat from attackers using dictionary-based password cracking tools is particularly severe.

It is also unrealistic to expect that all internal and external motivators or de-motivators can be eliminated. In either case, the goal is to minimise these factors to an acceptable level. An acceptable level varies from case to case and needs to be assessed based on the system and its context.

8.5 Application and Validation

To validate the threat model and process for evaluating security and usability, we conducted two case studies evaluating usability and security of HISPs. In this chapter, we only discuss details of one study while the other study is presented in Section 6.4.

8.5.1 Case Study: Security and usability study of HISPs in group scenarios

In order to evaluate security and usability of empirical channels in group scenarios, the threat model was used to identify possible sources of both usability and security problems. For usability, the following were identified: effectiveness (whether users can accomplish their intended goals using a particular empirical channel), efficiency (the speed with which security tasks where completed), satisfaction (users' perception of usability), and accuracy (how accurately users can transfer/communicate pieces of data from one to another). For security the following were identified: attention to the association process, conditioning, social context (group), vigilance (can participants be actively attentive to the association process throughout), and motivation. In the design and conducting of the study, particular attention was paid to these factors.

Upon identification of factors that may pose challenges to security or usability or both, the process for evaluating usability and security was employed. Rather than just focussing on usability in the design and conducting of the study, the process for evaluating usability and security enabled paying equal attention to security issues as well. Using this process, usage scenarios (scenarios that users can use in the real world rather than just security tasks) were identified. They included secure exchange of contact details, digital cash transfer, quiz contest using mobile devices, cinema ticket transfer, and secure group meeting. Identification of usage scenarios was followed by identification of threat scenarios. These are scenarios that a secure system must not allow to happen as they may break security. In device association of groups the following were identified; accepting a non-matching digest, initiator mistaking failure from one or more devices for success, intruder joining network without knowledge of initiator or other group members. Usage scenarios were used to determine difficult-of-use of security critical tasks in empirical channels and how they can be improved while threat scenarios were used to determine ease-of-use of actions that should be avoided by users.

Using prototype applications for usage scenarios as primary tasks, participants interacted with empirical channels only as a means to achieve their goals. Threat scenarios were also included in the prototype. User's attentiveness was tested by creating discrepancies in the digests displayed on devices while vigilance was tested by repeating such events. In addition, the study was designed such that in certain cases one device showed the number of devices connected to it that was greater than expected.

8.5.1.1 Results and conclusion from the study

The results from the study show that the model is robust in analysing a secure system to identify factors that may pause challenges to security and usability. As each factor was analysed against association methods for groups, the model enabled identification of all critical factors related to the system studied. The output of the analysis stage was a critical input to the design of the studies. Before a study is designed, a crucial question that needs to be asked is 'what has to be evaluated in the system?' The output of the analysis process provides an answer to this question.

The factors identified during the analysis stage are essential for designing and conducting a study. Moreover, they are crucial to analysing and understanding data collected during the study. A study that is designed to capture specific factors of usability and security will pay more attention to factors that are the focus of the study than those that are not. We found through the two case studies (and other studies) analysing data from the study is easier when evaluated factors are identified before the design of the study than otherwise.

While it is still possible for one to design the above study in a similar manner without using the model and process of evaluating usability and security, the approach in this chapter ensures that different researchers (regardless of experience) can analyse and evaluate secure systems in a similar manner. In addition, because the model forces one to identify factors that have to be evaluated, it means that comparison of results may become easier than is currently the case. This may also ensure conformity on how a study is reported as one is required to state what factors were evaluated and how they were evaluated.

8.6 Summary and conclusion

In this chapter, we presented a threat model for analysing usability and security of HISPs. The model is crucial for systematically analysing HISPs and identifying factors that may affect security or reduce usability. In addition, the model provides input to the evaluation process, also presented in this chapter. The factors identified in the model are employed in formulating usage and threat scenarios. During evaluation, these scenarios are used to identify and understand both system and external factors that are threats to a system's usability, security, or both. Usage scenarios are used to

identify areas that may hinder the usability of a system, whereas threat scenarios are used to identify areas that may help non-malicious users to break the security of a system. When a system's threat scenarios are more usable compared to usage scenarios, users are more likely to perform the former. External factors, too, may cause users to perform actions that they may not normally perform.

Both the model and evaluation process were validated using two case studies. The results of the case studies show that the model and evaluation process are robust for HISPs analysis and evaluation. We, however, believe that the model encompasses sufficient factors to be robust for analysing any secure system. In addition, the evaluation process relies on the output from the analysis and is not dependent on properties of a particular system.

Chapter 9

Conclusion and future work

9.1 The research problem restated

Security and usability seem to be at odds because designers consider one and not the other during system design. Bolting on usability to a finished product may compromise the security of that system. Conversely, bolting security to a finished product may make a system unusable. Design of an effectively secure system requires integrating security and usability requirements from the onset.

Usability and security of HISPs is compounded by a number factors. First, HISPs require users to perform security critical tasks correctly — creating extra work for them. In practice, this is difficult to implement given that HISPs run on devices that may have limited input/output interfaces and also the problems that users generally experience when using security mechanisms. Another issue is users' understanding of the need to have a more secure protocol. The current Bluetooth pairing is relatively straightforward, hence, users need to understand the importance of the information they want to protect in order to buy into the need for additional security. Lacking this incentive to adopt stronger security, should a new security mechanism require more effort than currently is being expended on the PIN mechanism, serious adoption issues may arise.

Second, different association scenarios and modes of authentication affect the security and usability of HISPs. In one-way authentication, a user is only required to confirm authenticity of information on a subset of devices. On the other hand, mutual authentication requires users to confirm authenticity of

information on all devices involved in an association. Either mode of authentication places different demands on users and understanding what these differences are and what is their effect on the usability and security of HISPs is critical to designing empirical channels that are effectively secure. In addition, the usability and security of an empirical channel is affected by whether the association scenario is pairwise or group, close or distant devices, single or multi-user.

Finally, usability and security of HISPs is contextual. A protocol may be used in stressful situations such as in military operations, peer-to-peer electronic payment systems, online payment systems, or in group settings such as meetings. Context include the social or physical environment. Users change behaviour depending on whether they are in a crowded or private environment. Device association also occurs in physical environments that have different variables such as lighting and noise levels. These contextual factors make different demands on users and certainly affect the usability and security of HISPs.

Despite the above challenges, HISPs present an opportunity. They enable users to effectively manage their own security by permitting users to actively participate in device associations and ensuring that only devices (or whose owners) they trust take part. This is crucial in scenarios where there is no infrastructure or third party to support device association, users do not trust a third party, or users want to retain control of the security of the associations in which their devices participate. This applies to applications involving exchange of sensitive information such as payment and medical data.

Given the above background, this thesis explored the research question,

> *"Given the different association scenarios and modes of authentication in Human-Interactive Security Protocols, how can we improve on existing, or design new, empirical channels that suit human and contextual needs to achieve acceptable effective security?"*

To answer the above research question, we first decomposed it into four specific research questions:

1. *Which methods for transferring or comparing digests are effectively secure?*
   This question targeted assessing the effectiveness of currently proposed empirical channels through identification of their weaknesses and strengths for both security and usability. To assess the effectiveness of empirical channels, we conducted user studies to gather empirical

data in different scenarios. User studies were chosen as a research methodology because it was necessary to gather empirical data on the different methods.

2. *In what association scenarios and modes of authentication can these methods work while achieving acceptable effective security?*

   A method that is acceptable and effectively secure in one mode of authentication and association scenario may not be in another. This research question aimed at identifying human, technical, and contextual factors that may impact the security and usability of empirical channels.

3. *How can we design or improve existing empirical channels to achieve acceptable effective security?*

   Currently proposed empirical channels force one to choose between security and usability. Studies have found that empirical channels are either usable and preferred but susceptible to security failures or vice versa. In addition, current proposals focus on specific contexts. For example, timing methods and short range directed channels may work well in pairwise associations but difficult to in group scenarios or where devices are a certain distance apart. While empirical channels rely on users to accurately perform security critical tasks, many do not compel users to perform required tasks. An example of such methods is *manual comparison*. A solution to this research focussed on developing design principles and methodologies for developing new, and improving existing, effectively secure empirical channels.

4. *How can we improve methodologies and procedures for usability and security evaluation of HISPs?*

   Usability evaluations of secure systems need methodologies that deviate from standard HCI techniques. A usability evaluation of secure software should not focus on usability to the exclusion of security. Moreover, because usability and security have a closely tied relationship, it is imperative to consider both factors when evaluating a system. Failures of HCI methodologies in HCISec user studies have manifested through studies that focus on usability only and make recommendations that are detrimental to the security of a system. Answering this research problem required developing new methodologies for user studies to fit HCISec requirements.

9.2 Research contribution

9.2.1 Analysis of the effectiveness of empirical channels for mobile device associations

Security and usability of single-user pairwise scenarios are affected by a number of factors. First, users get conditioned by repetitive tasks to which they can predict an outcome. Second, users must be motivated to effectively engage in a security task. Lack of motivation by users increases the chance of security or usability failures or both. Third, HISPs are secondary tasks but require users' attention. The more attention an empirical channel demands on users, the less usable it is and, possibly, insecure too. In addition, different environments put different demands on users' attention. A user carrying out device association in a quiet or well lit place may experience different attentional demands compared to one carrying out the same task in a noisy or dark environment. Fourth, affordances that a particular device offers determine possible user actions. In addition, differences in affordances among devices have an effect on the usability and security of HISPs. Finally, personal variables such as age, experience with mobile devices, knowledge and skills, and health conditions affect usability and security of empirical channels.

To design empirical channels that are both usable and secure for different contexts of use requires taking the above factors into account. The above factors are broad, within the context of HISPs, and it is unreasonable to assume that an empirical channel that suits all contexts of use and all users can be developed. The factors, however, must be reasoned about within the context of a specific application scenario and target users.

The challenges of security and usability of HISPs for group scenarios differ in many respects from those of single-user scenarios. In group scenarios, coordination among participating members is crucial to achieving required security. In addition, the number of participants in a group affects usability and security depending on the empirical channel used. While it has been believed that group settings may be more subject to failures during an association process compared to single-user associations, the findings in this thesis show the converse to be true. Group members feel the need to help each other and cover up the weaknesses of struggling members. Comparatively, there is no statistical significant difference between methods evaluated in terms of group completions times, ASQ scores (initiators), and overall rating scores (group members).

Security and usability of empirical channels in group association scenarios are affected by a number of factors. In these scenarios, security is a function of a sum of efforts rather than weakest link. Users tend to work as a team and weaknesses of a subset of members are covered by more active members. In addition, users designated as initiators take the responsibility of asking other member for the result the association. In addition to sum of efforts, security and usability of empirical channels are affected by context and users' perception of what security is. The context of operation dictates possible user actions while perception affects users' subjective usability and acceptance of empirical channels. Furthermore, users learn by trial and error only the parts of a system they consider important to their primary tasks. This has negative implications on security is which not a primary task in device associations.

The design of empirical channels for group scenarios should utilise the coordination effect of participants to effectively defend against attacks that may target potentially weak participants. Moreover, attention must be paid to context of operation, users' interpretation of security, conformity, perceptions, and learning behaviours. These factors have an impact on usability or security or both.

9.2.2 Design principles

Human-Interactive Security Protocols require users to carry out security critical tasks in order to establish secure device association. While device association is not usually a primary task for users, security critical tasks must be carried out correctly. Given users' lack of motivation and the demands of competing tasks, empirical channels must be designed such that users do not compromise the desired security.

We proposed principles for designing secure and usable empirical channels. The *principle of commitment* ensures that users do not compromise security by disclosing only partial information that a user employs to commit to a final outcome while the *principle of unpredictability* achieves the same goal by preventing users from predicting required action. To improve usability, thereby security, the *principle of single interaction path* requires implementations that provide a single ordered sequence of actions that users follow every time to achieve an intended goal. In addition, tasks must be concentrated on one device to minimise the number of inter-device interactions. In contrast to other principles that focus on empirical channels, *design by context* focusses on the context and application in which a method is applied.

We demonstrated the effectiveness of the principles by developing two empirical channels. We designed the empirical channels to satisfy all but the *deign by context* as this requires contextual information that is dependent on a specific application. The proposed methods were evaluated for usability and security using a laboratory experiment and results compared to those of previously proposed methods. The results of the comparison show that the proposed principles are robust and, if followed, can lead to developing secure and usable empirical channels.

We also discussed the scenarios to which proposed methods can be applied. The discussion highlighted the fact that it is possible to design methods that are secure, usable, and applicable to a wide range of scenarios. Unlike previously proposed methods that are either insecure, difficult to use, or limited to a single scenario, proposed methods ensure that users carry out security critical tasks accurately without demanding undue effort. Considering the range of scenarios that the proposed methods can apply, the principles are also useful in developing methods that work across context.

9.2.3 Framework for reasoning about empirical channels

Researchers are faced with the challenge of designing empirical channels that are effectively secure across context, hence, they need to understand and reason about the factors that may affect the security and usability in different contexts. Moreover, proposals of empirical channels take a single faceted approach that ignores crucial factors rendering them insecure or unusable in common use scenarios. System designers have needs to satisfy a specific application that operates in specific contexts. The challenge designers face is choosing from existing methods an empirical channel that is effectively secure within the context of the application.

The HISPs framework is designed to help both researchers and system designers in reasoning about empirical channels and developing/choosing methods that are effectively secure. The framework helps researchers to develop empirical channels that are effectively secure and work across contexts and different application scenarios. System designers, on the other hand, usually have specific requirements and want to choose an empirical channel that satisfies both technical security and contextual needs. However, designers are faced with empirical channels that only work in specific contexts with a likelihood of not suiting the application at hand. Both designers and researchers need to identify and be aware of factors that may affect usability and security of empirical channels.

The HISPs framework identifies these elements as human factors, technical security requirements, and context.

The HISPs framework can be applied during the analysis stage of a UCD process through which user, task, and context are analysed. We demonstrated the application of the HISPs framework, from a designer's perspective, by considering a specific application scenario and analysing it to arrive at a recommendation for a suitable empirical channel. It also discusses how the framework builds on previous work. The HISPs framework was validated using experts from industry and academia. Experts provided feedback on the strengths, contribution, practicality, weaknesses, improvements, and future direction of the HISPs framework. This feedback was used to improve on the initial proposal of the HISPs framework. We have also summarised expert feedback and provided our feedback to the criticisms by the experts.

9.2.4 Model and process for evaluating usability and security of empirical channels

Following HCI standard procedures and methodologies in conducting security usability studies has resulted in focusing entirely on usability issues (ignoring security) or lack of external validity because participants carried out security tasks as primary tasks or both. The security-usability threat model was designed to solve this problem specifically for HISPs.

To identify elements that are crucial to evaluating security and usability of secure system, a literature review of user studies was conducted. It was apparent from the literature review that different systems are affected by different elements depending on whether a system is an authentication, encryption, PKI, device pairing, security tool, or general secure system (See Chapter 8 for a discussion of this classification). The security-usability threat model is a collection of the elements identified during the review. It categorises these elements into security, usability, or both. Usability elements include effectiveness, satisfaction, efficiency, and memorability, accuracy, and knowledge/skill. Elements that affect security of a system include attention, vigilance, conditioning, motivation, context, memorability, and knowledge/skill.

A usability evaluation of a secure system, therefore, focusses on one or more of these elements — do users achieve their primary goals (effectiveness) within reasonable time (efficiency)? are they

happy with the use of the system (satisfaction)? are they able to recall required information? can they provide required information accurately? or what knowledge/skill level is required to use the system? For security evaluation of a system; does security depend on users not being disrupted (attention) every time (vigilance)? what level of user motivation is required? can users keep security secrets (memorability) without disclosing them (social context)? can users form a correct mental model of the system (knowledge/skill)? or can they execute security tasks without paying attention (condition)?

The model is useful for systematically analysing HISPs' security and usability by identifying elements that may aid in compromising security or reduce usability. The elements identified during the analysis using the model are employed in formulating usage and threat scenarios. Usage scenarios are used to identify areas that may hinder the usability of a system, whereas threat scenarios are used to identify areas that may help non-malicious users to break the security of a system. When a system's threat scenarios are more usable compared to usage scenarios, users are more likely to perform the former. A user study is recommended for evaluating usage and threat scenarios to identify problem areas in the system. It is possible, however, for an experienced researcher to reason about the scenarios and make general recommendation.

9.3 Future directions

The HISPs framework presented in Chapter 7 was solely validated using experts because it is currently difficult and prohibitively costly to carry out longitudinal studies. Longitudinal studies are will provide insight on changes in perception of empirical channels as users interact with them over time in different physical and social environments. It is difficult to capture longitudinal information in a laboratory experiment and, therefore, it is crucial that real world studies are conducted when HISPs are deployed or the cost of conducting such is justifiable. The information gained from such studies will be valuable to enhancing the framework.

There has been no work, prior to this thesis, proposing a methodology specifically for evaluating security and usability of secure systems. We have proposed the model for analysing, and process for evaluating, secure systems. We demonstrated how the model and process can be used to analyse and evaluate HISPs. Data that can enhance both the model and process can be obtained through

their application to different types of secure system. More effort should be put into this direction It, therefore, requires further analysis and validation for it to be robust. This may be made possible through input from other researchers as they apply it to different systems.

Bibliography

[1] A. Adams and M. A. Sasse. Users Are Not the Enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[2] I. Alexander and M. Neil. *Scenarios, Stories and Use Cases*. John Wiley, 2004.

[3] W.-F. Alliance. Wi-Fi CERTIFIED for Wi-Fi Protected Setup: Easing the User Experience for Home and Small Office Wi-Fi Networks. Technical report, Wi-Fi Alliance, 2007.

[4] R. Anderson. Why Cryptosystems Fail. *CCS '93: Proceedings of the 1st ACM conference on Computer and communications security*, pages 215–227, 1993.

[5] R. Anderson and T. Moore. The Economics of Information Security. *Science*, 314(5799):610–613, 2006.

[6] D. Balfanz, G. Durfee, D. Smetters, and R. Grinter. In Search of Usable Security: Five Lessons From the Field. *IEEE Security & Privacy*, 2(5):19–24, 2004.

[7] D. Balfanz, D. K. Smetters, P. Stewart, and H. C. Wong. Talking to Strangers: Authentication in ad-hoc Wireless Networks. In *Symposium on Network and Distributed Systems Security (NDSS '02), San Diego, California*, 2002.

[8] B. Beckles, V. Welch, and J. Basney. Mechanisms for Increasing the Usability of Grid Security. *International Journal of Human-Computer Studies*, 63(1-2):74 – 101, 2005. HCI research in privacy and security.

[9] A. H. Betiol and de Abreu. Usability Testing of Mobile Devices: A Comparison of Three Approaches. In *Human-Computer Interaction*. INTERACT, 2005.

[10] P. Beynon-Davies. *Information Systems: An Introduction to Informatics in Organisations*. Palgrave, 2002.

[11] J. Brooke. SUS: A Quick and Dirty Usability Scale. In P. W. Jordan, B. Weerdmeester, A. Thomas, and I. L. Mclelland, editors, *Usability Evaluation in Industry*, pages 189–194. Taylor and Francis, London, 1996.

[12] A. Brostoff. *Improving Password System Effectiveness*. PhD thesis, University of London, 2004.

[13] S. Brostoff and M. A. Sasse. Are Passfaces More Usable Than Passwords? A Field Trial Investigation. In *Proceedings of HCI 2000*, pages 405–424, Sunderland, U.K., Sept. 5-8 2000. Springer.

[14] S. Brostoff and M. A. Sasse. "Ten Strikes and you're out": Increasing the Number of Login Attempts can Improve Password Usability. In *Proceedings of CHI 2003 Workshop on HCI and Security Systems*, Ft. Lauderdale, Florida, 2003. John Wiley.

[15] M. Čagalj, S. Čapkun, and J. Hubaux. Key Agreement in Peer-to-Peer Wireless Networks. In *Proceedings of the IEEE (Special Issue on Cryptography and Security)*. IEEE, 2006.

[16] J. P. Chin, V. A. Diehl, and K. L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *CHI '88: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–218, New York, NY, USA, 1988. ACM.

[17] L. Cranor and S. Garfinkel. *Security and Usability: Designing Secure Systems That People Can Use*. O'Reilly Media, Inc., 2005.

[18] L. F. Cranor. A Framework for Reasoning About the Human in the Loop. In *UPSEC'08: Proceedings of the 1st Conference on Usability, Psychology, and Security*, pages 1–15, Berkeley, CA, USA, 2008. USENIX Association.

[19] K. Crisler, T. Turner, A. Aftelak, M. Visciola, A. Steinhage, M. Anneroth, M. Rantzer, B. von Niman, A. Sasse, M. Tscheligi, S. Kalliokulju, E. Dainesi, and A. Zucchella. Considering the User in the Wireless World. *Communications Magazine, IEEE*, 42(9):56–62, Sept. 2004.

[20] F. D. Davis. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 12(Sept.):319–339, September 1989.

[21] A. K. Dey. Understanding and Using Context. *Personal Ubiquitous Computing*, 5(1):4–7, 2001.

[22] R. Dhamija, J. D. Tygar, and M. Hearst. Why Phishing Works. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590, New York, NY, USA, 2006. ACM.

[23] P. DiGioia and P. Dourish. Social Navigation as a Model for Usable Security. In *SOUPS '05: Proceedings of the 2005 symposium on Usable privacy and security*, pages 101–108, Pittsburgh, Pennsylvania, 2005. ACM.

[24] D. Dolev and A. Yao. On the Security of Public Key Protocols. In *Information Theory*, volume 29(2), pages 198–208, 1983.

[25] P. Dourish, E. Grinter, J. Delgado de la Flor, and M. Joseph. Security in the Wild: User Strategies for Managing Security as an Everyday, Practical Problem. *Personal Ubiquitous Comput.*, 8(6):391–401, 2004.

[26] H. B.-L. Duh, G. C. B. Tan, and V. H.-h. Chen. Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. In *MobileHCI '06: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 181–186, New York, NY, USA, 2006. ACM.

[27] J. F. Dumas and J. C. Redish. *A Practical Guide to Usability Testing.* Greenwood Publishing Group Inc., Westport, CT, USA, 1993.

[28] J. S. Dumas. User-Based Evaluations. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, pages 1093–1117, 2003.

[29] L. M. Feeney, B. Ahlgren, and A. Westerlund. Demonstration Abstract: Spontaneous Networking for Secure Collaborative Applications in an Infrastructureless Environment. In *International conference on pervasive computing (Pervasive 2002)*, Zurich, Switzerland, 2002.

[30] A. Field and G. Hole. *How to Design and Report Experiments.* SAGE Publications Inc, 2003.

[31] I. Flechais. *Designing Secure and Usable System.* PhD thesis, University of London, 2005.

[32] I. Flechais, J. Riegelsberger, and M. A. Sasse. Divide and Conquer: The Role of Trust and Assurance in the Design of Secure Socio-Technical Systems. In *NSPW '05: Proceedings of the 2005 workshop on New security paradigms*, pages 33–41, California, USA, 2005. ACM.

[33] I. Flechais and A. Sasse. *Security and Usability: Designing Secure Systems that People Can Use*, chapter Usable Security: Why Do We Need It? How Do We Get It? O'Reilly, 2005.

[34] I. Flechais, M. A. Sasse, and S. M. V. Hailes. Bringing Security Home: A Process for Developing Secure and Usable Systems. In *NSPW '03: Proceedings of the 2003 workshop on New security paradigms*, pages 49–57, Ascona, Switzerland, 2003. ACM.

[35] S. Gaw, E. W. Felten, and P. Fernandez-Kelly. Secrecy, Flagging, and Paranoia: Adoption Criteria in Encrypted Email. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 591–600, Montréal, Québec, Canada, 2006. ACM.

[36] C. Gehrmann, C. J. Mitchell, and K. Nyberg. Manual Authentication for Wireless Devices. In *RSA Cryptobytes*, volume 7(1), pages 29–37. RSA Security, Spring 2004.

[37] D. Gollmann. *Computer Security.* John Wiley & Sons, Ltd, 2006.

[38] N. S. Good and A. Krekelberg. Usability and Privacy: A Study of Kazaa P2P File-sharing. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 137–144, Ft. Lauderdale, Florida, USA, 2003. ACM.

[39] M. Goodrich, M. Sirivianos, J. Solis, G. Tsudik, and E. Uzun. Loud and Clear: Human-Verifiable Authentication Based on Audio. In *Proc. 26th IEEE International Conference on Distributed Computing Systems ICDCS 2006*, pages 10–10, Lisbon, Portugal, 04–07 July 2006.

[40] B. S. I. Group. Simple Pairing White Paper. www.bluetooth.com/NR/rdonlyres/0A0B3F36-D15F-4470-85A6-F2CCFA26F70F/0/ SimplePairing_WP_V10r00.pdf.

[41] P. Gutmann. Security Usability Fundamentals: http://www.cs.auckland.ac.nz/ pgut001/pubs/usability.pdf. Accessed 16-06-2010, 2010.

[42] N. M. Haller. The S/KEY One-time Password System. In *Proceedings of the Symposium on Network and Distributed System Security*, pages 151–157, 1994.

[43] V. Henderson-Summet and J. Clawson. Usability at the Edges: Bringing the Lab into the Real World and the Real World into the World. In *INTERACT*, 2007.

[44] A. I. From Intentions to Actions: The Theory of Planned Behavior. In J. Kuhl and J. Beckmann, editors, *Action Control: From Cognition to Behavior*, pages 11–39. Springer Verlag, New York, 1985.

[45] I. Ion, M. Langheinrich, and P. Kumaraguru. Influence of User Perception, Security Needs, and Social Factors on Device Pairing Method Choices. In *SOUPS '10: Proceedings of the 5th symposium on Usable privacy and security*, Redmond, USA, 2010.

[46] M. Jakobsson and S. Wetzel. Security Weaknesses in Bluetooth. In *Lecture Notes in Computer Science*, volume 2020, pages 176+, 2001.

[47] S. Jeff. and E. Kindlund. How Long Should a Task? Identifying Specification Limits for Task Times in Usability Tests. In *In Proceeding of the Human Computer Interaction International Conference HCII 2005), Las Vegas*, 2005.

[48] U. Jendricke, U. Jendricke, and D. Gerd tom Markotten. Usability meets security - the identity-manager as your personal security assistant for the internet. In D. Gerd tom Markotten, editor, *Proc. 16th Annual Conference Computer Security Applications ACSAC '00*, pages 344–353, Miami Beach, Florida, 2000.

[49] M. Jones and G. Marsden. *Mobile Interaction Design.* John Wiley & Sons, 2006.

[50] P. Juola. Whole-word Phonetic Distance and the PGPfone Alphabet. In *Fourth International Conference on Spoken Language Processing (ICSLP) 96*, volume 1, pages 98–101, Oct 1996.

[51] A. Kaikkonen, A. Kekalainen, M. Cankar, T. Kaillio, and A. Kankainen. *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, chapter Will Laboratory Test Results be Valid in Mobile Contexts? IGI Global, 2008.

[52] R. Kainda. Human Factors in HCBK Protocols. Master's thesis, Oxford University, Oxford, UK, 2007.

[53] R. Kainda, I. Flechais, and A. Roscoe. Usability and Security of Out-Of-Band Channels in Secure Device Pairing Protocols. In *SOUPS '09: Proceedings of the 5th symposium on Usable privacy and security*, Mountain View, USA, 2009. ACM.

[54] R. Kainda, I. Flechais, and A. Roscoe. *Information Security Theory and Practice. Security and Privacy of Pervasive Systems and Smart Devices*, volume 6033 of *WISTP 2010, Lecture Notes in Computer Sciences*, chapter Secure and Usable Out-Of-Band Channels for Ad hoc Mobile Device Interactions, pages 308–315. Springer, Passau, Germany, 4 2010.

[55] R. Kainda, I. Flechais, and A. Roscoe. Security and Usability: Analysis and Evaluation. *International Conference on Availability, Reliability and Security Krakow, Poland*, pages 275–282, 2010.

[56] R. Kainda, I. Flechais, and A. Roscoe. Two Heads are Better Than One: Security and Usability of Device Associations in Group Scenarios. In *SOUPS '10: Proceedings of the 6th symposium on Usable privacy and security*, Redmond, USA, 2010.

[57] R. Kazman, G. Abowd, L. Bass, and P. Clements. Scenario-Based Analysis of Software Architecture. *IEEE Softw.*, 13(6):47–55, 1996.

[58] J. Kirakowski and M. Corbett. SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology*, 24(3):210–212, 1993.

[59] J. Kjeldskov, J. Kjeldskov, M. B. Skov, B. S. Als, and R. T. Høegh. Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Proceedings of the 6th International Mobile HCI 2004 conference*, pages 61–73. LNCS, Springer-Verlag, 2004.

[60] A. Kobsa, R. Sonawalla, G. Tsudik, E. Uzun, and Y. Wang. Serial Hook-ups: A Comparative Usability Study of Secure Device Pairing Methods. In *SOUPS '09: Proceedings of the 5th symposium on Usable privacy and security*, Mountain View, California, 2009.

[61] A. Kumar, N. Saxena, and E. Uzun. Alice Meets Bob: A Comparative Usability Study of Wireless Device Pairing Methods for a "Two-User" Setting. *CoRR*, 2009.

[62] A. Kumar, G. Tsudik, and E. Uzun. Caveat Emptor: A Comparative Study of Secure Device Pairing Methods. In *International Conference on Pervasive Computing and Communications*, pages 1 – 10, Galveston, Texas, 2009.

[63] C. Kuo, S. Romanosky, and L. F. Cranor. Human Selection of Mnemonic Phrase-Based Passwords. In *SOUPS '06: Proceedings of the second symposium on Usable privacy and security*, pages 67–78, Pittsburgh, Pennsylvania, 2006. ACM.

[64] C. Kuo, J. Walker, and A. Perrig. Low-Cost Manufacturing, Usability, and Security: An Analysis of Bluetooth Simple Pairing and Wi-Fi Protected Setup. In S. Dietrich and R. Dhamija, editors, *Financial Cryptography*, volume 4886 of *Lecture Notes in Computer Science*, pages 325–340. Springer, 2007.

[65] J. Leach. Improving User Security Behaviour. *Computers & Security*, 22(8):685 – 692, 2003.

[66] J. R. Lewis. Psychometric Evaluation of an After-Scenario Questionnaire for Computer Usability Studies: the ASQ. *SIGCHI Bull.*, 23(1):78–81, 1991.

[67] J. R. Lewis. Psychometric Evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 36:1259–1263(5), 1992.

[68] J. R. Lewis. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *Int. J. Hum.-Comput. Interact.*, 7(1):57–78, 1995.

[69] J. Mangan, C. Lalwani, and B. Garner. Combining Quantitative and Qualitative Methodologies in Logistics Research. *International Journal of Physical Distribution & Logistics Management*, 34(7):565–578, 2004.

[70] R. Mayrhofer and H. Gellersen. Shake Well Before Use: Authentication Based on Accelerometer Data. In *Proc. Pervasive 2007: 5th International Conference on Pervasive Computing*, volume 4480 of *LNCS*, pages 144–161. Springer-Verlag, May 2007.

[71] R. Mayrhofer and M. Welch. A Human-Verifiable Authentication Protocol Using Visible Laser Light. In *ARES '07: Proceedings of the The Second International Conference on Availability, Reliability and Security*, pages 1143–1148, Washington, DC, USA, 2007. IEEE Computer Society.

[72] J. McCune, A. Perrig, and M. Reiter. Seeing-is-Believing: Using Camera Phones for Human-Verifiable Authentication. In *Proc. IEEE Symposium on Security and Privacy*, pages 110–124, 8–11 May 2005.

[73] G. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information, 1956. One of the 100 most influential papers in cognitive science: http://cogsci.umn.edu/millennium/final.html.

[74] A. Minke. Conducting Repeated Measures Analyses: Experimental Design Considerations. Technical report, Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 23-25, 1997), 1997.

[75] K. D. Mitnick and W. L. Simon. *The Art of Deception: Controlling the Human Element of Security.* John Wiley & Sons, Inc., New York, NY, USA, 2003.

[76] W. Moncur and G. Leplâtre. Pictures at the ATM: Exploring the Usability of Multiple Graphical Passwords. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 887–894, New York, NY, USA, 2007. ACM.

[77] N. P. Moray and B. M. Huey, editors. *Human Factors Research and Nuclear Safety.* National Academy Press, Washington, DC, USA, 1988.

[78] B.-Y. Ng, A. Kankanhalli, and Y. C. Xu. Studying Users' Computer Security Behavior: A Health Belief Perspective. *Decision Support Systems*, 46(4):815 – 825, 2009. IT Decisions in Organizations.

[79] L. H. Nguyen and A. W. Roscoe. Efficient Group Authentication Protocol Based on Human Interaction. In *Proceedings of the Workshop on Foundation of Computer Security and Automated Reasoning Protocol Security Analysis (FCS-ARSPA)*, pages 9–33, 2006.

[80] L. H. Nguyen and A. W. Roscoe. Authenticating ad hoc Networks by Comparison of Short Digests. In *Journal of Information and Computation. Special Issue of Information and Computation on Computer Security: Foundations and Automated Reasoning*, 2007.

[81] C. M. Nielsen, M. Overgaard, M. B. Pedersen, J. Stage, and S. Stenild. It's Worth the Hassle!: The Added Value of Evaluating the Usability of Mobile Systems in the Field. In *NordiCHI '06: Proceedings of the 4th Nordic conference on Human-computer interaction*, pages 272–280, New York, NY, USA, 2006. ACM.

[82] J. Nielsen. *Usability Engineering*. Boston; London : Academic Press, 1993.

[83] J. Nielsen. Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier. *Cost-justifying usability*, pages 245–272, 1994.

[84] J. Nielsen and J. Levy. Measuring Usability: Preference vs. Performance. *Commun. ACM*, 37(4):66–75, April 1994.

[85] R. Nithyanand, G. Tsudik, and E. Uzun. Groupthink: On the Usability of Secure Group Association of Wireless Devices. In *15th European Symposium on Research in Computer Security (ESORICS'10)*, Anthens, Greece, 2010.

[86] T. I. S. Organisation. Ergonomic Requirements for Office Work with Visual Display Terminals, ISO 9241-11, 1998.

[87] Owen. Zxing: Multi-format 1d/2d barcode image processing library with clients for android, java, and iphone project: http://code.google.com/p/zxing/.

[88] A. Perrig and D. Song. Hash Visualization: a New Technique to improve Real-World Security. In *International Workshop on Cryptographic Techniques and E-Commerce (CrypTEC '99)*, pages 131–138, 1999.

[89] D. E. Perry, A. A. Porter, and L. G. Votta. Empirical Studies of Software Engineering: A Roadmap. In *ICSE '00: Proceedings of the Conference on The Future of Software Engineering*, pages 345–355, New York, NY, USA, 2000. ACM.

[90] R. Prasad and N. Saxena. Efficient Device Pairing Using "Human-Comparable" Synchronized Audiovisual Patterns. In *ACNS*, 2008.

[91] K. Ricks and B. Arnoldy. How to Conduct Your Own Usability Study. In *Proc. IEEE International Professional Communication Conference IPCC 2002*, pages 115–126, 17–20 Sept. 2002.

[92] A. W. Roscoe. Human-Centred Computer Security. Unpublished draft, 2006.

[93] A. W. Roscoe, S. Creese, M. Goldsmith, and I.Zakiuddin. The attacker in ubiquitous computing environments: formalising the threat model. In *Proceedings of FAST 2003, Pisa*, 2003.

[94] A. W. Roscoe, S. J. Creese, M. H. Goldsmith, and M. Xiao. Bootstrapping Multi-Party Ad-Hoc Security. In *Proceedings of SAC*, 2006.

[95] J. Rubin. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[96] M. A. Sasse. Computer Security: Anatomy of a Usability Disaster, and a Plan for Recovery. In *Proceedings of CHI2003 Workshop on Human-Computer Interaction and Security Systems*, 2003.

[97] M. A. Sasse. Red-Eye Blink, Bendy Shuffle, and the Yuck Factor: A User Experience of Biometric Airport Systems. *IEEE Security and Privacy*, 5(3):78–81, 2007.

[98] M. A. Sasse, S. Brostoff, and D. Weirich. Transforming the 'Weakest Link' — a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal*, 19(3):122–131, 2001.

[99] J. Sauro and E. Kindlund. A Method to Standardize Usability Metrics into a Single Score. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–409, New York, NY, USA, 2005. ACM.

[100] N. Saxena, J.-E. Ekberg, K. Kostiainen, and N. Asokan. Secure Device Pairing based on a Visual Channel (Short Paper). In *SP '06: Proceedings of the 2006 IEEE Symposium on Security and Privacy*, pages 306–313, Washington, DC, USA, 2006. IEEE Computer Society.

[101] N. Saxena, B. Uddin, and V. Jonathan. Universal Device Pairing Using an Auxiliary Device. In *Symposium on Usable Privacy and Security (SOUPS)*, Pittsburgh, Pennsylvania, July 2008. ACM.

[102] N. Saxena and M. B. Uddin. Secure Pairing of "Interface-Constrained" Devices Resistant against Rushing User Behavior. In *ACNS*, pages 34 – 52, 2009.

[103] N. Saxena and J. Voris. Pairing Devices with Good Quality Output Interfaces. In *ICDCS Workshops*, 2008.

[104] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The Emperor's New Security Indicators. In *SP '07: Proceedings of the 2007 IEEE Symposium on Security and Privacy*, pages 51–65, Washington, DC, USA, 2007. IEEE Computer Society.

[105] B. Schneier. Biometrics: Truths and Fictions. Crypto-Gram Newsletter, August 15, 1998.

[106] B. Schneier. Security in The Real-World: How to Evaluate Security Technology, 1999.

[107] B. Schneier. *Secrets & Lies: Digital Security in a Networked World*. John Wiley & Sons, Inc., New York, NY, USA, 2000.

[108] B. Schneier. *Beyond Fear: Thinking Sensibly about Security in an Uncertain World*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

[109] D. Schuler and A. Namioka. *Participatory Design: Principles and Practices*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1993.

[110] B. Shackel. Usability—context, framework, definition, design and evaluation. *Human factors for informatics usability*, pages 21–37, 1991.

[111] H. Sharp, Y. Rogers, and J. Preece. *Interaction Design: Beyond Human-Computer Interaction.* Wiley, 2 edition, March 2007.

[112] B. Shneiderman and C. Plaisant. *Designing the User Interface: Strategies for Effective Human-Computer Interaction (4th Edition).* Pearson Addison Wesley, 2004.

[113] C. Soriente, G. Tsudik, and E. Uzun. BEDA: Button-Enabled Device Association. In *International Workshop on Security for Spontaneous Interaction (IWSSI)*, Innsbruck, Austria, 2007.

[114] C. Soriente, G. Tsudik, and E. Uzun. HAPADEP: Human-Assisted Pure Audio Device Pairing. In *ISC '08: Proceedings of the 11th international conference on Information Security*, pages 385–400, Berlin, Heidelberg, 2008. Springer-Verlag.

[115] J. H. Spyridakis. Conducting Research in Technical Communication: The Application of True Experimental Designs. *Technical Communication*, v39:p607–624, 1992.

[116] R. Stallman. Can You Trust Your Computer. http://www.gnu.org/philosophy/can-you-trust.html, 21 October 2002. Accessed 15 March 2010.

[117] I. Standards. QRCODE Standard: ISO/IEC18004.

[118] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cronor. Crying wolf: An Empirical Study of SSL Warning Effectiveness. In *Usenix Security*, 2009.

[119] S. Systems. JavaME at a Glance: http://java.sun.com/javame/index.jsp.

[120] S. Systems. MIDP Specification: http://java.sun.com/products/midp/.

[121] R. W. Taylor and Y. Yu. An SMS history. In *In Mobile world*, pages 75–91. Springer, 2005.

[122] T. Tullis and B. Albert. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics (Interactive Technologies): Collecting, Analyzing, and Presenting ... Kaufmann Series in Interactive Technologies).* Morgan Kaufmann, 2008.

[123] E. Uzun, K. Karvonen, and N. Asokan. Usability Analysis of Secure Pairing Methods. In *Financial Cryptography and Data Security*, pages 307–324, 2007.

[124] S. Vaudenay. Secure Communications over Insecure Channels Based on Short Authenticated Strings. In *Lecture Notes in Computer Science*, volume 3621, pages 309–326, November 2005.

[125] M. Čagalj, N. Saxena, and E. Uzun. On the Usability of Secure Association of Wireless Devices Based on Distance Bounding. In *CANS*, 2009.

[126] D. Weirich and M. A. Sasse. Pretty Good Persuasion: A First Step Towards Effective Password Security in the Real World. In *NSPW '01: Proceedings of the 2001 workshop on New security paradigms*, pages 137–143, New York, NY, USA, 2001. ACM.

[127] M. Wenger and J. Spyridakis. The Relevance of Reliability and Validity to Usability Testing. *Professional Communication, IEEE Transactions on*, 32(4):265–271, Dec 1989.

[128] A. Whitten. *Making Security Usable.* PhD thesis, Carnegie Mellon University, 2004.

[129] A. Whitten and J. Tygar. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *Proceedings of the 8th USENIX Security Symposium, August 1999, Washington*, pages 169–183, 1999.

[130] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon. Authentication Using Graphical Passwords: Effects of Tolerance and Image Choice. In *SOUPS '05: Proceedings of the 2005 symposium on Usable privacy and security*, pages 1–12, Pittsburgh, Pennsylvania, 2005. ACM.

[131] W. E. Woodson. *Human Factors Design Handbook: Information and Guidelines for the Design of Systems, Facilities, Equipment, and Products for Human Use.* McGraw-Hill, 1981.

[132] J. Yan, Alan, Ross, and Alasdair. Password Memorability and Security: Empirical Results. *IEEE Security and Privacy*, 2:25–31, 2004.

[133] K.-P. Yee. User Interaction Design for Secure Systems. In *ICICS '02: Proceedings of the 4th International Conference on Information and Communications Security*, pages 278–290, London, UK, 2002. Springer-Verlag.

[134] K.-P. Yee. Aligning Security and Usability. *IEEE Security & Privacy*, 2(5):48–55, 2004.

Appendix A

HISPs framework validation - expert feedback

Expert feedback - strengths

- Considers all channels/threats to security e.g. technology, user
- Taking into account the mental load of users performing an authentication procedure. Placing users/human factors in the focus (Fig. 1). Talking about social norms and including it in the environment part of the framework - I would like to see this expanded, as it seems highly relevant. Multi-user scenarios might pose novel issues, so having them in the model is good.
- The framework offers a comprehensive enumeration of contextual factors that may influence the effectiveness of a human interactive security protocol (HISP). It offers designers and researchers a concise checklist of issues that should be considered during the design and/or evaluation of a HISP.
- Generally its a good idea to have methodologies that can help to people to focus on BOTH usability and security - think AEGIS. Very good that this is adaptable to different scenarios and devices. Excellent that it makes the point (well known but not always followed) that users failure to follow unusable technical fixes is not a human problem the user is not the problem. So, its good to keep technical security separate its needed, but its not the answer, and its useless if a user cant use it.
- A major strength is application of the UCD method to a security problem. This has not been done before. Frameworks so far have been very abstract, and actual research results have been point solutions for a specific problem in a specific scenario. The work has great potential for being used in practice. Device manufacturers have already problems setting up secure usable pairing between two devices, let alone between many devices. Id like to note that the background work is very thorough
- The proposal is based on a solid understanding of the current problems in how security is offered to end users. It aims at improvements on both the underlying technical solutions as well as on the user interface level; a combined approach is probably the best way to go. If the underlying technical solutions are crafted with an understanding of the cognitive issues that effect their understandability and easiness of use in short, the human aspects in order to reflect on these issues, the better. The framework aims to bring together and add to the existing body of work on understanding the related issues in secure device pairing and its usability aspects
- The main strengths are that the framework involves a comprehensive set of factors involved in the success of a system that uses empirical channels for security in ad-hoc networks. In particular, I was pleased to see the human factors identified at some length, including even social and affective issues. To date, the human side of ad-hoc network security has been dealt with only lightly, and work is still being done that does not adequately address important factors that will limit any practical success. This framework may increase awareness of the issues, and help researchers and practitioners better determine the success of new proposals before actual trials. The framework might also be used to construct heuristic evaluation methods like those of Nielsen for methodical but quick and low-cost evaluation of proposals.

Expert feedback - contribution

- Contributes by including more information with regard to user behaviour
- The framework is an important step in summarizing existing work on empirical authentication protocols, and offers researchers new to the field a concise introduction into the various relevant factors
- Useful attempt to apply HCISec thinking to non-desktop scenarios, and timely, with new applications requiring secure communication. Eg.if payment using devices such as RFID tags becomes more common, how do I know I am really sending my micropayment to the PoS terminal and not to some rogue device? This is my first introduction to HISP. From what I have seen, Im not convinced that HISP is the solution to the problem of securing ad-hoc networks, but the problem is likely to become more urgent, so applying HCI thinking to the problem is welcome.
- Structuring the thinking around secure pairing taking into account different external factors. Ideally it could develop a method to give a numerical value to a security method as used in the wild. Another major contribution may come from doing solid user studies.
- The proposal sets up a few questions on the usability issues of the research problem that if answered in a good way through the work, will greatly benefit the field of usable security, as they represent major challenges. Will users compare the values accurately? Currently they dont in such situations. Will they bother to compare and not skip this step? No. Can they be duped into pairing devices whose displayed information does not match? Yes, easily. How large a value of b can they effectively deal with? Remember the magical number 7 to start with, also understanding of underlying cognitive issues should be well taken into account. Building the technical solution up with solid user research work is likely to lead to a more usable security management, which would be highly desirable from the research point of view.
- This appears to be a reasonable scheme for evaluation and for preparation for further work that involves comparison, design, exploratory study, and experiments. As such, I think the contribution is primarily in its role of supporting other work. By itself, the contribution does not appear strong, relying as it does on assembling standard elements from other work, and with only the UCD model to tie them together.

Expert feedback - practicality

- Practical for both as highlights issues that need to be considered from different groups
- Practicality for researchers: For researchers, this model might help to compare new methods with previously published ones, that is, to structure related work. I think it works well for this case. Practicality for designers: The current discourse misses more explicit design decisions, so that it is doubtful if system designers without in-depth knowledge of authentication protocols and the issues listed in the model/framework can derive hints from it. Researchers already aware of the issues benefit from the structure, but application designers/developers who only "need some security" will probably not benefit from it in its current state.
- While its application seems relatively straightforward, the actual weighing of the different factors is rather undefined. This might actually be a strength, as an experienced researcher and/or designer might understand how to do this. However, a beginner in the field might not have enough knowledge to understand the implications and importance of some of these choices.
- You should ask some designers how practical it is! It seems quite practical although it looks like it will nearly always produce MCE as the best solution. Maybe the framework should help people to consider that the best solution would be context-dependent? Eg., in the meeting case study, they could use SDC when they are in one room and only have to resort to MCE when in different cities. Hence, Im not sure that the idea of always taking the worst-case is appropriate.
- As alluded to, if successful the results will be very valuable to device manufacturers and UI designers. Usefulness will critically depend on thorough user studies
- The work would seem to be useful for both researchers and designers: for researchers, there could be a deeper understanding of the underlying issues and how the current methodologies work; for designers, the work could provide tools and guidelines to implement more usable security. The mobility aspect has been under researched in current body of work, so adding to that corpus of research and hopefully some guidelines will benefit both communities.
- This will be helpful, but primarily for those researchers and designers working from a technical security perspective, because it will alert them to the numerous human-factors issues involved in the success of empirical as an approach to ad-hoc network security. For HCI researchers or industrial product designs, I would expect more familiarity with the issues highlighted here, and I think the main benefit of the framework would be to serve as a kind of checklist.

Appendix B

Proforma for collection of expert feedback

Criteria	Description	Comment
Strengths	What the are the strengths of the proposed framework that will make it achieve its intended goal of helping in designing or choosing OOB channels that fit human and contextual needs	
Weaknesses	What the are the pitfalls of the proposed framework that will prevent it from achieving its intended goal	
Contribution	As an academic/research contribution, what is the significance of the proposed framework	
Practicality	The framework is aimed for both researchers and designers - how practical is it to both user groups	
Improvements	How can the proposal be improved to achieve its goals	
Future direction	How can the proposal be extended/modified to apply to other areas of security other than device associations	
Other comments		

Appendix C

ISUT: A Tool for Security and Usability Testing of Device Association Protocols

C.1 Introduction

The small size of mobile device interfaces and the distributed nature of device association makes it nearly impossible to employ direct observation or video recording to accurately capture user interactions. In this chapter, we present the Integrated Security and Usability Testing (ISUT) tool that automates the capturing of user interactions in device associations.

The Integrated Security and Usability Testing tool was developed using Java 2 Micro Edition (J2ME) for use in security and usability testing of empirical channels. It generates test cases, logs user events, presents participants with realistic usage scenarios, and is configurable to specific test requirements, test design, security protocol, and prototype applications.

C.2 ISUT main components

The tool is divided into four layers:

- protocol - a layer that implements HISPs
- framework - a layer that implements communication, configuration, and logging modules
- evaluation - a layer that binds user interfaces with specific evaluation criteria
- application - a layer that implements prototype application

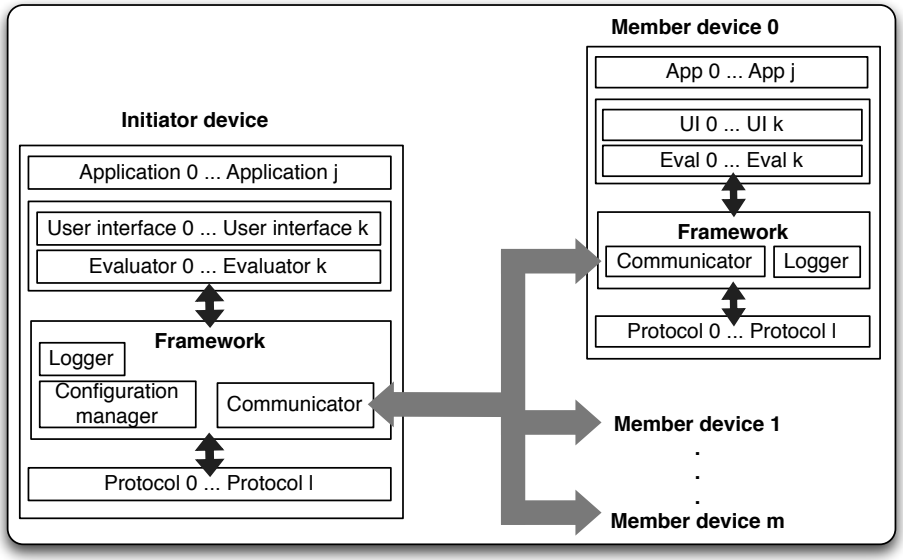

Figure C.1: ISUT: tool support for security and usability testing

C.2.1 Protocol layer

The protocol layer enables the testing of the performance of specific HISPs. In its current state, the tool implements the SHCBK protocol as the protocol was our focus and a representative of HISPs in our studies. Other protocols, however, can be implemented on the same layer and, through the configuration user interface, one can choose which protocol to use for a particular case.

C.2.2 Framework layer

The framework has two main functions; handling communication between devices and coordinating how other components of the tool work together. The current implementation of the framework only supports Bluetooth communication. It handles Bluetooth device and service discovery (on device configured as initiator), responds to connection requests (when configured as member device), and handles subsequent communication between devices including that requested by protocols and applications.

The framework coordinates other components of the tool. A user passes configurations to the framework through a user interface. The framework uses these configurations to determine what protocol to run, how tested methods are presented, and what application should be presented to participants as a primary task. It chooses a user interface—through which the user interacts with the protocol—based on the empirical channel tested.

C.2.3 Evaluation layer

C.2.3.1 User interface

An interface through which users interact with protocols and applications, and through which configurations are set. The framework determines which interface to invoke based on the type of digest transferred or compared and permits individual applications to control the user interface. Each interface requires an *evaluator* that correctly evaluates user's interactions with that interface.

C.2.3.2 Evaluator

An evaluator captures specific user interactions with an interface and evaluates the correctness or otherwise of those actions based on specified criteria. Interactions that are of particular interest to the evaluator are those related to the transferring or comparing of digests. The evaluator captures these actions, evaluates them, and passes the outcome to the framework. The framework, based on the configuration, either saves this information into a log file or passes it to a framework running on a different device that eventually save the data.

C.2.4 Application layer

Users invoke prototype applications through the framework based on configurations and the result of the evaluator. A test conductor configures what applications can be invoked through out the running cycle of the tool. The framework determines whether to invoke an application or not based on the results of the evaluator–if the evaluator indicates that an association failed, the framework may prompt the user to restart the association. A test conductor may also configure the framework to let a study participant decide whether to invoke an application or not. This is important for methods that require to confirm that a secure communication has been established between devices.

C.2.5 Configuration manager

The strength of the ISUT implementation is also in its dynamic configuration capability. Through a user interface, a test conductor can configure the tool according to requirements. While the current implementation covers basic configurations, it can be extended to cover fine grained configurations to satisfy specific requirements.

Through the user interface, a test conductor can configure what protocol (if there is more than one implementation) the testing should use and can specify whether it is a single user or multi-user scenario. Specifying single or multi-user scenarios allows for custom messages and also how initiator device behaves. The tool can also be configured according to test design; whether the order of the tests is randomised or counterbalanced (for within-subject designs), or specify what tests a group of participants is able to see (for between-subject designs). In addition, one can also specify what applications users interact with. For example, one may wish to test how a specific pairing method may be viewed by participants in a payment application.

C.3 Main features of ISUT

C.3.1 Workload management

ISUT tool draws its strengths from three criteria; distributed workload management, support for usability testing within context, and automated testing. Mobile devices are lacking in terms of computational power and battery life compared to desktop counterparts. Taking this into account, it is essential that computational workload is shared among devices involved in collaborative work such as device association. With this view, the tool distributes evaluation and event logging tasks to all devices involved. This avoids potential bottlenecks that may be introduced by having a single device perform the evaluation and logging of events from all other devices especially when several devices are involved.

C.3.2 Usability testing within context

Usability testing is meaningful if it is conducted within context. While many user studies of device association are conducted without specific context, our studies found that the context within which usability testing is conducted has an effect on how participants judge methods tested. The tool supports usability testing within context by adding an application layer above the framework. For usability testing purposes, applications need not be fully functional. They may only be prototypes with fully developed interfaces for participants to interact with. For example, in one of our studies, we used a prototype implementation of a peer-to-peer payment system in which participants played a role of payee making a payment from a personal device to another.

C.3.3 Large scale testing

As mobile devices are now pervasive, there is need for large scale usability testing of applications related to them. This is common practice especially for usability testing of websites. However, large scale testing can only be possible if tests can be undertaken by participants without requiring presence of a test monitor. The tool supports large scale testing by automating test sessions. People will participate in the testing not simply because they want to but because they want to use underlying applications developed on top of the framework. While these applications are used, the tool creates a log of user interactions with tested methods.

C.3.4 Protocol performance analysis

Current usability studies of device associations assume that a protocol has no effect on the usability of methods that users directly interact with. This assumption, however, is flawed. For example, during our usability testing participants have expressed concern on the length of time that Bluetooth takes to complete device and service discovery processes. Considering the processing power of most mobile devices, protocol execution time may not be negligible. The tool provides a platform through which usability testing can be conducted more realistically than other tools. It is now possible, using the tool, to conduct a user study with an actual implementation of the protocol and a specific application through which users interact.

C.4 Limitations

Though Java applications are the most portable on mobile devices, they are not the most efficient. This has an impact especially on the underlying protocol implementations and less so on user interfaces. The other limitation is that the system currently relies on Bluetooth for communication. Bluetooth is not as fast as WiFi and the discovery process leaves something to be desired. However, since WiFi is now able to support peer to peer communication between devices, it is possible to implement support for it.

C.5 Related work

There has been little work in the development of tools to support security and usability evaluations of systems. This may partly be because the field of Human Computer Interaction Security (HCISec) is still developing. On the other hand, it may be that no attention is being paid to developing such tools.

One notable work is that of Uzun *et al.* [123]. In their work, they developed a tool for usability analysis of distributed applications. The tool provides a framework for device communication, test evaluation, and event logging. While this tool demonstrates how usability evaluation of distributed applications can be supported, it has a number of shortcomings:

- Lack of flexibility: the tool lacks the flexibility of adapting to different experimental designs. In a usability evaluation, one may choose to randomise tests, present them in a specific order, or use counterbalancing for repeated measures design.
- Focus on security tasks: a well acknowledged property of security is that it is a secondary goal for most users. Users, for example, associate devices only because they want to exchange files or access a service, for example. A usability evaluation of a secure system should focus users on primary tasks, generated from usage scenarios, rather than on security tasks. Presenting security tasks as primary tasks forces participants to focus more on the former task than they normally would in a real world situation.
- Performance and event isolation: the tool uses a client-server model where only one device is responsible for test generations, evaluation, and event logging. There are two problems with this approach. First, as mobile devices have limited memory and processing power, performance is likely to be an issue especially in cases where a large number of devices is involved. Second, in a usability evaluation of a group of people using devices, it is important that events are isolated so as to identify exactly where they occurred. A central logging system may not be able to provide such information which is crucial to identifying usability issues.

Appendix D

Sample test plan

D.1 Introduction

In conducting a usability study, it is crucial to develop a test plan that outlines the main tasks that need to be undertaken before and during the study. It helps to adhere to initial objectives of the study and also to keep track of the tasks that have been completed and those that are still pending. The sample test plan presented here was used for the group association scenario.

D.2 Test plan

D.2.1 Purpose of study

The goal of the study is to evaluate effectiveness of empirical channels in group device association scenarios of Human-Interactive Security Protocols (HISPs). The study will focus on usability and security of methods.

D.2.2 Problem statement

HISPs require users to compare or transfer digests among devices. While studies of the effectiveness of empirical channels have been conducted in single-user pairwise device associations, there has been none on multi-user scenarios. We consider a case of more than 2 devices with each managed by its own user. A multi-user scenario has different challenges on the usability and security of empirical channels. While the tasks may be shared among participants, increasing the number of user/device pairs may increase the number of points at which security may fail. In this study, we aim to study the effectiveness, in terms of usability and security, of empirical channels in group scenarios.

D.2.3 User profiles

In this study, we want to target participants who use mobile devices on a daily basis so that lack of familiarity with devices does not affect the results of the study. We intend to target participants of varying age, education, and professional background.

D.2.4 Study design

To reduce the number of participants required and increase internal validity, a within-subject, counter-balanced, design will be used. Participants will randomly be allocated to one of 10 groups and each group will be given similar primary tasks and test conditions. Participants will be presented with scenarios that require secure device association. Primary tasks will be derived from these scenarios. There will be five primary tasks and 7 test conditions. Test conditions are method-representations and their variants. For each primary task, groups will perform all test conditions.

D.2.4.1 Initiator

Each primary task will have a different initiator —a person who initiates the connection and whose device will display the digest. In all tasks, the initiator's task will be the same: reading or showing a digest to other group members and indicating on a personal device whether all participating devices indicate success or otherwise. The use of different initiators for different primary tasks is aimed at reducing effects that a single initiator may introduce in the association process. With different initiators, effects of a single initiator on the performance of a group are less significant. This way, results of one group can be compared to other groups.

D.2.4.2 Failure conditions

Out of 7 test conditions for each primary task, 2 will be failure conditions. Failure conditions are specifically designed to make the association fail and are used to determine whether the initiator and other members can make the correct decision of aborting and restarting the association. For each method-representation, there will be exactly two error conditions and 5 normal conditions for all the 5 primary tasks. Failure conditions are fewer than normal conditions because, in practice, the former are likely to be much less frequent compared to the latter.

D.2.4.3 Failure presentation:

In a group association, there are three ways in which failure conditions may be presented.

1. All devices have different digests, that is, no two devices have the same digest. This scenario, however, is not realistic in a practical sense and will not be tested in the study.
2. A group is split into two subgroups with each having the same digest that is different from other subgroup. A subgroup may comprise a single participant, which may be initiator or any other member. The chances of an attack being detected when initiator has a different digest from the rest of the group are higher compared to other cases especially if the initiator reads out the digest to other members. In this study, we will use failure conditions where initiator

and all but one members' digests match. To detect and defeat an attack, a member with a non-matching digest has to be alert, paying attention to initiator when a digest is shown or read, and initiator also has to ensure that every member's device indicated success after association. Other scenarios involving different subgroup sizes are future work.

3. A group split into two subgroups with one subgroup consisting of matching digests while the other having devices each with a non-matching digest. This is a slight variation of the previous scenario and will not be tested in the current study.

D.2.4.4 Learning effect

Considering that same test conditions will be used in all 5 primary tasks, it is an opportunity to learn how much of a learning effect that will have on participants and whether it changes their initial reactions to the test conditions. We will capture participants satisfaction scores and preferences ratings of test conditions in the first and final primary task. This data will provide information on whether repetitive use of a method has an effect users' satisfaction and preferences.

D.2.4.5 Independent variable

One independent variable is considered; method-representation. Methods considered under the study are *manual comparison* and *manual copying and entering*. Manual comparison will have three representations —numeric, images, and double-numeric. *Manual copying and entering* will have two representations —word-matching and number-typing.

- *Manual comparison—numeric:* Though *manual comparison* has been found vulnerable to security failures in previous studies of pairwise associations, it is included in this study as a control variable to which other methods can be compared against. This is due to the fact that the method is the most preferred and the highest rated in terms of easy of use.
- *Manual comparison—images:* In our previous study, image comparison was among the methods that participants liked least. It appeared, however, that this was partly due to instructions that were given to participants; to look for differences in the displayed images. This made it a lot harder for similar images as opposed to distinct images—differences were distinct in dissimilar images. While this type of instruction worked very well with other methods, we will change it specifically for images and observe any change in the rating of images.
- *Manual comparison—double comparison:* Based on previous performance, we have included this method to assess its performance in group scenarios.
- *Manual copying and entering—numeric: Manual copying and entering* is not vulnerable to security failures and numeric is the highest rated representation for this method. However, its performance in group associations has never been assessed.
- *Manual copying and entering—word-matching and number-typing:* In a more recent study of pairwise associations, this method was the most preferred and we want to assess its effectiveness in group scenarios.

D.2.4.6 Dependent variables

- Failures - both security and non-security failures

- Efficiency – completion times
- Satisfaction – satisfaction scores
- Effectiveness/accuracy – task success

D.2.5 Participant tasks

Each participant will be required to complete an enrolment questionnaire for demographic information, After Scenario Questionnaire to rate each method representation, and After Experiment Questionnaire to give overall scores and preferences for each method-representation. Each group will perform the same primary tasks and test conditions. Each participant will play the role of initiator in one of the tasks.

D.2.5.1 Primary tasks

- **Messaging**
  Security scenario– You are leaders of an organisation campaigning for large corporate organisations to invest more in green technology and reduce their reliance on activities that worsen environmental pollution. You have been involved in large protests and law enforcement forces are monitoring every move you make for purposes of gathering information on your planned future protests. You have now decided to use your mobile devices to exchange information in a secure way. Because you are worried that the Internet is monitored, you have decided to exchange information using a local network created among your devices. Each of you has specific information to share. This information is useful for organising your next protest event.
- **Ticketing**
  Security scenario– You have randomly been given a cinema ticket each. The issue, however, is that some of the tickets are invalid but none of you knows which is which. Your devices cannot determine the validity of their own tickets but that of tickets from other devices. So the only way to find out which tickets are valid is for each of you to send your ticket to other members of the group for their devices to determine whether your ticket is valid or not. You are concerned that sending your ticket to a wrong device may result in your tickets being stolen. You decide to take turns in validating tickets. All you need is to find that valid ticket that all of you can use to watch one movie at cinema of your choosing.
- **Quiz**
  Security scenario– You are participating in a quiz contest similar to University Challenge. In this particular case, however, each group is given the same set of questions and the quiz is taken using personal mobile devices. As a group, each of you will receive the same question on your mobile device and you are allowed to discuss before submitting a final answer. Each of you is required to submit the agreed answer. Because you are doing the questions on your mobile devices, you are worried that other groups may read your answers. You decide to create a secure association among your devices.
- **Contacts**
  Security scenario– As leaders of an organisation campaigning for large corporate organisations to actively participate in reducing environmental pollution, you want to exchange contact details. Contact details are not only personal but you do not want the police to get hold of them. You decide to share these details securely. Even though you are confident that the security you establish among your devices is good enough for this purpose, you decide to exchange one contact at a time so that should one of your communications be readable to outsiders, only one contact detail will be exposed.

- **Money transfer**
  Security scenario– As a group you have won a sum of money. The money has been electronically transferred to one of you and you decide to share equally. The person with the full amount has agreed to transfer the money to your devices. Though you have been guaranteed security for such mobile to mobile transfers, you decide to be cautious and transfer the money in smaller bits.

D.2.6 Task list

The task list constitute all preparatory work that has to be completed before the actual study commences. Other tasks may be carried out but these are the critical ones. They include preparation of tools to be used and participant recruitment.

D.2.6.1 Preparation of software

We have developed Integrated Security and Usability Testing (ISUT), a tool for conducting usability and security testing of empirical channels. The tool is developed in Java 2 Micro Edition. The decision to develop it in Java was based on the portability of Java on mobile platforms. The system has been tested on Symbian OS based and Blackberry devices. The system supports the testing process by generating test conditions and logging user events. The current implementations relies on Bluetooth communication. As the primary goal of participants will not be to pair devices securely but to some other task, we have developed, on top of the tool, various applications that can be used as primary tasks for participants. Details of current implementation of ISUT can be found in a technical report to be made available soon.

D.2.6.2 Pilot study

A pilot study will be conducted prior to the actual study. This is aimed at ensuring that unclear instructions, flawed error logging, and other unanticipated conditions are identified and corrected before the actual study. The testing will involve recruiting 'casual' participants to carry out the intended usability testing tasks and analysing the results in order to see whether the system behaves correctly. This testing will be repetitive until there is a high level of assurance that the system will behave as expected. Participants at this stage may be employed repetitively as the focus is not to analyse the usability but the correctness of the software and test documents. These participants will be disqualified for the actual study for obvious reasons.

D.2.6.3 Preparation of questionnaires

In order to capture information other than errors and completion times, questionnaires will be used. Three questionnaires will be used;

- Enrolment or pretest questionnaire — to capture participant demographic information.
- After Scenario Questionnaire (ASQ) — to capture participants' rating of methods on three criteria; efficiency, effectiveness, and satisfaction.
- Post-Test or End of Experiment Questionnaire —to capture participants' overall ratings for each method.

D.2.6.4 Recruitment of participants

Participants will be recruited through mailing lists, notice board and web advertisements.

D.2.6.5 Preparation of testing environment

In the initial usability testing, a lab testing environment will be used. Follow up experiments may be taken in other environments, such as coffee shops, streets etc. This task will involve making the test room available and accessible to participants.

D.2.6.6 Testing

This task involves carrying out the actual usability testing. Each group will be allocated a time slot at which they have to come and do the study. Participants will be briefed and given a simple tutorial on the use of mobile devices to be used in the study. During the study, there will be video recording to capture interactions among participants. There will also be a brief discussion after all tasks have been completed.

D.2.7 Data collection and analysis

Quantitative data will largely be collected through logs created by the system. This will constitute time to complete device association—for individual group members—and errors from each of the devices. Data will also be gathered through observation, video and audio recording. Demographic information, participants' ratings of each method-representation, satisfaction, and preferences will be collected through questionnaires. Each participant will contribute 4 scores—as a group member—and 1 score—as initiator—to each method-representation. With 10 groups, there will be 200 member scores and 50 initiator scores.

Data will be analysed first by group per group basis. This will allow to observe group performances and allow for comparison between groups. Given that initiators' effects are minimised by using different initiator for different tasks, groups will be assumed independent. This will allow for amalgamation of data from all groups and analysed as a single dataset. A one-way repeated measure analysis of variance (ANOVA) will be used to analyse the data.

Appendix E

Questionnaires

E.1 After Scenario Questionnaire[1]

MANUAL COMPARISON - NUMBERS
For each of the statements below, circle the rating of your choice.

1. Overall, I am satisfied with the ease of comparing numbers.

STRONGLY AGREE STRONGLY DISAGREE

1 2 3 4 5 6 7

2. Overall, I am satisfied with the amount of time it took to complete comparing numbers.

STRONGLY AGREE STRONGLY DISAGREE

1 2 3 4 5 6 7

3. Overall, I can effectively carry out tasks using a system based on comparing numbers.

STRONGLY AGREE STRONGLY DISAGREE

1 2 3 4 5 6 7

E.2 After experiment questionnaire

Thank you for participating in the study. Could you please answer the following questions.

[1]Developed by Lewis [68]

1. Please rank each of the following methods

Comparing numbers

VERY EASY						VERY DIFFICULT
1	2	3	4	5	6	7

Typing a number

VERY EASY						VERY DIFFICULT
1	2	3	4	5	6	7

Word-matching and number typing

VERY EASY						VERY DIFFICULT
1	2	3	4	5	6	7

Double comparison

VERY EASY						VERY DIFFICULT
1	2	3	4	5	6	7

Image comparison

VERY EASY						VERY DIFFICULT
1	2	3	4	5	6	7

2. Which methods do you think are difficult to use? Please tick as many as applicable.
☐ Comparing numbers ☐ Typing a number ☐ Word-matching and number typing ☐ Double comparison ☐ Image comparison

3. Which method do you think is the most difficult? Please tick one.
☐ Comparing numbers ☐ Typing a number ☐ Word-matching and number typing ☐ Double comparison ☐ Image comparison

4. Which method would you avoid? Please tick one.
☐ Comparing numbers ☐ Typing a number ☐ Word-matching and number typing ☐ Double comparison ☐ Image comparison

5. Which methods do you think are to use? Please tick as many as applicable.
☐ Comparing numbers ☐ Typing a number ☐ Word-matching and number typing ☐ Double comparison ☐ Image comparison

6. Which method do you think is the easiest? Please tick one.
☐ Comparing numbers ☐ Typing a number ☐ Word-matching and number typing ☐ Double comparison ☐ Image comparison

7. Which method is your personal preference? Please tick one.
☐ Comparing numbers ☐ Typing a number ☐ Word-matching and number typing ☐ Double comparison ☐ Image comparison

Thank you for your participation.

E.3 Enrolment questionnaire

Age (Please tick ✓)

[ ] 16 – 25 [ ] 26 – 35 [ ] 36 – 45
[ ] 46 – 55 [ ] 56 – 65 [ ] 66 – 75 [ ] 76 – 85

Highest level of education attained (Please tick ✓)
[ ] High school [ ] College [ ] Graduate [ ] Post-graduate

Education major e.g. Chemistry, Business studies
.....

Gender (Please tick ✓)
[ ] Male [ ] Female

Do you have any personal mobile device such as cell phone, PDA, pocket pc, smart phone?
[ ] YES [ ] NO

TYPE(if applicable e.g Nokia N73).....

For how long have you been using mobile phones?.....

On a typical day, how many text messages do you send on your mobile device?
[ ] Less than 5 [ ] 5 to 10 [ ] 10 to 20 [ ] 20 or more

Is your mobile device capable of establishing Bluetooth, Infra-red or WI-FI connection?
[ ] YES [ ]NO [ ]N/A

Do you use any of its Bluetooth, Infra-red or WI-FI functionality on a regular basis?
[ ]YES (how often?)..... [ ]NO [ ]N/A

Please check the corresponding box if you answered YES to question 9:
[ ] Playing two-player mobile phone games

[ ] Using a wireless headset with your mobile phone
[ ] Connecting your computer or PDA to the internet using your mobile phone
[ ] Wirelessly synchronising your mobile phone calendar with your computer calendar
[ ] Other (specify) .....

In general, I am concerned about security while using wireless communication
[ ] Agree [ ] Disagree