Investigation of the Security and Usability of Visual and Verbal Public Key Fingerprint Verification Methods

Lee William Livsey

Doctor of Philosophy
University of York
Computer Science

April 2023

Abstract

Modern end-to-end-encrypted (E2EE) applications include an optional key fingerprint verification which allows users to establish the authenticity of a received key, and provide assurance that all subsequent communication is confidential. An under-explored aspect is the impact of verification mode upon user performance and perceived usability. Key fingerprints can be verified either visually or verbally, which present very different tasks to the user. Modern applications tend to support verification using a verbal verification, yet previous research has largely investigated visual verification. Users may also possess a pre-existing preference in how they prefer to process auditory–visual information, which may in turn affect their performance.

This thesis reports the results of a systematic investigation of the impact of verification mode upon user performance and perceived usability, with the evidence suggesting that visual verification is more efficient and provides increased usability. A robust usability effect was observed, with participants found to make more non-attack errors when using both word-based and numerical fingerprints. A surprising result was the absence of a security effect related to effectiveness, with participants found to be proficient in identifying non-identical attack fingerprints. The impact of a participant's auditory–visual information processing preference was also not significant, with the impact of verification mode instead appearing to be the dominant factor.

These results demonstrate the advantages in providing users the option to verify fingerprints visually. Visual verification appears to provide reduced ambiguity about the correctness of a received fingerprint, and though information processing preference was not found to be an indicator of performance, participants did report a clear preference for use of a visual verification mode. This should motivate E2EE applications to increase their support for utilisation of a visual verification, for those users who prefer to use it.

Contents

List of Figures

List of Tables

Acknowledgements

As with any project of this complexity, there are a number of people who have been instrumental in the development and completion of this thesis.

- I am extremely grateful to my supervisors Dr. Siamak F. Shahandashti and Prof. Helen Petrie for their support, guidance and mentorship throughout my time at the University of York. Their assistance in development of both this research and my skill as a researcher were vital.
- I would also like to thank Dr. Vasileios Vasilakis who served as a member of my thesis advisory panel and internal examiner, and provided an insightful external prospective on the progress of my research.
- A special thanks to all participants who contributed to this research. Though they will remain anonymous, their contribution was invaluable.
- On a personal note, I would like to thank my partner Leanne for all of her support and motivation which helped me complete this work. She has sacrificed a large of amount of her own time to enable me to focus on writing this thesis, and I could not have done it without her. In addition I also wish to thank the rest of our families who have helped in various ways, including by providing childcare or a quiet place to work.

Declarations

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references. Part of this thesis has been used in the following publication:

- Livsey, L., Petrie, H., Shahandashti, S. F., & Fray, A. (2021). Performance and Usability of Visual and Verbal Verification of Word-based Key Fingerprints. In Human Aspects of Information Security and Assurance: 15th IFIP WG 11.12 International Symposium, HAISA 2021, Virtual Event, July 7–9, 2021, Proceedings, pp. 199–210. Springer International Publishing.

This research was funded via an EPSRC scholarship and was conducted entirely on my own.

Chapter 1

Introduction

1.1 Problem Background

A recent trend in instant messaging applications has been to adopt end-to-end encryption (E2EE), to provide universal access to secure communication and detect whether an active attacker is eavesdropping on a users conversations [14, 66]. Additional examples of activities that have increased the demand for inclusion of E2EE include disclosure of the mass surveillance activities of a number of nation states in 2013, followed by the subsequent acknowledgement of the large scale harvesting and misuse of users personal data by Cambridge Analytica in 2018 [5, 52]. At the time of writing, the most popular E2EE instant messaging application is WhatsApp which has been downloaded over five billion times from the Google Play Store, and in 2020 claimed to posses over two billion users [24, 41]. Similar applications include Viber (1.2 billion users) and Signal (40 million users) [13, 57].

E2EE utilises public key encryption to protect message confidentiality. Users are assured that all messages are encrypted whist in transit with only the intended recipient able to decrypt and read them. Its inclusion is increasingly non-negotiable, and the user base may desert a messaging application if it is not securely implemented. As an example, in early 2020 a variety of organisations banned the use of the Zoom application after reports emerged that both messages and public keys were exchanged via severs located in China, potentially breaking the guarantee of E2EE and introducing the potential for interference by the Chinese intelligence services [6, 36, 43, 64].

A vital preliminary task to achieve E2EE is an authenticated key exchange between a pair of intended recipients, which then enables users to encrypt and sign messages for each other. A lack of authentication during the key exchange introduces the potential for a determined attacker to to implement a man-in-the-middle (MitM) attack. In such attacks an attacker intercepts the initial exchange of public keys and replaces them with keys that they themselves control. If the intended users are unable to identify this manipulation, and accept the keys received from the attacker as genuine, then the attacker can gain full control of the message exchange and eavesdrop on all messages without breaking the underlying encryption. Thus, the communication would appear to remain secure to individual users.

An attacker who seeks to implement such an attack against a real instant messaging application would need to gain access to a host positioned between the targeted user's

device and the application's back end infrastructure to facilitate interception and manipulation of packets at the network level. There a variety of potential methods that could accomplish this, including persuading a target to install a custom malicious application upon their device, exploiting a separate vulnerability to gain access to a device located within the same subnet as the target device, collusion with the target's internet service provider or identifying and exploiting an issue within the messaging application's own back end infrastructure.

Though such attacks are theoretically possible, they are difficult to implement in practise and will require a well resourced and highly motivated attacker, for example a state level actor. Consequently, it is difficult to ascertain from public sources if the specific attack envisaged within this thesis has occurred in practice as any such reports are likely to remain classified. However, there is publicly disclosed evidence of nation states attempting to obtain access to the mobile devices of high value targets, for example Saudi Arabia's usage of NSO's Pegasus spyware against Jamal Khashoggi prior to his murder in October 2018 [29]. These factors suggest that attacks of this type may be of real concern for members of society who are perceived to be the target of state level actors, for example investigative journalists or protesters against authoritarian states.

Prevention of MitM attacks is typically achieved by obtaining a public key certificate from a designated Certificate Authority (CA), which acts as a trusted third party (TTP) and attests to the security and validity of an specified public key. This model is commonplace in a variety of different areas of communication, including internet traffic encrypted using TLS and email messages encrypted using the S/MIME protocol [50, 54]. The main advantage is that it enables authentication to be automated without the need for any direct user interaction, but the trade-off is that users must place their trust into the honesty and integrity of the certification authority, which cannot always be guaranteed [60].

Furthermore, provisioning of a CA introduces significant overheads, both in terms of cost and maintenance. Instead, instant messaging applications typically operate within a decentralised model and incorporate a peer-to-peer human interaction within the initial key exchange. Current solutions begin with the exchange of a key-dependent verification message via an out-of-band channel (OOB), which assures the integrity of "short" messages [27]. Note that the OOB channel does not need to provide any guarantee of confidentiality, and so the attacker may observe all messages but crucially is unable to manipulate them. Clearly an exchange of messages within the application cannot be classed as OOB since the attacker is assumed to be able to intercept and manipulate these messages, but there exists a number of potential options.

If users can meet in person, they may create an OOB channel directly between their devices and then automatically verify the authenticity of each other's public key material (e.g. through NFC or scanning a QR code). This solves the problem for the in-person context, yet such applications are mainly intended for remote communication as it is not always feasible for users to meet in person.

In the remote setting, the OOB channel cannot be directly implemented between devices as communication is required to traverse a public channel (e.g the internet or SMS). The solution is to directly involve users in determining the authenticity of an exchanged key via a manual inspection. Practical examples of user-facilitated OOB channels include phone calls, text messages or publishing the fingerprint on a trusted website, e.g. The

Representation	Example public key fingerprint
Hexadecimal	83D1 3AFF D71F 8918 5F37
Trustwords	ORIN EFFEMINATE ASHER PEEING INTENTION
Numerical	33745 15103 55071 35096 24375
PGP Words	Mohawk scavenger cleanup Yucata stopwatch, businessman nightbird borderline eyetooth consensus

Table 1.1: Examples of a variety of different public key fingerprint representations generated from the first 80 bits of the same underlying hash value.

Guardian's journalist PGP key fingerprints page[1]. In practise secure public keys are at least 2000 bits long and impractical for inclusion in a verification by users. Instead, a cryptographic hash function is used to generate a short key dependent value called a public key fingerprint, which is of the order hundreds of bits in length. These fingerprint bit strings are encoded into a human readable format and users are asked to verify that the fingerprint displayed upon their device matches that received from some verifiable and trusted source. If it does, then the user is assured of the confidentiality of their communications, so long as the cryptographic hash function used to generate the fingerprint remains collision resistant[2]. If the fingerprint values differ then this provides an indication that the received public key may have been manipulated in transit and that they may be the target of a MitM attack.

Fingerprint values are usually encoded into easy-to-use formats such as chunked numbers (e.g. in Signal/WhatsApp), or dictionary words (e.g. in Pretty Easy Privacy (PEP)[3]) for Pretty Good Privacy (PGP) keys. We provide some examples of public key fingerprint representations included in real-world secure messaging applications in Table 1.1.

Traditionally fingerprints were often exchanged in person or printed upon a user's business card, which formed an asynchronous visual verification. However, modern secure messaging applications have increasingly encouraged users to exchange the fingerprint string via a synchronous phone call, which introduces a verbal verification mode. The shift towards use of a verbal verification appears to provide clear benefits, particularly that the synchronous nature of the exchange allows the users to participate in a negotiation as to whether they both believe the fingerprints to match or not.

Regardless of the chosen verification mode, manual verification of the received fingerprint is a challenging task that includes significant potential for human error and introduces an interesting new threat vector. Attackers could adapt their attack strategy, and attempt to generate a public key whose fingerprint is highly similar to that of a target. Such near pre-image fingerprints would be difficult to distinguish, and there exists a real possibility that users may accept the received fingerprint as genuine; the user may make a mistake, rush the verification or only compare a section of the two fingerprints. A real-world example of a near pre-image fingerprint encoded using the Trustwords representation of the PEP secure email application is provided in Table 1.2[4]. The underlying

[1] `www.theguardian.com/pgp`

[2] Collision resistance ensures that different keys are mapped to different fingerprints, i.e. identical fingerprints guarantee identical original keys.

[3] `www.pep.security`

[4] This pair forms one of the potential attacks of the study described in Chapter 3.

Original fingerprint	ALIKEE FLATTERER CURD ROAN CALCOMP
Attack fingerprint	ALIKEE FLATTERER COMPLECTED FEELINGLY CALCOMP

Table 1.2: Example of a near pre-image displayed using the Trustwords representation of PEP.

public keys of the two fingerprints are available in publicly accessible PGP servers. The first, second and fifth words are in agreement, and the differences in the central section may be difficult for users to identify during a casual verification.

Previous work has identified significant differences in user performance and perceived usability between different representations, including investigation of the ability to identify near pre-image fingerprints [16,58]. But there are a range of additional factors which may affect both performance and perceived usability, including:

- The verification mode.
- The length of fingerprint.

The impact of the fingerprint length appears clear, with shorter fingerprints easier to compare than longer ones [61]. However, the impact of verification mode represents an as yet unexplored aspect within the literature.

1.2 Research Aims and Objectives

The research presented in this thesis aims to investigate the impact of verification mode upon the verification of key fingerprints, and aims to answer the following main questions:

Is there a significant difference in user performance and perceived usability between visual and verbal key fingerprint verifications?

Previous work has identified challenges related to the impact of different representations upon the usable security of visual verifications of key fingerprints, but there has been limited investigation of the alternative synchronous verbal verifications. Verbal verifications represent a very different task with different types of complexity, for example verifications which include homophones would be much more difficult to verify during a verbal verification. It is reasonable to hypothesise that users may also find verbal verifications to be a challenging task, and that one of the two modes may possess a clear increase in both security and usability.

What is the impact of a user's auditory–visual information processing preference upon a key fingerprint verification?

It is widely agreed that users possess a personal preference for how the receive information, with an auditory–visual preference forming a common dimension. A useful example is how some people prefer to read a book, whilst others prefer to listen to an audiobook. However, the impact of such a preference upon subsequent task performance is a controversial topic, particularly in the context of how users learn [69].

Yet the task under investigation within this research requires the use of working memory instead of learning. The impact of a user's information processing preference upon working memory tasks is an open question, particularly in the context of a key fingerprint verification task. It is of interest to determine whether improved performance and usability is observed when the verification mode is aligned to the users information processing preference. It is feasible that users who display a strong auditory preference for receiving information may display improved performance using a verbal verification mode, and vice versa.

The investigation of the two main research questions generated additional questions, the results of which helped to improve the research methodology and produce a cohesive narrative. Specifically these were:

How can a user's auditory–visual information processing preference be measured?

Though an auditory–visual information processing preference is widely accepted, there is no agreed method to measure this preference. Previous instruments have tended to focus upon assessment of a participants learning style and included a range of aspects which distract from the phenomena of interest (e.g. environmental factors). Development of such a measure will aid the methodology of this research, as well as assisting future researchers who may seek to investigate this specific scenario in other fields.

What is the optimal level of fingerprint similarity achievable from a computational search performed by a well resourced and highly motivated attacker?

Previous work had investigated the threat of MitM attacks within a range of different representations, but these simulations did not always include practical levels of similarity. Hence, an important part of this research is to consider the number of errors users make when they encounter attacks which possess levels of similarity that are likely achievable from a large brute force effort. This is first investigated using fingerprints that are accessible within public key servers, but these were found to be ineffective. Hence, a brute force pre-computation is simulated to assess the achievable attack similarity, with the results providing clarity and also provide a template for subsequent work.

What is the optimal method to assess the similarity of two key fingerprints?

At first glance, the concept of "similarity" of two fingerprints appears to be straightforward – similar fingerprints include a greater proportion of identical characters than non-similar ones. A useful related measure is the "edit distance", i.e. the minimum number of edits required to transform one fingerprint to the other. Though it is in the application of the term "distance" where small nuances arise, particularly in the context of word-based fingerprints. There are two distinct distances that can be assigned to a pair of words, an orthographical distance and a phonological distance, and it is unclear whether the type of fingerprint similarity impacts an attacker's success rate. Determination of the impact of this difference would help raise awareness of an attacker's optimal

attack strategy, and provide an interesting insight into the types of differences that users find difficult to identify.

The answers to these questions will provide an important contribution that improves the understanding of the manual key fingerprint verification task, both in terms of the properties of the verification modes and the impact of individual differences of users. Though use of a verbal verification is encouraged by applications, their lacks sufficient evidence within the academic community to determine if this is the most appropriate method for users, in terms of both the provided security and usability of the task. The answers to these research questions will provide a clearer picture about the role of verification mode upon the fingerprint verification task, and help to determine if a lack of encouragement of visual verification methods is a sensible design decision. This may motivate instant messaging applications to provide improved support and integration for both visual and verbal verification modes.

1.3 Research Contributions

The main contributions of this research include:

- Identification of a robust usability effect between the two verification modes. In all studies of this research, participants were observed to make significantly more non-attack errors when they used a verbal verification mode, including during tasks that implemented different fingerprint representations.
- Development of a custom auditory–visual information processing scale (IPP-AV). A detailed search of the literature found that such a scale was previously absent, and its development may aid future research within this area.
- The research found that a participants auditory–visual information processing preference did not play a significant role in their performance. There was a lack of correlation between a participants IPP-AV score and their performance on the verification task, with the impact of verification mode appearing to be the dominant factor.
- Description of an attack simulation which seeks to model the fingerprint similarity potentially achievable from a simulated brute force pre-computation by a well resourced and highly motivated attacker, such as a state level actor. This ultimately identified attacks with significantly greater similarity than those extracted from public key servers.

The direct beneficiaries of these findings are the developers of instant messaging applications. The results provide guidance upon the design a fingerprint verification task, and suggest that both visual and verbal verification modes should be supported. Members of the academic community can also benefit. If developers incorporate this guidance within their future designs, then these results may also be of benefit to the general user base as they will be provided with the option to utilise a verification mode with improved usability.

Development of the IPP-AV scale to enable measurement of a user's information processing preference may be of benefit to future researchers who seek to explore this influence upon other tasks in scenarios where auditory and visual verifications are pertinent. Such work would also help to establish the validity and reliability of the scale.

The lack of effect related to a user's information processing preference is also of interest to the academic community. It was reasonable to suggest that there may be some kind of effect, but the findings suggest that users perform as well using either verification mode regardless of their personal information processing preference. This appears to share parallels with the results of investigation of the meshing hypothesis within the education field.

1.4 Thesis Overview

This thesis consists of seven chapters, with the remaining chapters organised as follows:

- Chapter 2 considers previous related work including an overview of usability issues related to secure messaging applications, a review of previous research related to key fingerprint verification tasks and a summary of previous human factors studies that investigated secure device pairing methods.

- Chapter 3 describes an exploratory study that aimed to provide an initial assessment of the performance and perceived usability of a word-based key fingerprint verification. The study identified an interesting usability effect related to the effectiveness of the verifications, but included a number of design limitations that prevented a full assessment of the intended research questions.

- Chapter 4 details the development of an improved experimental methodology that intended to resolve the design limitations of the exploratory study. The chapter describes the development of the custom IPP-AV scale, a simulated brute force fingerprint generation which identifies attack pairs which possess increased similarity, and an assessment of the impact of different similarity measures upon user performance.

- Chapter 5 applies the improved experimental methodology within a further investigation of the usable security of word-based key fingerprints. The improved methodology enabled a coverage of all research questions, and identified both a security and usability effect related to the effectiveness of the two modes. However, participants scores on the IPP-AV scale did not correlate with their performance, indicating that the impact of a user's information processing preference is not a significant factor.

- Chapter 6 reports the results of a subsequent study that applied the same experimental methodology to the investigation of numerical fingerprints. The study also identified a usability effect between verification mode and effectiveness, indicating that this is a robust property of a fingerprint verification task. However, in contrast to the previous chapter, there was no related security effect indicating that this difference may instead be a property of the chosen representation. The relationship between a participant's information processing preference and performance was again not significant.

- Chapter 7 provides a summary and discussion of the results of this research, potential avenues for future work and an overview of additional factors which affect a fingerprint verification in addition to verification mode. The chapter closes with some final conclusions and provides a high-level overview of potential practical applications of this research.

Chapter 2

Background and Context

2.1 Usability Issues in Secure Messaging Applications

Usability issues related to secure messaging applications were highlighted by Whitten and Tygar, who found that design limitations of the application's user interface caused non-specialist users to be unable to successfully use PGP 5.0 to send secure emails, particularly those who lacked a working knowledge of public key cryptography. They reached the conclusion that "the standard model of user interface design is not sufficient to make computer security usable for people who are not already knowledgeable in that area" [67].

Recent work has identified usability issues specific to the authentication procedures of modern secure messaging applications. Schröder et al. investigated the usability of Signal's key fingerprint verification ceremony. The study found that from a sample of 28 computer science students, 21 failed to "compare encryption keys to verify the identity of other users" even though they "believed they had succeeded" [55]. Often participants confused verification of the identity of their intended recipient with the required task of verification of the received key material. Related work identified similar issues within WhatsApp, Viber and Telegram, finding that participants were both unaware of the need to verify their recipient's key and unable to do so without additional instruction [26, 63].

2.2 The Effect of Fingerprint Representation

Previous work has included considerable investigation of the effect of encoding upon key fingerprint verification tasks, which identified significant differences in performance and perceived usability (see Table 2.1). These works have lead to a number of recommendations regarding the fingerprint representations that provide increased security and usability, which have been reflected in the design of the latest versions of the task. The clearest example of this is that developers have largely dropped usage of the hexadecimal representation in favour of either a word-based or numerical representation.

	Tan et al. [58] (a)	**Dechand et al.** [16]	**Kainda et al.** [27]	**Shirvanian et al.** [56]
Hexadecimal	21%	11% (a) 0% (n)	-	-
Sentences	6%	3% (a) 2% (n)	10% (a) 17% (n)	-
Words	14%	9% (a) 0% (n)	0% (a) 17% (n)	-
Numbers	35%	6% (a) 0% (n)	0% (a) 3% (n)	23% (n) 40% (a)
Images	10% -SSH 12% - Unicorns 54% - Vash	-	3% (n)	13% (n) 19% (a)
Auditory	-	-	3% (a) 0% (n)	27% (a) 0% (n)

Table 2.1: Error rates observed in previous studies of key fingerprint verification. Separate non-attack and attack error rates are indicated with (n) and (a), respectively.

An Empirical Study of Textual Key-Fingerprint Representations

Dechand et al. [16] performed a detailed investigation of textual fingerprint representations, including hexadecimal, numerical, Base32, PGP, Peerio, and randomly-generated English sentences (see Table 2.2). The study tasked participants to perform 46 fingerprint verifications, although 6 were discounted from the analysis as they acted as either attention checks or training tasks. The remaining 40 tasks were split across four different representations (i.e 10 verifications per assigned representation). Each participant encountered four near pre-image attacks, one per representation. The study was completed within a web browser, and included functionality which simulated a scenario that required a participant to verify that a fingerprint displayed in their secure messaging application matched that displayed on the business card of their hypothetical intended recipient (see Figure 2.1).

The results identified that word-based formats led to higher usability scores and increased attack detection rates than the traditional hexadecimal format. A sentence-based representation performed best empirically with error rates of less than 3%, but these were also perceived to be less trustworthy than other word-based representations (see Table 2.3). One explanation may be that the sentence's lack of semantic meaning made their inclusion confusing, and not something that would be expected from a security check. The study included a hypothetical attacker able to control 80 bits of a 112-bit fingerprint, with all differences confined to central fingerprint chunks. Such fingerprints are likely infeasible in practise, which may have impacted the study's ecological validity.

Scheme	Example
Hexadecimal	18e2 55fd b51b c808 601b ee5c 2d69
Base32	ddrf l7nv dpea qya3 5zoc 22i
Numeric	2016 507 6420 1070 394 1136 2973 991 70
PGP	locale voyager waffle disable Belfast performance slingshot Ohio spearhead coherence hamlet liberty reform hamburger
Peerio	bates talking duke rummy slurps iced farce pound day
Sentences	Your line works for this kind power cruelly. That lazy snow agrees upon our tall offer.

Table 2.2: Examples of the textual fingerprint representations investigated by Dechand et. al. This table was presented as Table 1 within [16].

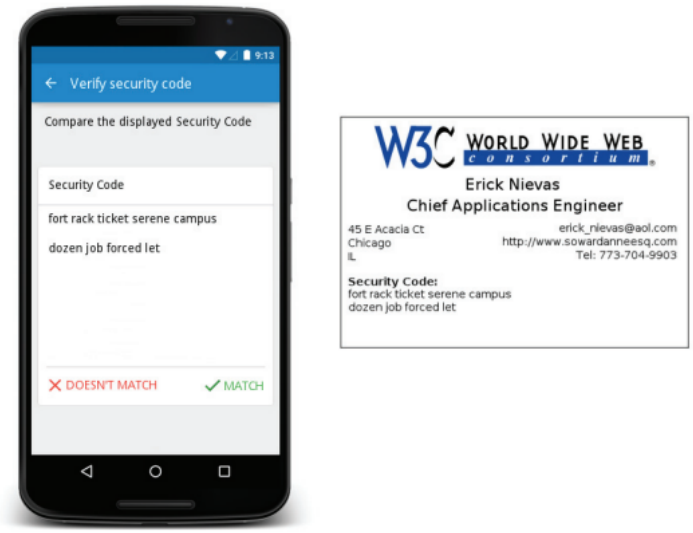


Figure 2.1: An example of the task that Dechand et. al asked participants to complete. This figure appeared as Figure 3 in [16].

	Trustworthiness	**Usability**
Hexadecimal	82%	63%
Sentences	67%	85%
Words	75%	80%
Numbers	76%	76%

Table 2.3: Trustworthiness and usability scores obtained by Dechand et al. [16]

Scheme	Example
Hexadecimal	BAAA 9AE6 7B8B 0D41 BD83 05E7 5209 8EDF 1058 41F6
Alt vow./cons.	bunu difu tura wefi wiwe haqe tano haco qevu cori qife nufi
Words	learning equal education bent collar religion new shelf angle table train sad keep meal thing punishment
Numbers	7748 5689 7453 6977 5604 5939 2765 8791 5022 4957 3805 0309
Sentences	The basket ends your right cat on his linen. Her range repeats her nerve. The smile tells secretly. My clean cake pulls your waiting pocket.

Table 2.4: Examples of the textual fingerprint representations investigated by Tan et. al. This table was presented as Table 2 within [58].

Can Unicorns Help Users Compare Crypto Key Fingerprints?

Tan et al. [58] investigated textual fingerprint formats alongside a range of visual representations. Like Dechand et al., the study investigated a variety of textual representations including hexadecimal, words, numerical strings and sentences. Tan et al. also investigated a pseudo-word representation generated by strings of alternation vowels and strings (see Table 2.4). The study also investigated the potential usage of image based fingerprints as an alternative to the standard textual approach (see Figure 2.2). The hypothesis was that the image structure may facilitate a more usable solution if participants were able to easily identify differences between the two displayed images.

The study asked participants to perform 30 fingerprint verifications. Of these 28 were identical non-attack verifications, one was a fully mismatching attention check and

(a) OpenSSH Visual Host Key (b) Vash (c) Unicorn

Figure 2.2: Examples of the textual fingerprint representations investigated by Tan et. al. This table was presented as Figure 2 within [58].

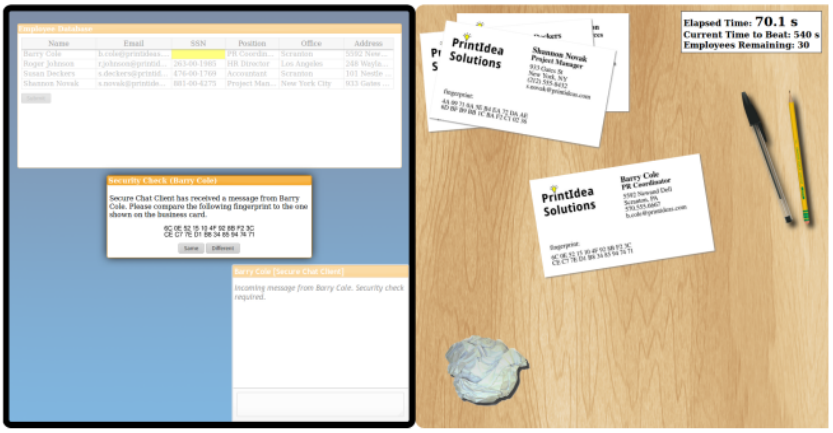


Figure 2.3: An example of the task that Tan et. al asked participants to complete. This figure appeared as Figure 1 in [58].

the other acted as a near pre-image attack randomly positioned between the 25^{th} and 29^{th} tasks. Participant's performed the study from within the web browser, and the scenario again required verification of a fingerprint displayed within the secure messaging application with one printed on a business card (see Figure 2.3).

The results found that text-based formats achieved some of the lowest error rates with the sentence based approach again found to generate the least number of errors, with only 6% of participants failing to identify the near pre-image attack. The performance of the image–based formats varied with 54% of participants that made verifications using the unicorn representation failing to identify the near pre-image attack. The paper also considered other challenges caused by the use of an image–based representation. Chiefly that there will always remain a degree of uncertainty when making an image–based fingerprint verification as it is infeasible for a user to compare individual pixels of two images, a characteristic that is not shared by the textual representations. Given the hypothesised target user of this thesis, a user that is aware of the threat of a MitM attack and is deemed to be of sufficient value to be targeted by a well resourced and highly motivated attacker, it is unlikely that they will be willing to accept use of a representation that is unable to remove all such uncertainty.

A strength of the study was the inclusion of an entropy-based attack model, which simulated an attack set generated from a large brute force computation. The majority of conditions simulated a 2^{60} computation, which enabled an attacker to control 60 bits of the underlying 160-bit SHA-1 hash value. However, the study included only a single attack towards the end of the experiment, and it is possible that participants had by this point become fatigued and made more errors within the attack verifications than would be observed among real users. Furthermore, an artificial time pressure encouraged users to rush their verifications, which may have artificially increased the error rate. These factors may help explain observation of a surprisingly large numerical attack error rate.

Both Dechand et al. and Tan et al. simulated only visual verifications, with the received fingerprint displayed on a business card (see Figures 2.1 and 2.3), a method commonly used as part of the Web of Trust [62]. Though verbal verifications were mentioned by Dechand

et al. neither study investigated differences between auditory and visual verifications.

Moreover, both studies included a wide range of conditions within their independent variable and performed extensive statistical tests. This reduces the strength of any identified results as the experiments fails to form a severe test of a hypothesised real world phenomena and instead form only an exploration. This is something that the research described within this thesis strove to avoid, by generating concise research questions that could be analysed through application of a limited number of statistical tests.

Usability and Security of Out-Of-Band Channels in Secure Device Pairing Protocols

Kainda et al. [27] performed a laboratory study which in addition to investigation of textual representations (e.g numbers, words and sentences) also included other methods, including auditory verification of numeric and alphanumeric strings. The study used two real Nokia mobile devices to simulate usage on a peer-to-peer payment mechanism, with a custom application developed to meet the requirements of this specific task. At the start of the study, each participant was provided with two devices. One was identified as their personal device an the other the device of the payee. The goal of the task was to use the devices to successfully perform secure payment transaction within the application, which required that participants verified that the fingerprints of both devices were identical. However, this task lacked external validity as participants only acted as individuals and were assumed to be in possession of both their own device and that of the payee.

The study included nine conditions within its independent variable, with each condition including three distinct tasks: an attack, a non-attack, and an attention check. This methodology is unlikely to produce clear evidence of a real effect as the small number of verifications per condition provide insufficient opportunity for observation of user variation within each task.

Overall, numerical verifications were found lead to the lowest attack and non-attack error rates, with similar results observed in both visual and auditory verifications. However, each verification task included fingerprints that were only 20 bits long, which is insufficient to provide security against MitM attacks in real applications. Consequently, it is not possible to be certain if an auditory verification provides equivalent effectiveness to a visual verification within real fingerprint verification tasks.

On the Pitfalls of End-to-End Encrypted Communications: A Study of Remote Key-Fingerprint Verification

Shirvanian et al. [56] investigated the performance and perceived usability of real messaging applications upon physical mobile devices, specifically WhatsApp, Viber and Telegram. This was a strength of the study, as their results are reflective of real tasks that users are likely to encounter in practise.

The study was implemented within a laboratory setting, with users simulating both remote and in-person fingerprint verifications, with a member of the study team acting as the intended recipient. As the recipient had pre-existing knowledge of the task, they were stipulated to act as a passive observer who read, transferred or showed the fingerprint to the participant. The ability to perform the verification with a real pair of users facilitates a more accurate investigation of the phenomena of interest, but this would be further improved if the pairs consisted of two participants who are both allowed to perform an active role within the verification task.

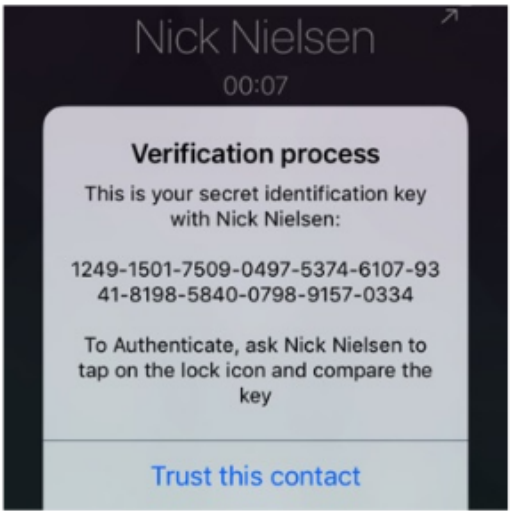
The remote verifications, which are of direct interest to this research, included three conditions. The first used the Viber application to perform an auditory verification of a 48 digit numerical string. The second used the Telegram application to perform verification of an image-based fingerprint that was shared via a text message. The final condition used the WhatsApp application to perform verification of a 60 digit numerical string shared via text message. Examples of these tasks is provided in Figure 2.4.

Within each condition participants were asked to perform five verifications. Two consisted of identical non-attack verifications and the remaining three acted as attack verifications: a fully mismatching fingerprint, an attack where all but one block was identical, and an attack where all but a single character was identical. However, the final two attacks display levels of similarity that are likely to be infeasible for an attacker to produce in practise which impacts the external reliability of their results.

All three applications were perceived to provide insufficient security and usability, with participants found to make both attack and non-attack errors (see Table 2.1). Furthermore, these error rates were significantly higher when compared with the investigated in-person tasks, during which no users made non-security errors and the highest security error rate was 2.67%.

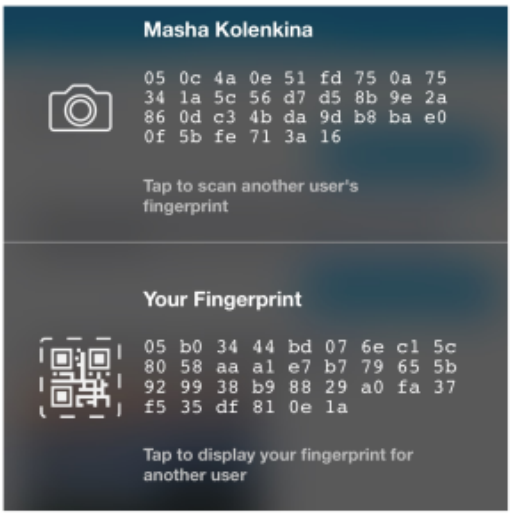
Though the study included discussion of the options to implement both visual fingerprint exchanges via a text message or email and verbal exchanges via a phone call, there was no direct investigation of the impact of the differences between these two verification strategies upon user performance and perceived usability. This is an gap that this research intended to address.

(a) WhatsApp Security Code.



(b) Viber Secret Identity Key.

(c) Telegram Encryption Code.



(d) Signal Fingerprint.

Figure 2.4: Examples of the fingerprint tasks investigated within the study of Shirvanian et al. This figure appeared as Figure 8 in [56]

2.3 Fingerprint Verification in Deployed Applications

This section provides an overview of the fingerprint verification interfaces found within real-world applications. There is significant variety in the design of these interfaces, with differences in both the chosen representation and the recommended method to perform the verification. Some have made modified their designs to increase their usability, in line with the findings of the research discussed in Section 2.2, yet surprisingly some have seen very little development and continue to implement tasks which users are known to find difficult, such as verification of Hexadecimal strings.

2.3.1 WhatsApp and Signal

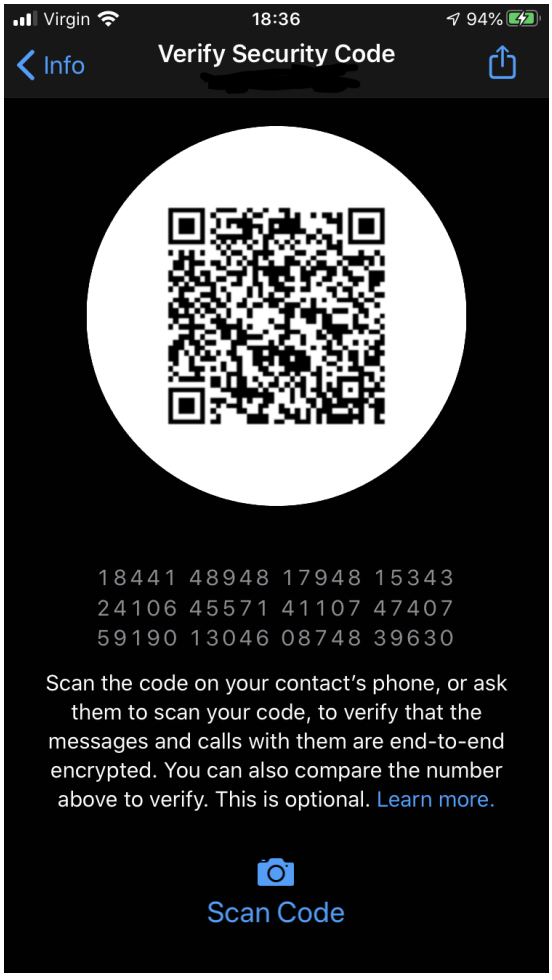


Figure 2.5: WhatsApp verification interface

Whatsapp and Signal implement end-to-end-encryption (E2EE) by default. Both applications utilise the same manual key verification functionality, with the interface supporting both in-person (QR-code) and remote verification via a 60 digit fingerprint (see Figure 2.5). The fingerprint consists of a concatenation of truncated hash value of each user's public key, with each user providing 30 digits each. Both applications recommend

that remote users share their fingerprints visually and includes integration with a device's text and email functionality to easily facilitate visual verifications. This is a significant strength, particularly when compared with the functionality of similar applications which do not always facilitate an embedded method for a visual exchange of fingerprints (see Section 2.3.2).

A limitation is that this interface is hidden away within the application and requires multiple clicks to access. Furthermore, notification that a recipient's key has changed are disabled by default, and so it remains unlikely that users will routinely complete these verifications unless they possess some specialist knowledge of its advantages.

Unique to this interface is inclusion of a tabular structure to display the fingerprint. This is an interesting deign choice as, in addition to improving the visual design of the interface, the inclusion of line breaks complicates the challenge for an attacker seeking to perform a near pre-image attack as more chunks occur at the ends of lines.

2.3.2 Viber

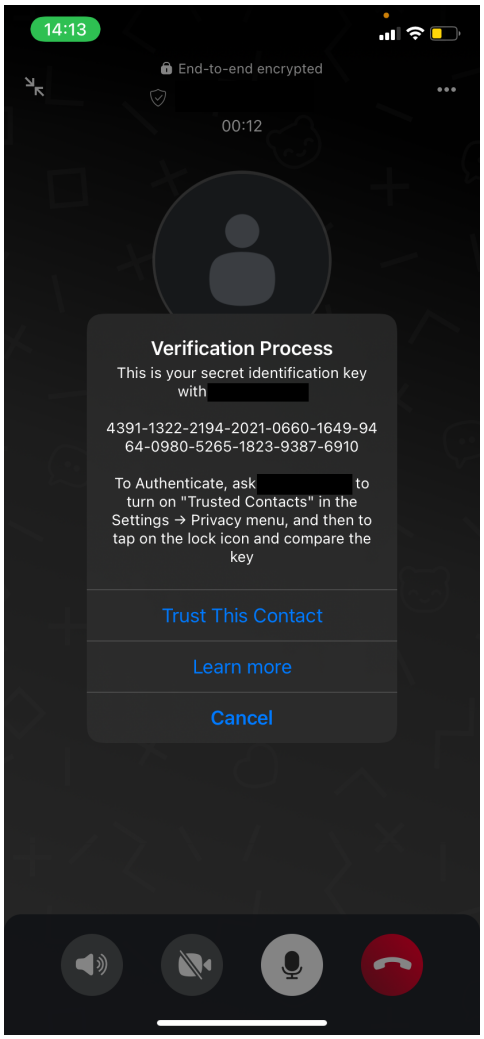


Figure 2.6: Viber verification interface

Viber also implements E2EE by default, and provides functionality to mark a recipient as a "trusted contact". This process involves completion of a special synchronous verification during a video call with an intended recipient. During the call it is possible to display a 48-digit fingerprint which users then read out and compare collaboratively (see Figure 2.6), and served as inspiration for investigation of differences between visual and auditory verification methods. Though the ability to form a collaborative conclusion is likely beneficial, the interface does not include any facility to easily share the fingerprint visually and the copy and paste functionality is disabled. This is a restriction that may cause challenges for some users, particularly if they find auditory tasks difficult.

2.3.3 Pretty Easy Privacy (PEP)

Pretty Easy Privacy[1] is an application which integrates with PGP to provide encrypted email messaging, seeking to automate many of the configuration tasks that users are known to find difficult, such as the initial key generation. Key verification remains a manual task, but in an effort to improve its usability they developed a custom word base which they called Trustwords, with each word corresponding to a 16-bit chunk of the hash value. In the standard configuration, the first 5 words of the fingerprint would then be displayed for comparison upon the interface, with the option to verify all 10 words if desired (see Figure 2.7). This was of particular interest to this research, as it served as a practical example of word-based fingerprint verification which had been seen to perform well within the literature.

One limitation of the PEP interface was its lack of clear support for visual verification of a fingerprint. This appears to be the case within its current interface (see Figure 2.8), but it was not possible to verify this directly as the application now utilised a licensing model.

[1]During the course of this research it appears PEP has been rebranded to Planck security (`https://www.planck.security/`). It now operates under a license based subscription model aimed at securing the email communication of corporate clients.

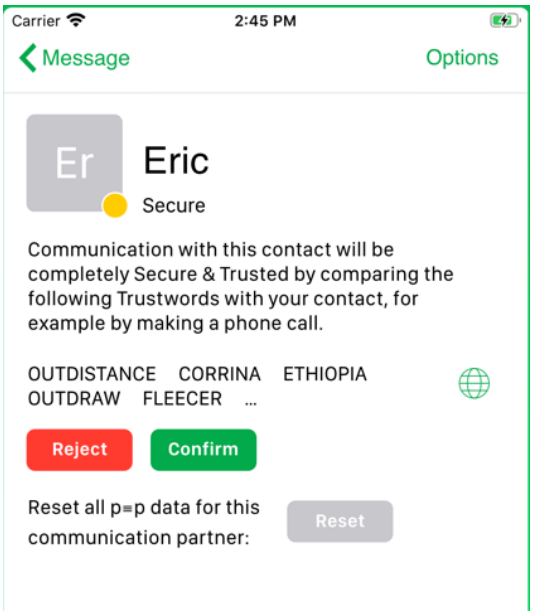


Figure 2.7: Original PEP interface.

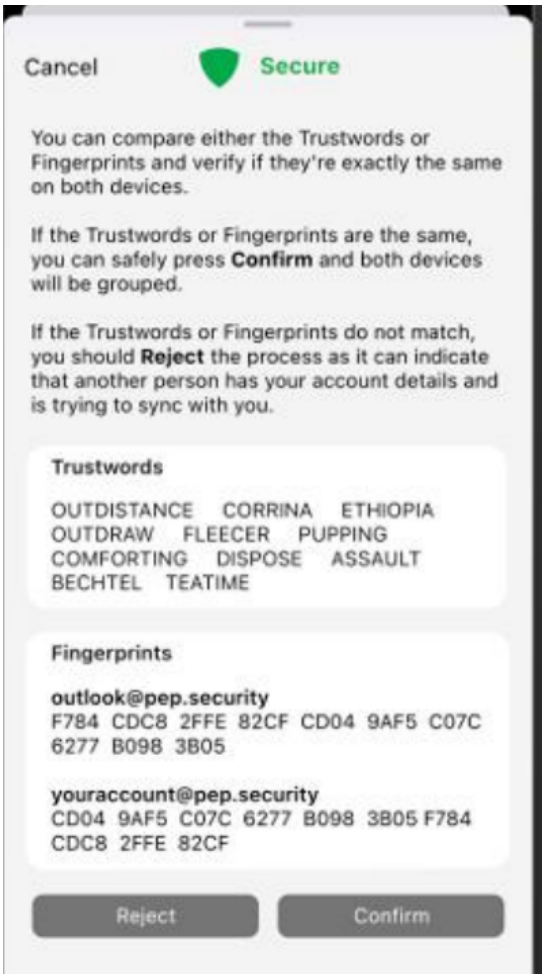


Figure 2.8: Current PEP interface, after rebranding to Planck Security.

2.4 Modality Effects in Secure Device Pairing Methods

Studies investigating a range of existing device pairing methods identified interesting differences in usability between visual and verbal verifications, but they involve substantially shorter fingerprints that provide sufficient security only for short-range device pairing scenarios [31, 39]. Goodrich et al. described a method to facilitate human assisted device pairing via an auditory channel, a method that they entitled "Loud and Clear". Previous secure device pairing methods had largely utilised the visual displays of each screen, but these can be limiting as not all devices include a visual display. Loud and Clear provides a method to extend device pairing to those that include only speaker functionality, with the sentence based audible fingerprint compared with that displayed on the screen of a second device [23]. However, though the study considered the potential for users to perform either a visual or verbal verification, they did not perform a related human factors study and so it was not possible to determine if users would display any differences between visual and verbal Loud and Clear verifications.

2.5 User Mental Models

A additional factor highlighted in a range studies was the impact of incomplete mental models of encryption and the types of attack that a fingerprint verification can protect against [1, 15, 26].

Abu-Salma et al. [1] implemented a qualitative study with 60 participants that aimed to develop understanding of user's "experience with different communication tools and their perceptions of the tools' security properties". The paper provided a thorough analysis of the motivations and mental models of users when using messaging applications, with the analysis of the sample's knowledge of encryption and the potential threat of MitM attacks of particular interest to this research. The first 10 interviews were unstructured which enabled thorough investigation of the factors which motivated the adoption of a secure messaging application and the perceived capabilities of a hypothetical attacker. These interviews identified a set of common themes which formed the areas of exploration in the remaining 50 semi-structured interviews.

They study "found that the adoption of secure communication tools is hindered by fragmented user bases and incompatible tools" and that "the vast majority of participants did not understand the essential concept of end-to-end encryption, limiting their motivation to adopt secure tools". A main factor for selection of an instant messaging application was its popularity with "the ability to reach their intended communication partners the primary communication goal of participants", rather than the relevant security guarantees provided by developers.

A common belief was that with sufficient effort any message could be decrypted and read by an attacker, particularly intelligence agencies and the application service providers. References were also made of acceptance of privacy notices causing users to feel that applications already have access to their data anyway, as evidenced by targeted adverts within social media applications. Consequently, "secure communications were perceived as futile" with the paper drawing the conclusion that "if the perception that

secure communications are futile persists, this will continue to hinder adoption". This is an understandable conclusion, as if users fail to trust the claims of secure messaging applications that they can keep their messages secure, then it is unlikely that users will be motivated to use them.

Interestingly a majority of participants "agreed that integrity is an important property a secure communication tool must offer", but "only three participants discussed man-in-the-middle attacks and digital signatures". This again highlights a blind spot in the mental models of typical users of secure messaging applications. Given that they appear to be largely unaware of the risks posed by a MitM attack against the initial exchange of public keys, then it is highly unlikely that they will be motivated to perform a key fingerprint verification.

Consequently, given the lack of knowledge or motivation among users to perform a key fingerprint verification, it appears likely that in the vast majority of cases an attacker would be able to successfully implement a MitM attack using an arbitrary public key whose fingerprint was fully mismatching when compared to the authentic one Furthermore, the previous research discussed in this chapter, and the clear design improvements that have been identified, are of limited impact unless users are educated about the limits of third parties to intercept and access messages sent within secure messaging applications, and also encouraged to protect themselves from the threat of MitM attacks.

Though these factors are deemed out of scope of this research, they form an important component that needs to be addressed to increase the rate at which users complete a fingerprint verification and improve their overall security posture. Abu-Salma et al. referenced a collection of efforts that have produced educational materials to explain existing security tools. However, as they note within their paper "documentation only helps the users who read it and are already motivated enough to adopt a new tool". The solution to this problem likely requires investment to include this specialist knowledge within the school curriculum, so as to engage with users at a very early age before these inaccurate and insecure mental models take root.

Chapter 3

Exploration of Word-Based Verification Modes and the Effect of Learning Style

3.1 Introduction

Previous investigation of the security and usability of public key fingerprints has tended to focus upon verifications made using only the visual verification mode. A common scenario was to facilitate the fingerprint verification through a simulated exchange of business cards with the fingerprint printed upon [16,58]. Prior to the work reported in this chapter, there had been no direct investigation of differences in user performance and perceived usability between visual and verbal verifications. Previous investigation of verifications facilitated using a verbal channel were restricted to short-range secure device pairing methods [23]. The results found a verbal exchange to perform well, but the fingerprints were much shorter in length and verifications did not always require direct user interaction. Thus, it was not clear if the same results would be replicated within the context of this research.

Modern instant messaging applications tend to encourage use of a verbal exchange and verification via a direct call with their intended recipient (e.g. WhatsApp and Viber). This provides many obvious advantages; it is more efficient, enables identification of familiar contacts via voice recognition, and allows for a challenge and response repartee to enable users to agree upon an accurate exchange. However, verbal verifications represent a significantly different task for users, and there exists the possibility that some may actually find this method to be more difficult. Potential challenges include the large number of homophones within spoken language (e.g English and French), or if the intended recipient possesses a particularly strong accent which may be very difficult for the user to accurately verify. This challenge is likely further exacerbated for non-native speakers who may be more likely to encounter unfamiliar words which increases their general uncertainty.

A related question concerned the impact of a user's personal preference to receive information upon their performance. A range of theories suggest that user's posses an innate preference for receiving information either visually or verbally, and it is reasonable to hypothesize that such a preference may impact their performance, particularly if the verification mode was aligned to their preference. However, this was yet to be explored within the existing literature.

This chapter reports the results of an exploratory study that aimed to investigate the impact of verification mode and information processing preference upon both user performance and perceived usability when performing word-based key fingerprint verification tasks [1]. The results provided valuable initial insight, and appeared to describe a complex picture. It was not possible to definitively identify an optimal verification mode, with the more effective visual mode also perceived to be less usable in two out of six usability dimensions. Subsequent evaluation of the study design also identified a number of aspects for improvement (see Section 3.4).

3.2 Method

3.2.1 Design

The study involved a within-participants design with two conditions; the visual and verbal verification modes. The visual verification mode simulated verification of a fingerprint via a text message exchange, whilst the verbal verification mode used a pre-recorded audio clip to simulate a phone call (see Figures 3.1 and 3.2). A within-participants design was chosen to enable participants to interact with both verification modes, facilitating investigation of the impact of their information processing preference upon their performance.

The study asked participants to simulate an authentication task by performing a verification of a received five word fingerprint. The five words were selected from the Trustwords word list [37], a real word example used in the Pretty Easy Privacy (PEP) extension to PGP. Each participant compared 20 pairs of key fingerprints in each condition, with the order of taking conditions counterbalanced to minimise the impact of learning effects. Apart from an attention check that occurred in position 3, each verification could either be an attack pair which included differences in the third and fourth words, or a fully identical non-attack pair. This attack strategy was motivated by the results of a previous eye tracking study which found that participants who only looked at only the start and end of a fingerprint pair were more likely to make an attack error than those participants who looked at a range of sections across the full string [42]. This attack strategy was also implemented in the related work which investigated the usable security of different fingerprint representations [16, 58]. Further detail about the attack set implemented in this study is provided in Section 3.2.3.

[1] The results of this study have been previously reported at HAISA 2021 [35]

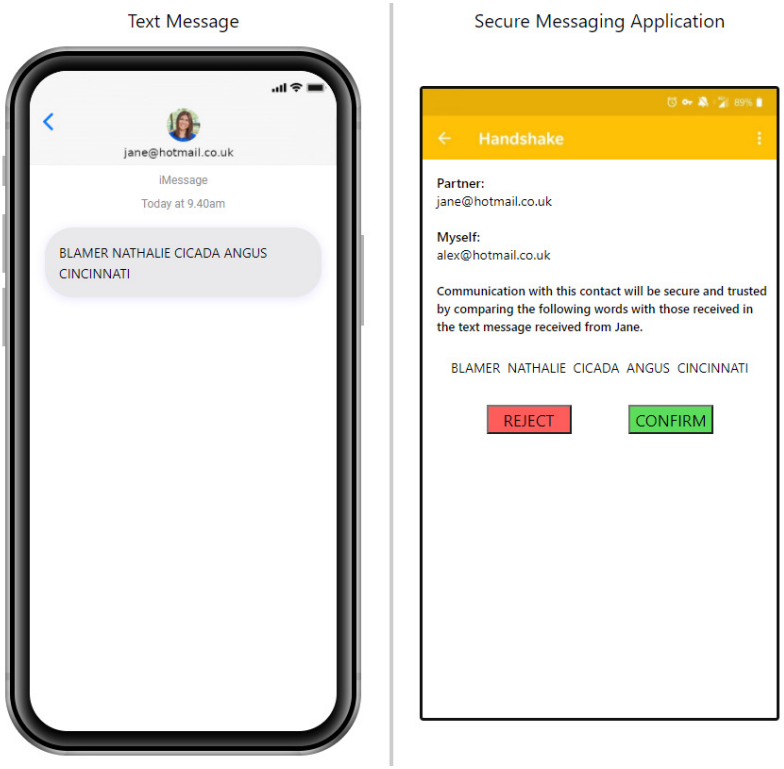


Figure 3.1: Visual verification task interface.

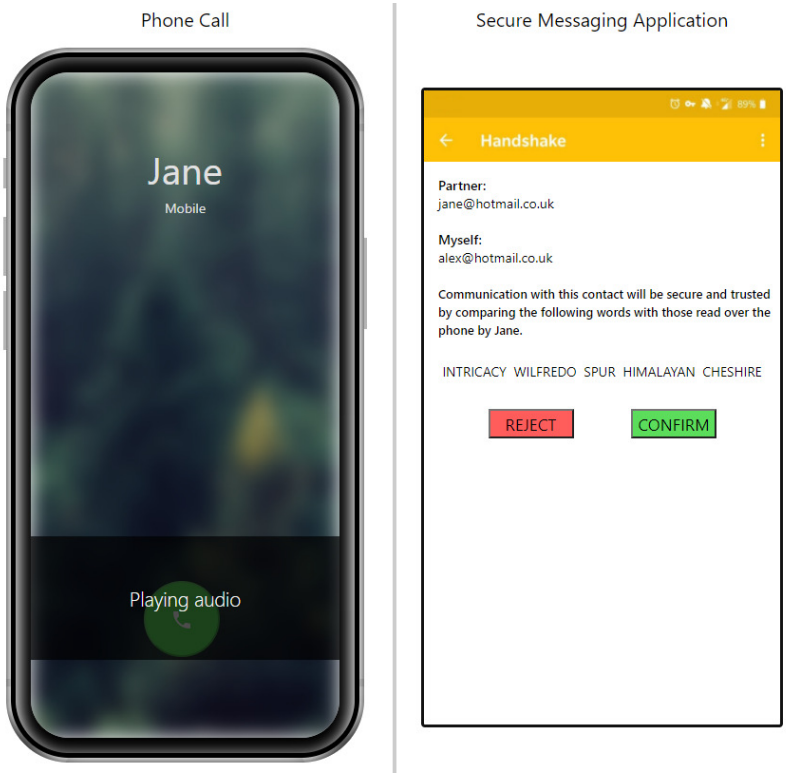


Figure 3.2: Verbal verification task interface.

Each set of verifications consisted of the following structure:

- Verifications 1, 2, 4 and 5 were all identical non-attack verifications. This was intended to enable the participant to develop some familiarity with the task before encountering attack scenarios.
- The third verification displayed a pair of totally mis-matching fingerprints. This enabled identification of any participants that did not provide their full attention, with the data from any who failed this check being eliminated.
- Two of the subsequent 15 verifications were randomly selected to simulate an attack. A low attack rate of 10% was used because attacks are uncommon in practice. The low attack rate also limited a participant's awareness of the possibility of attack. If attacks occurred with high frequency, then participants may begin to anticipate them and consequently display atypical behaviour when compared with the general user base, an effect that had been observed within earlier work [16].
- The other 13 verifications were identical non-attack verifications.

Performance was measured by time to make correct verifications and the number of errors. An attack error occurred if a participant accepted a non-matching verification, whilst a non-attack error occurred if an identical fingerprint pair was rejected. Perceived usability was measured on a set of five-level Likert items, with options from "Strongly Disagree" to "Strongly Agree". Standard usability instruments such as the System Usability Scale (SUS) [7] do not capture all the aspects of the user experience of interest, e.g. trust that the verification provides security and confidence in one's judgement. Therefore, a specific set of 12 questions was developed that measured six distinct areas of interest. A range of sources were used to develop the chosen question set, with the final including three from [16], three from the SUS and six original questions (see Table 3.1). The question set was designed to include two questions per dimension, with one each asked from a negative and positive context. The scores of the negative questions will be reversed so that high scores corresponded to high usability for all 12 questions. Justification that this question set is a good measure of perceived usability in this specific context is provided in Section 3.3.2.

The study investigated the following hypotheses:

H_1 There is a significant difference in the number of errors made using the visual and verbal fingerprint verifications.

H_2 There is a significant difference in time to make the correct decision between the visual and verbal fingerprint verifications.

H_3 There is a significant difference in perceived usability ratings between the visual and verbal fingerprint verifications.

H_4 Participants perform significantly better when the verification mode aligns with their preferred method to receive and process information.

Ethical principles of no harm, informed consent and data protection were followed and formal ethical approval was obtained from the author's departmental ethics committee.

Dimension	Rating items
Efficiency	I was able to do the verifications very quickly with this method*. Verifications using this method were unacceptably long.
Ease of use	The method was easy to use+. The method was unnecessarily complex+.
Low mental workload	The verifications did not need much mental effort. I needed to concentrate a lot.
Confidence	I would need a lot of technical support to be able to use this method+. I am confident that I can make verifications using this method* without making mistakes.
Repeat use	Completing the verifications using this method was annoying. Using this method is worth it for the additional security it provides.
Trust	Making verifications using this method would keep my communications secure*. I would not trust this method when sending confidential information.

Table 3.1: Dimensions of perceived usability and related concepts. Questions from [16] are marked(*) and questions from the SUS are marked(+).

3.2.2 Materials and Task

A single page web application was developed and embedded within the study to enable participants to interact with mockups of two mobile devices and perform a simulated fingerprint verification. The application was written using the Python Flask framework, and was based upon an earlier application developed by Fray during a master's project in 2021. Once deployed, the application can be accessed within a web browser and includes functionality that enables users to report if they perceive a pair of fingerprints to be identical or not. Upon clicking the relevant button, the application makes a microservice request to the backend server, which stores the response and related metadata (such as the button click time) within an SQLite database for downstream analysis. The source code of the latest version of the word-based application is available at `https://github.com/lli90/WordsExperimentApp`.

The web application did not allow study completion using mobile devices as their small screen dimensions could not adequately display the two virtual devices side by side. The PEP (`pep.security`) application was used as a template for the secure messaging application. PEP uses a bespoke word list called Trustwords to replace every 16 bits of the hashed key with one word from Trustwords, resulting in five-word fingerprints which are equivalent to 80-bit hashes [37]. PEP is supported by popular email clients such as Mozilla Thunderbird. PEP was chosen for the following reasons:

- It uses a word-based fingerprint representation, which have been shown to provide high usability and low error rates [16, 58].
- It introduces an interesting fingerprint combination method via a bitwise exclusive-or of the underlying fingerprint hashes before encoding the resulting 80 bit value into a five word string. This motivated investigation of the feasibility of identifying near pre-images with greater similarity than possible via a brute force search. However, no noticeable gains were identified.

The 11 question Visual–Verbal subscale of Felder's Index of Learning Styles (ILS) was used to measure an individual's preference for receiving and processing information. The Index of Learning Styles (ILS) was developed to gain insight into the preferred learning styles of engineering students and provide recommendations of how teaching can be adapted accordingly [19]. The ILS is a reliable and valid instrument to assess learning styles, and each of its four dimensions display high test-retest correlation coefficients after intervals of between four weeks and eight months [20]. The Visual–Verbal subscale assesses individual preference to receive and process information visually (e.g., through pictures and diagrams) or verbally (e.g., through written or spoken-aloud text). The subscale consists of 11 forced-choice questions and scored from -11 (if all questions are answered with a verbal preference) to +11 (if all questions are answered with a visual preference). The full set of questions is provided below, with visual and verbal responses marked as (vis) and (ver), respectively.

1. When I think about what I did yesterday, I am most likely to get
   (vis) picture.
   (ver) words.
2. I prefer to get new information in
   (vis) pictures, diagrams, graphs, or maps.
   (ver) written direction or verbal information.
3. In a book with lots of pictures or charts, I am likely to
   (vis) look over all the pictures and charts carefully.
   (ver) focus on the written text.
4. I like teachers who
   (vis) put a lot of diagrams on the board.
   (ver) who spend a lot of time explaining.
5. I remember best
   (vis) what I see.
   (ver) what I hear.
6. When I get directions to a new place, I prefer
   (vis) a map.
   (ver) written instructions.
7. When I see a diagram or sketch in class, I am most likely to remember
   (vis) the picture.
   (ver) what the instructor said about it.
8. When someone is showing me data, I prefer
   (vis) charts or graphs.
   (ver) text summarising the results.
9. When I meet people at a party, I am more likely to remember
   (vis) what they looked like.
   (ver) what they said about themselves.

10. For entertainment, I would rather

    (vis) watch television.
    (ver) read a book.

11. I tend to picture places that I have been

    (vis) easily and fairly accurately.
    (ver) with difficulty and without too much detail.

A post-task questionnaire assessed the perceived usability of each condition. Six dimensions of usability and related concepts were identified as being of interest and two five-level Likert items were used to measure each dimension (see Table 3.1). The scoring of items was reversed as appropriate so that a high score always indicates high usability. A post-study questionnaire asked participants which condition they preferred, their previous experiences using secure messaging applications, and also collected demographic information (gender, age, location, education level). The full list of questions is provided in Appendix F.

3.2.3 Security Assumptions

The study assumed the attacker randomly generates a large set of public keys before implementing a MitM attack. During the attack, they replace the authentic keys with ones from this set that display maximal similarity to the target fingerprint. This study simulated such an attacker using $2^{21.8}$ distinct PGP public keys scraped from publicly available PGP key servers. Previous work has identified security issues within these keys, including a number of RSA moduli which possess a common prime factor that enabled the efficient recovery of the associated secret key, but the effective security of their associated fingerprints was yet to be investigated [25, 34]. An analysis of the collected fingerprints found optimal attacks which possessed three out of five identical words which subsequently formed the attack set (see Table B.1).

Meylan et al. performed an eye tracking study with 40 participants which aimed to identify the areas of the fingerprint that were routinely looked at during a verification, finding that participants who looked only at the beginning and end of the sequence made an increased number of security errors [42]. Consequently, this structure was applied to the attacks of this research, with all differences confined to the third and fourth words. This is also consistent with the previous investigations described in the literature [16, 27, 58]. The attacker was also assumed to be unable to manipulate any messages exchanged over the OOB channel.

3.2.4 Procedure

Before running the main study, a pilot study was conducted with four participants recruited from within the author's research group. This led to improvements in the explanation of the task (e.g. to clarify that participants were expected to make multiple verifications in each condition). Several issues identified in the web application were also resolved, including increasing the size of the mock-up devices to improve the clarity of the task, and resolving a bug which occasionally caused one of the devices to be obscured from view. The main study procedure was as follows:

1. An information sheet explained the aims of the study, described the tasks participants would undertake and the data to be collected (see Appendix E). Participants were asked to confirm that they were over 18 and to consent to participation.
2. Participants were asked two screening questions: if they could view an image displayed upon their device and if they could play and hear a sound clip. This ensured that participants' devices supported the experimental conditions.
3. Participants then completed the Visual–Verbal subscale of the ILS.
4. Participants were randomly assigned to complete either the visual or verbal condition, compared the 20 fingerprints in that condition, and answered a post-task questionnaire to assess the perceived usability of that condition.
5. The above step was then repeated for the other condition.
6. Participants then answered the post-study questionnaire.
7. Participants were then thanked and provided with a reward of USD 2.00 (MTurk) or given the chance to enter into a prize draw (Local Networks).

3.2.5 Participants

Several methods of participant recruitment were used: through the University of York network, the author's personal contacts, and Amazon Mechanical Turk (MTurk). Participants recruited from local networks were entered into a prize draw, whilst participants from MTurk were paid USD 2.00 based on the pro-rated US Federal Minimum Wage.

Some researchers have raised doubts about the care with which MTurk participants undertake tasks [8], but others have found that MTurk participants produce data of equal quality to those recruited in more traditional ways [59]. Therefore, it was decided to use both more traditional recruitment methods and MTurk and compare data from the two sources. No differences in responses were detected between the two groups (verifications were made on times, errors and responses to rating questions), so results are presented for the whole sample.

In total, 75 people responded to the study, but data from 13 participants were eliminated: two experienced network errors, eight provided a partial response, and one failed to identify a totally mismatching attention check. Data from two participants who are dyslexic was also eliminated, since both verification modes involve reading words, including unusual words, which may be difficult for people with dyslexia to compare. All participants whose data were excluded were still rewarded for their time.

Data from 62 participants were analysed, 25 men (40%), 36 women (58%) and one who identified as non-binary. Age ranged from 18–24 to over 65, with the majority being in the 25–44 years range (71%, see Table 3.2). Educational level ranged from high school education to postgraduate degree, with the majority having a bachelors or postgraduate degree (73%, see Table 3.3). As the experimental task involved reading and listening, participants were asked whether they had a visual or hearing impairment, none reported any. For the same reason, participants were asked about their proficiency in English; 98% (61/62) rated it as good or excellent, and one as average. There were 29 participants recruited via the local networks, all located in the UK except one from the USA. There were 33 participants recruited via MTurk, all in the USA. Participants were also asked about their previous usage and attitudes towards secure messaging applications. Responses showed 94% (58/62) use at least one secure messaging application, and 60%

Age	Count
18–24	2
25–34	22
35–44	22
45–64	14
65 and over	1
Prefer not to say	1

Table 3.2: Age distribution.

Highest Education level	Count
High School education	9
Vocational training	4
Bachelors degree	32
Postgraduate degree	13
Other	3
Prefer not to say	1

Table 3.3: Education background.

(37/62) do so every day. Furthermore, 87% (54/62) of participants agree that "it is important to be able to have private conversations using secure messaging applications", yet 82% (51/62) of participants have never performed a fingerprint verification.

3.3 Results

Data did not meet the requirements for parametric statistics. Parametric analysis requires sample data to be normally distributed and satisfy the assumption of homogeneity of variance. Homogeneity of variance assumes that "the population variances (i.e. the distribution, or 'spread', of scores around the mean) of two or more samples are considered equal" [53]. The collected data did not possess these properties (see Figures 3.3 and 3.4), and so non-parametric statistics were used, with medians and semi-interquartile range (SIQR) reported as measures of central tendency and spread.

To compare between conditions, Wilcoxon related samples non-parametric tests were used. These test the difference between repeated measurements on a single sample to assess whether their population mean ranks are significantly different, which is most appropriate for non-parametric analysis of an experiment using a within participants design [68].

3.3.1 Performance

Effectiveness

In general, participants did not make many errors (i.e. identifying a non-attack verification as an attack or missing an attack verification). There were only 2 attack verifications in each condition, so errors could range from 0 to 2. There were 17 non-attack verifications, so errors could range from 0 to 17. Figures 3.3 and 3.4 show the distribution of errors for the non-attack and attack verifications. There was a difference in the effectiveness of the two verification modes, with participants making significantly more errors in the verbal non-attack condition than in the visual non-attack condition, see Table 3.4. There were no significant differences observed within the attack condition. Thus H_1, that there will be a difference between the errors on the two conditions, was supported for the non-attack condition.

	Verbal	Visual	Wilcoxon W	p-value
Attack verifications	0 (0.0)	0 (0.0)	1.19	0.23
Non-attack verifications	1 (0.5)	0 (0.0)	4.84	<0.01

Table 3.4: Median errors on correct verifications and SIQR for verbal and visual verification conditions with Wilcoxon Signed Rank tests of differences between conditions.

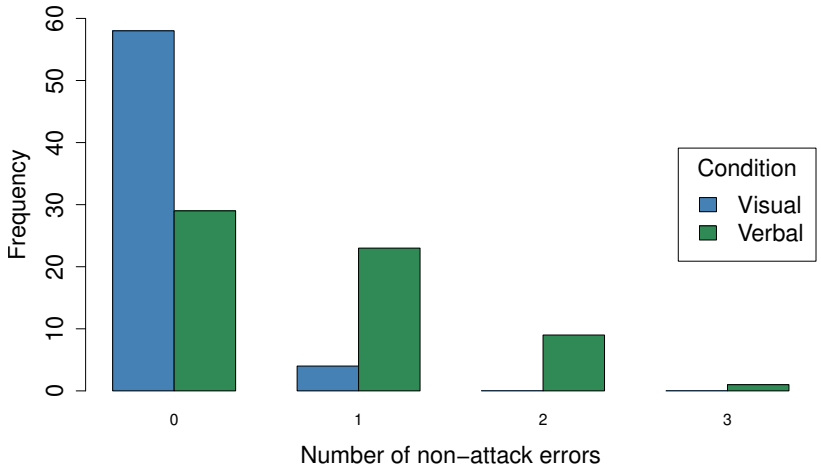


Figure 3.3: Number of errors by each participant on 17 non-attack verifications.

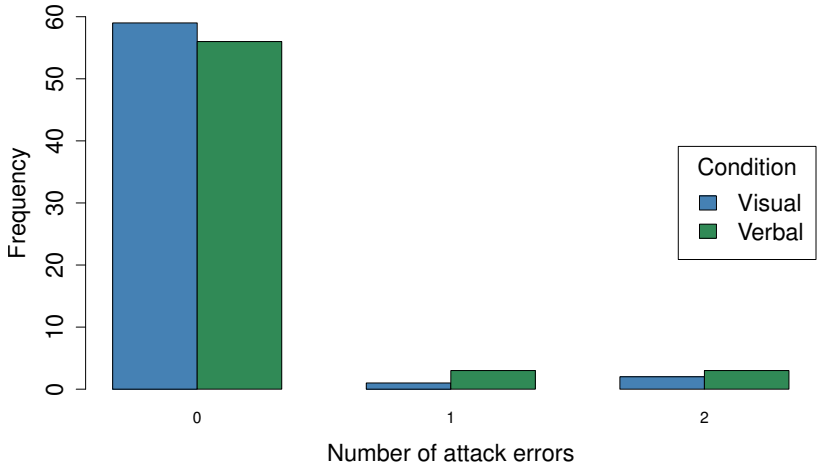


Figure 3.4: Number of errors by each participant on 2 attack verifications.

Efficiency

The time to complete correct verifications did not differ significantly between the visual and verbal modes for either the attack or non-attack verifications, as tested by Wilcoxon signed-rank tests for related samples (see Table 3.5). Thus H_2, that there is a difference between the times on the two conditions, was not supported.

	Verbal	**Visual**	**Wilcoxon W**	**p-value**
Attack verifications	5.49 (0.75)	5.50 (1.04)	0.22	0.83
Non-attack verifications	6.15 (0.55)	6.52 (1.96)	1.20	0.23

Table 3.5: Median times (seconds) and SIQR on correct verifications for verbal and visual conditions with Wilcoxon signed rank tests of differences between conditions

3.3.2 Perceived Usability and Related Concepts

The ratings on the two items for all six dimensions of perceived usability and related concepts were all highly correlated, (Spearman's ρ between 0.31 and 0.82, all $p < 0.05$ - see Table 3.6). Furthermore, Tables 3.7 and 3.8 provide an analysis of the mean scores on each dimension for the verbal and visual verification modes. All dimensions correlate at the 0.01 significance level for the visual verification mode. The same observations were largely repeated for the verbal verification mode, but the "Low Mental Workload" dimension did not correlate with "Trust", and "Efficiency" was only found to correlate with "Trust" at the 0.05 significance level. Consequently, median scores were calculated for each dimension and used in subsequent analyses.

Dimension	**Questionnaire for Verbal Condition**	**Questionnaire for Visual Condition**
Efficiency	0.604**	0.758**
Difficulty	0.457**	0.683**
Low Mental Workload	0.656**	0.820**
Confidence	0.365**	0.312*
Repeat Use	0.551**	0.539**
Trust	0.650**	0.539**

Table 3.6: Spearman's rank correlation coefficient of each dimension for both the verbal and visual verification mode (**: p < 0.01, *: p < 0.05).

	Efficiency	Difficulty	Low Mental Workload	Confidence	Repeat Use	Trust
Efficiency		**	**	**	**	*
Difficulty			**	**	**	**
Low Mental Workload				**	**	
Confidence					**	**
Repeat Use						**
Trust						

Table 3.7: Mean scores on each dimension for the verbal verification mode (**: $p < 0.01$, *: $p < 0.05$).

	Efficiency	Difficulty	Low Mental Workload	Confidence	Repeat Use	Trust
Efficiency		**	**	**	**	**
Difficulty			**	**	**	**
Low Mental Workload				**	**	**
Confidence					**	**
Repeat Use						**
Trust						

Table 3.8: Mean scores on each dimension for the visual verification mode (**: $p < 0.01$, *: $p < 0.05$).

Dimension	Verbal	Visual	Wilcoxon W	p-value
Efficiency	4.00 (1.00)	4.00 (1.00)	0.22	0.83
Ease of use	4.50 (0.50)	4.25 (0.75)	1.84	0.06
Low mental workload	4.00 (0.82)	3.00 (1.00)	4.21	< 0.01
Confidence	4.50 (0.50)	4.50 (0.75)	2.39	0.02
Repeat use	4.00 (0.75)	3.50 (1.00)	1.35	0.18
Trust	4.00 (0.82)	4.00 (1.00)	0.76	0.45

Table 3.9: Median ratings (with SIQR) of the perceived usability dimensions for verbal and visual conditions and Wilcoxon Signed Rank tests of differences between conditions

Table 3.9 shows participants' median ratings for the six dimensions for the visual and verbal conditions. There was a significant difference on the low mental workload dimension ($p < 0.01$), with the verbal condition perceived to require less mental workload than the visual condition. There was a strong trend towards a difference on the ease of use dimension ($p = 0.06$), with the verbal condition rated as easier than the visual condition. There was also a significant difference on the confidence dimension ($p = 0.02$). Although the median ratings were the same, inspection of the distributions showed that more participants had confidence in the visual condition than the verbal condition. These results show partial support for H_3, that there is a difference in the perceived usability of the two conditions, with the verbal condition being perceived as more usable on two out of six dimensions. In addition, at the end of the study, participants were asked which verification mode they would prefer to use, verbal or visual. There was an almost even split between preferences for each system, with 53.2% choosing verbal and 46.8% choosing visual. This was not a significant difference ($\chi^2 = 0.26$, $p = 0.61$).

3.3.3 Effect of Preferred Information Style: Auditory vs Visual

The participants' scores on the Visual–Verbal subscale of the ILS were skewed towards the visual end of the scale (see Figure 3.5). To create groups of approximately equal size for analysis, participants were divided into three groups:

- Very Visual (scores 7 to 11, 23 participants).
- Moderately Visual (scores 1 to 5, 21 participants)
- Verbal (scores -1 to -9, 18 participants)

To compare participants across the three different information processing preference groups, Kruskal-Wallis tests were used. Kruskal-Wallis tests are a non-parametric method for testing whether samples originate from the same distribution, with null hypothesis that the mean ranks of the groups are the same [32]. Hence, this test was most appropriate in determining if there was a significant difference in either effectiveness or efficiency between the three identified groups, which would be identified through rejection of the underlying null hypothesis.

There were no significant differences in time to complete correct verifications in either the verbal or visual conditions between the three groups of participants. Nor were there any significant differences in the errors made on the attack verifications. However, for non-attack verifications, all three groups made significantly more errors in the verbal

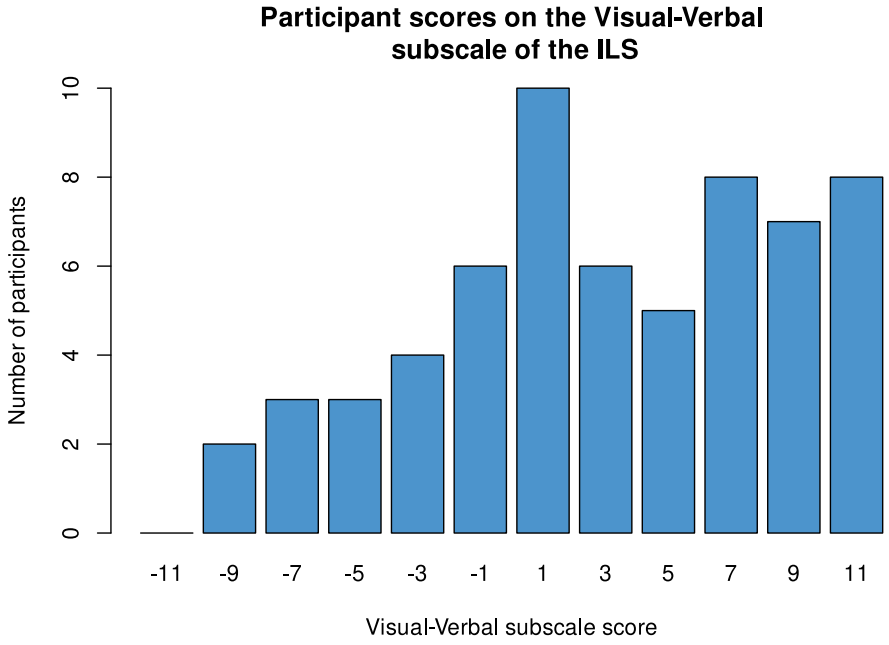


Figure 3.5: Participants scores from the 11 question Visual–Verbal subscale of the ILS.

	Non-attack errors
Very Visual	$W = 2.95, p < 0.01$
Moderately Visual	$W = 2.88, p < 0.01$
Verbal	$W = 2.64, p < 0.01$

Table 3.10: Results of Wilcoxon related samples tests for each group.

condition than in the visual condition (see Table 3.10). This does not support H_4, which predicted verbal users make more errors on the visual condition and visual users make more errors on the verbal condition.

3.4 Discussion

This study represented an exploratory investigation of the differences in effectiveness, efficiency and perceived usability between visual and verbal verifications of word-based key fingerprints. The study also sought to investigate the impact of a user's information processing preference, as measured by the Visual-Verbal subscale of the ILS.

Participants were found to make an increased number of non-attack errors using the verbal verification mode. One explanation for this result is that it is easier to mishear than misread a word. Without asking for the word to be spelt out, users are unable to check the spelling of any unfamiliar spoken words, and this uncertainty may cause users to reject fingerprints that they would otherwise accept if a visual verification mode was used. This explanation gains further support since participants perceived that the visual condition provided increased confidence that they were getting the verifications correct. In contrast, the verbal condition was perceived to require less mental effort and be easier

to use. This may be explained by participants viewing listening to the words as an easier task than a visual verification across two screens. Since fingerprint verifications are a secondary task to actual communication, these factors may be of greater importance to real users, and may motivate them to choose a verbal verification mode even though visual verifications would provide increased effectiveness and confidence.

A surprising result was the lack of effect between verification mode and the ILS Visual–Verbal subscale score. In fact, for the non-attack verifications all users made more errors when using the verbal mode, conflicting with the underlying hypothesis that users with a verbal preference to receive information would make more errors when using the visual verification mode. One interpretation is that the main effect of verification mode dominates, and visual verifications are significantly more effective against non-attack errors for all users. However, care must be taken before reaching this conclusion given the sample's skew towards participants with a visual preference to receive and process information. Further investigation, that includes a greater proportion of participants with a verbal preference, is required to clarify this. Another explanation is that the Visual–Verbal subscale does not measure the intended phenomena and an alternative scale may be more appropriate. This is discussed further in the following section.

All fingerprints were based on the Trustwords representation of PEP over PGP. The Trustwords word base contains many unusual and unfamiliar words which may have contributed to the increased number of non-attack errors in the verbal condition as participants were unfamiliar with their spelling or pronunciation. Additional investigation of fingerprint verifications that use a non-word based representation (e.g. the numeric representation used by Signal/WhatsApp) is required to determine if the effects observed in this study are specific to the Trustwords representation or fundamental properties of a fingerprint verification (see Chapter 6).

3.5 Reflections on Study Design Effectiveness

Though this study was able to identify an interesting usability effect related to the effectiveness of the two verification modes, it was unable to determine definitive answers to all of the intended research questions. Particularly understanding of differences on attack verifications between the two modes and the impact of a users underlying preference to receive information remain unresolved. Thus, a detailed evaluation of the implemented study design was performed to identify areas that may be improved to enable coverage of all intended research questions. The evaluation identified the following elements which required additional refinement:

Measurement of a Participant's Information Processing Preference

In retrospect, the Visual–Verbal subscale Felder's ILS was not an appropriate measure of a participant's information processing preference. Despite the rather promising subscale name, 7 of the 11 questions did not provide a clear differentiation between a verbal and visual option. Instead they tended to provide two visual responses (e.g. written text or diagrams), with listening to words and reading words included within the same dimension.

However, listening to a string of words is a very different task to reading them, and it is widely accepted that users have a preference for working with one or the other

medium [46]. A modern example is that some users may prefer to listen to an audiobook rather than reading the same novel in paperback. Furthermore, it is realistic to predict that users may show improved performance when they use a mode that is aligned with their own underlying preference to receive information.

Hence, it was determined that an instrument which implements an auditory–visual scale should be identified. Thus, a further review of the related literature was performed which investigated a wide range of possible instruments, see Section 4.1.

Effectiveness of Attacks

The effectiveness effect on the non-attack verifications indicates an underlying usability issue. This was surprising in many ways, but most significantly as the initial premise of this study had predicted a security related issue. However, users appeared to relatively proficient at identifying the non-matching fingerprint pairs included within this study, with relatively few errors observed overall. There are two obvious explanations for this. Either participants are very good at identifying such differences, or the included attack fingerprints possessed insufficient similarity. It was determined that it was likely the latter, and even if it was the former then including fingerprints with even greater levels of similarity was a sensible next step. However, it was important to ensure that the levels of similarity included within the attack fingerprints could feasibly be produced by a well resourced and highly motivated attacker, so as to not produce results which lack ecological validity.

Floor Effect on Attack Verifications

Another explanation for the lack of user errors on attack verifications, is that participants had few opportunities to make such errors. The study included only two attack verifications within each set of 20, which ultimately introduced a floor effect. Though there were good reasons for the low attack rate, motivated by a desire to prevent users from becoming familiar with the attacks and as attacks are uncommon in practise, it made identification of a significant effect between conditions difficult. As a consequence the study would be unlikely to observe an effect related to the attack verifications even if one were to exist. Thus, it was decided that all subsequent related studies would include an increased number of attacks.

These limitations combined to create two new research questions that are investigated as part of the following chapter:

1. How can a user's auditory–visual information processing preference be measured?
2. What is the optimal level of fingerprint similarity achievable from a computational search performed by a well resourced and highly motivated attacker?

3.6 Conclusions

Though exploratory and not without its limitations, this initial study identified some interesting results. Visual verifications were found to be more effective against non-security errors and perceived to provide increased confidence, yet verbal verifications were perceived to be easier and require less mental effort. Though participants often displayed a preference for a particular verification mode (based on measures of both performance and perceived usability), this did not correlate with their score on the Visual–Verbal subscale of the ILS. This suggests that the associated research questions may have some merit, and are worth of further investigation.

The post study analysis of the experimental design identified a range of factors which limited its effectiveness. However, this evaluation signifies one of the major milestones of this thesis, as it motivated efforts to formulate an improved methodology so as to implement an experimental design that had the ability to fully investigate the research questions described in Section 1.2. In Chapter 4, each of the limitations identified above in Section 3.5 are investigated and resolved to produce an improved methodology that is then applied to the subsequent studies described in Chapters 5 and 6.

Chapter 4

Improving The Experimental Design

This chapter details the results of the efforts to develop an improved experimental design by addressing the limitations described in Section 3.5. This work consisted of the following phases:

- Section 4.1.1 reviews the pre-existing information processing preference instruments included within the literature and aims to identify a validated measure that includes a Auditory–Visual dimension. This was ultimately unsuccessful, in large part due to inclusion of verbal dimensions that included both written and spoken text.
- Section 4.1.2 describes development of a custom scale to measure a participants auditory–visual information processing preference. This was formed from a collection of suitable questions that were identified within previous work, followed by a principal components analysis and statistical testing of the suitability of the custom scale.
- Section 4.2 details the implementation of a simulated brute force pre-computation, to enable identification of fingerprint pairs with increased similarity that were expected to cause an increased number of errors within the attack condition.
- Section 4.3 reports investigation of the impact of different similarity metrics upon user performance. It occurred that there were two plausible methods to measure the similarity of two words, either by their orthographical edit distance or their phonological edit distance. Given that it was unclear which type of similarity would cause users to make the most errors, a further human factors investigation was implemented to determine the impact of this difference.

These results identified significant improvements of the underlying experimental design, which were subsequently integrated into the later studies described in Chapters 5 and 6.

4.1 Development of A Custom Information Processing Preference Scale

As discussed in Section 3.5, reflection upon of the design of the exploratory study found the Visual–Verbal subscale of the ILS to be an ineffective measure of a participant's preferred method to receive information. The subscale includes the processing of spoken words and written words within its visual dimension, but this research requires these two aspects to be considered separately.

Paivio's Dual Coding Theory predicted that users possess distinct channels for the processing of verbal and non-verbal information and hypothesised that the generation of mental pictures may aid learning [10,44]. Baddeley's Model of Working Memory proposes a similar structure for working memory, introducing the concepts of "articulatory loop" for the processing of word-based information, and the "visuo-spatial scratch pad" for the processing of visual images related to the physical world. The model also predicted differences in how the human brain passes word-based information to the articulatory loop depending upon whether it is presented audtorily or as written text, with auditory data processed by the phonological input store whilst written text utilises the articulatory rehearsal process [3].

Penney investigated modality effects related to the retention differences between information presented either visually or verbally, and subsequently introduced the "separate streams" hypothesis which predicts "that the processing of auditorily and visually presented verbal items is carried out separately in short-term memory". Penney extended Baddeley's theory by arguing "that the short-term memory trace laid down when the subject silently articulates a visually presented item or imagines the sound of that item does not contain the same information as the trace resulting from auditory presentation of the item" [48].

In addition, Mayer's Cognitive Theory of Multimedia Learning predicts users would demonstrate improved learning when the material is simultaneously provided via multiple mediums, for example as both pictures and text. The model "assumes that the human processing system includes dual channels for visual/pictorial and auditory/verbal processing" and observes that "processing of spoken words occurs mainly in the auditory/verbal channel, however processing of printed words takes place initially in the visual/pictorial channel and then moves to the auditory/verbal channel" [38].

These theories combine to support the prediction that the processing of spoken words and written words is different. Given this difference, it is realistic to hypothesise that users may possess a preference for processing either written or spoken text, and that this preference may impact their performance within a key fingerprint verification. The challenge lies in identifying an effective measure of such a preference.

4.1.1 Review of Existing Cognitive Style Instruments

A common approach to the measurement of information processing preference is the consideration of a user's learning or cognitive style. Anastasi et al. state "cognitive systems refer essentially to one's preferred and typical modes of perceiving, remembering, thinking, and problem solving" [40]. "They are regarded as broad stylistic behavioural characteristics that cut across abilities and personality and are manifested in many activities and media" [2]. There has been considerable psychological and educational research into the concept of different cognitive or learning styles, with many different dimensions and models proposed. However, one of the more robust is visual-verbal processing. While the concept of learning style is controversial [69], and people are undoubtedly flexible in the ways they can process information, they may have preferences which would affect their perception of the usability of an authentication system.

Investigation of learning styles has been particularly prominent within education, and such research has inspired the creation of a large number of learning styles instruments. Coffield et al. provided an extensive assessment of these instruments, but ultimately conclude that that "too much is being expected of relatively simple, self-report tests", that "learning style researchers do not speak with one voice; there is widespread disagreement about the advice that should be offered to teachers, tutors or managers" and "there is a dearth of rigorously controlled experiments and of longitudinal studies to test the claims of the main advocates" [11]. Further criticism stems from application of the related "matching hypothesis", which predicts that students will demonstrate improved learning when the teaching method is tailored towards their individual learning preference [12,46]. Furthermore, Penney made a similar observation, finding that "in spite of the large and robust effects of presentation modality found in short-term memory tasks, there was no evidence of any permanent effects on learning, and modality effects in long-term memory tasks were conspicuously absent" [48]. However, it is important to note that the underlying scenario of this research involves a short term memory task rather than one of learning. The information shared within the fingerprint is ephemeral and need only be stored within the users consciousness for the duration of the verification task. This is in contrast to a passphrase where the words must be memorised and retrieved repeatedly.

Thus, a detailed search of previously proposed information processing instruments sought to identify an effective measure of a user's preference to receive information upon an auditory–visual scale. A wide range of instruments were identified, yet for a variety of reasons, none were deemed to be appropriate for the specific context of this research. However, there were examples of individual questions that could be of potential use. An overview of each instrument and their limitations is provided in the following sections.

Visualiser-Verbaliser Hypothesis

The majority of pre-existing instruments within the literature tend to investigate the visualiser-verbaliser hypothesis, which predicts that users possess a preference for processing words or pictures [45,51]. However, as with the Visual–Verbal subscale of the ILS, these instruments combine the processing of written words and spoken words into a single dimension, and are consequently not appropriate for this research. Other instruments included additional aspects such as how users prefer to share information with others, distracting from the phenomena under investigation [9].

VARK

Based upon his personal experience as an educational inspector, Fleming perceived the visual dimension to be more nuanced, and divided visually presented information into two distinct categories:

- **Visual**: preference for graphical and symbolic ways of representing information.
- **Read/Write**: preferences for information printed as words.

Subsequently, the 13 multi-response item VARK[1] questionnaire was developed which aimed to enable learners to develop a personal understanding of their own metacognition [21]. Fleming also developed a variety of "help sheets" that are intended to help learners develop study strategies that are compatible with their preferred method to receive information[2].

This questionnaire appeared to possess some promise for use within this research, particularly as the Auditory and Read–Write dimensions relate to the information processing in the two verification conditions. It appeared that participants' different responses to the questionnaire would enable identification of their auditory–visual information processing preference, particularly if the answer options were restricted to these two dimensions.

Unfortunately, there were a number of unforeseen complications:

- Leite et al. investigated the reliability and validity of the VARK questionnaire, finding support for the claimed four dimension model, particularly when viewed as "as a low-stakes diagnostic tool by students and teachers". However, they caution about its use as a predictor of performance within research studies without further investigation [33].
- Only 6 of the original 13 questions relate to the scenario of receiving information. Fleming found this to be too great a restriction, but the inclusion of additional scenarios, such as presenting information, again distracts from the specific phenomena of interest and at worst may introduce a confound.

Moreover, The VARK questionnaire is copyrighted, and permission must be obtained before use. Permission is only provided if researchers agree to very specific permissions about its use. Specifically:

- Modification of the questionnaire was prohibited. This meant that all four responses would need to be provided for each question.
- Analysis of the raw scores was prohibited. Instead the analysis must investigate differences between groups, using a proprietary algorithm to convert a participant's raw score to one of 25 possible learning modalities.

A request was made for an exemption from these restrictions, but it was denied. Ultimately, it was decided not to use the VARK questionnaire, as these restrictions would prove incompatible with the intended research methodology.

[1] VARK is an acronym for Visual, Auditory, Read/Write, Kinaesthetic

[2] See `https://vark-learn.com/`

The Learning Style Inventory (LSI) and Productivity Environmental Preferences Survey (PEPS)

Dunn et al. developed a pair of preference scales that sought to identify individual differences across a wide range of tasks:

- **The Learning Style Inventory (LSI) [18]:** a 72-item inventory aimed at school children.
- **The Productivity Environmental Preferences Survey (PEPS) [17]:** a 100-item inventory aimed at adults.

Both inventories seek to identify differences across over 20 elements. This included 13 scenarios which investigated preferences of receiving information either auditorily or visually, the underlying phenomena of interest. But both also investigated other aspects, including environment, emotionality, sociological preferences and psychological processing inclinations [49]. Consequently, the large number of items within each inventory and the range of aspects unrelated to this research meant neither was deemed to be directly compatible with the aims of this research. However, the identification of a subset of questions from within these pre-existing inventories motivated a hypothesis that it may be feasible to develop a custom inventory that is suitable for measuring a user's information processing preference on an auditory–visual scale. These efforts are discussed in further detail in the next section, ultimately leading to the development of a custom scale which meets the requirements of this research.

4.1.2 The IPP-AV Scale

Though the extensive search of the literature was unable to identify a validated scale that facilitated measurement of a user auditory–visual information processing preference, the search did identify a subset of questions that included a scenario that investigated a user's preference to interpret information either auditorily or visually. It was hypothesised that a suitable selection of these questions could be combined to produce a custom scale to facilitate measurement of the users individual auditory-visual information processing preference, facilitating investigation of its effect upon fingerprint verification performance and within later studies.

The remainder of this section describes the process taken to develop this scale, which is subsequently named the auditory–visual information processing preference (IPP-AV) scale, and some initial analysis of its reliability.

Question Selection

The two main sources of questions were the LSI and PEPS. As discussed above, the earlier literature review identified a collection of 13 questions which investigated a users preference to receive information either auditorily or visually. Subsequent analysis identified seven distinct questions that were chosen for inclusion within the initial version of the IPP-AV scale. Of these seven questions, six were found to form obvious pairs of scenarios, with each pair including a question that described an auditory preference and one that described a visual preference (see questions 1–6 of Table 4.1). This observation ultimately

formed the methodology employed to develop the scale, with any other questions included only if they shared the same characteristics.

In addition to the initial six questions there was another question contained within the LSI that did not immediately possess a contrasting partner question to enable inclusion within the custom scale:

- "If I have something new to learn, I would rather talk with someone to learn about it".

However, earlier review Richardson's Verbalizer-Visualizer Questionnaire (VVQ), formualted from evaluation of Paivio's Individual Differences Questionnaire [51], did identify a question that appeared to show some promise as a contrasting partner:

- "I prefer to read instructions about how to do something, rather than have someone show me."

Though not quite aligning with the intended auditory–visual dimensions, this could be easily resolved by changing "show me" to "tell me", and enabling generation of the missing converse question. These two questions were also added to the initial scale.

Finally, Kirby et al. extended upon the work of Richardson's VVQ to produce three scales that considered information processing as a combination of verbal processing, dream vividness and mental imagery. Though again not directly applicable to measurement of auditory–visual processing, as auditory and visual textual processing is again combined within the scales verbal dimension, the paper did introduce a new question which included an auditory–visual information processing scenario:

- I have a hard time remembering words to songs.

However, none of the other identified questions considered a similar scenario from a visual perspective. To resolve this challenge, a bespoke question was developed:

- I have a hard time remembering quotes from books I've read.

The full list of initial 10 questions of the custom IPP-AV scale, and their source, are provided in Table 4.1. To produce an IPP-AV score participants were asked to rate their agreement with each statement upon a 7-level Likert items (with options from "Strongly Disagree" to "Strongly Agree"). The scores of five questions were reversed so that high scores correspond to an auditory preference to receive information (see Table 4.1), facilitating assessment of a user's underlying auditory–visual information processing preference.

Though this set of 10 questions appeared to show promise as a measure of a participants auditory–visual information processing preference (IPP-AV), it was yet to be determined if this custom scale aligned with a participant's pre-existing preference to receive information. A human factors study with 75 participants aimed to determine this, through application of a principal components analysis, and comparison of their self reported preference to receive information with their IPP-AV score.

[3]The original question used the suffix "than have someone show me". This was modified to align to the intended auditory–visual dimensions.

No.	Question	Source
1	The things that I remember best are the things that I hear.	[17]
2	The things that I remember best are the things that I see or read*.	[17]
3	I learn better by reading than by listening to someone*.	[17]
4	*I learn more by listening to someone explain something than by reading about it.*	[18]
5	*I remember things better when people tell me them rather than when I read about them.*	[18]
6	*I remember things better when I read, rather than when someone tells me them*.*	[18]
7	I have a hard time remembering words to songs*.	[28]
8	I have a hard time remembering quotes from books I've read.	Bespoke
9	I prefer to read instructions about how to do something rather than someone tell me*[3].	[51]
10	If I have something new to learn, I would rather talk with someone to learn about it.	[18]

Table 4.1: The initial 10 questions of the custom IPP-AV scale. Questions for which the scoring was reversed are marked *. Italicised text denotes questions that were dropped after a principal components analysis.

Materials and Task

A Qualtrics questionnaire was created to enable participants to answer the custom set of 10 questions. Before answering the 10 questions, participants were asked which method they preferred to receive information, with the following options provided:

- Visual.
- Verbal.
- Both.
- Don't know/ Never thought about this.

After answering the 10 questions, participants were asked standard demographics questions (gender, age, location, education level). The full list of questions is provided in Appendix F.

Procedure

1. An information sheet explained the aims of the study, described the task that participants would undertake and the data to be collected (see Appendix E). Participants were asked to confirm that they were over 18 and to consent to participation.
2. Participants were then asked to report what they perceived to be their preferred method to receive information.
3. Participants then completed the 10 questions of the proposed IPP-AV scale (see Table 4.1.)
4. Participants then answered the demographic questions.
5. Participants were then thanked and provided with a £1.00 reward.

Participants

All participants were recruited via the Prolific platform and were paid £1.00 upon successful completion of the study. On average study completion took 167 seconds (SD = 57), so participants were paid for 5 minutes based on Prolific's payment model. Data from all 75 participants who responded to the study were analysed, 38 (50.7%) men and 37 women (49.3%). Participants ages ranged from 18–81, with a mean age of 38 (SD = 14.8). Prior to beginning the study participants were asked about their perceived preference to receive information; 24 answered that they were "visual" people (32.0%), 2 said they were verbal (5.3%) and 47 (62.7%) said they were "both". No one answered "Don't know/ Never thought about this".

4.1.3 Results

Preference	N	IPP-AV Score (median, SIQR)	Wilcoxon W	p-value
Visual	24	19.0 (8.25)	−2.96	0.003
Verbal	4	39.0 (3.0)	+2.64	0.059
Both	47	25.0 (19.00)	−1.09	n.s.

Table 4.2: Results of one sample Wilcoxon signed rank tests for each group.

A principal components analysis (PCA) was conducted (with direct Oblimin rotation) and found that a subset of 7 questions produced a one factor solution which accounted for 51.4% of the variance, with the three questions highlighted in italics within Table 4.1 not included. Therefore, IPP-AV score was calculated for each participant which was the total of the ratings on the 7 questions. IPP-AV scores could range from 7 (strong auditory preference) to 49 (strong visual preference).

Non-parametric statistics were used to analyse the relationship between a participants self reported preference to receive information and their score on the custom IPP-AV scale. Medians and semi-interquartile ranges (SIQR) are reported as measures of central tendency and spread, with one sample Wilcoxon signed rank tests used to compare the median of each group with the midpoint score of 28.

The results, presented in Table 4.2, show that participants with a self-reported visual preference score significantly below the midpoint score, those with a self-reported verbal preference nearly score significantly above the midpoint score (there are only 4 of them), and those who reported a "both" preference score not significantly different from the midpoint. All of these results align with our expectations and provide support for the suitability of using the custom IPP-AV scale to measure a participant's preference to receive information upon an auditory–visual scale within subsequent studies.

4.2 Simulated Brute Force Pre-Computation

A further limitation of the exploratory study discussed in Chapter 3 was that the selected attack fingerprints did not possess enough similarity and lead to few errors on attack verifications (see Table B.1). Thus, an important addition to the existing experimental method would be the inclusion of attacks which possess increased levels of similarity to an authentic fingerprint, which in turn would be expected to cause participants to make an increased number of attack errors. Yet it was important to ensure that the included attacks remained feasible for a well resourced and highly motivated attacker so as to ensure that the experimental task replicated that which users may face during real attacks. This was a limitation of previous research [16]. These factors generated the following research question:

- What is the optimal level of fingerprint similarity achievable from a computational search performed by a well resourced and highly motivated attacker?

The remainder of this chapter describes the assumed capabilities of a hypothetical attacker and the results of a simulated large computational effort based on their optimal attack strategy. The levels of similarity of the resulting attack set are then compared with those utilised in Chapter 3, observing a significant increase.

4.2.1 Method

An obvious approach for an attacker seeking to implement the attack described within this research would be to compute and store a large number of key pairs and their associated fingerprints, with 2^{60} chosen as a realistic upper bound. The upper bound was chosen as it is likely feasible for the envisaged attacker, for example a nation state actor. Furthermore, though difficult and expensive to implement, there are examples of such large computational efforts that have been performed within the academic community, for example the efforts to factor RSA-768 [30], and it is reasonable to suggest that there are attackers willing to implement a similar effort to facilitate an attack against a targets of sufficiently high value. The computationally expensive step would only need to be performed once, with the attacker replacing the intercepted public key with one from their set whose fingerprint possesses maximal "similarity" to the target. It is assumed that the attacker has an efficient method to both store the attack set and can efficiently identify an optimal attack.

Consideration of methods to maximise the similarity of the central fingerprint section is the crucial aspect of the enhancement of the included attack set. Previously attacks implemented an "IIDDI" model[4] with the attacker assumed to have limited control of the non-matching chunks. This improved simulation instead seeks a "IISSI" structure, with effort placed in generating attacks which limit the differences in non-matching sections, and should ultimately cause participants to make more attack errors.

Given the associated cost of a practical simulation of such a search, and as it was not intended to target real users, it was decided to simulate the set of fingerprints that such a computation may generate. To maintain comparability with the attacks included within Chapter 3, it was decided that fingerprints should:

[4] I: Identical, D: Different, S: Similar

- Be 80 bits long.
- Consist of five equal sized chunks, which can be subsequently encoded as words or numbers.
- Contain three identical chunks, in positions 1, 2 and 5.

It is important to quantify the concept of similarity between two fingerprints.

- Given that the attack will possess an identical prefix and suffix, consideration of similarity is restricted to the third and fourth chunks.
- Individual chunks may be called "similar" if they possess a low "edit distance", which quantifies the distance between two chunks as the number of letter edits to transform one to another. As an example "trust" and "just" have an edit distance of 2: deletion of the "t" and replacement of the "r" with "j". A numerical example is "12345" and "12675".
- The impact of chunk length also plays a factor, and it may be argued that longer chunks with the same edit distance are in some way more similar than the shorter chunk (e.g "translator" and "transactor").
- It is possible that the simulation could generate a central section with one highly similar chunk and one that is quite different to the target – an "IISDI" structure. These would likely prove less effective than attacks which evenly distribute similarity across both chunks, and so the overall similarity of a central section was set to be the average similarity of the two chunks.

These factors led to generation of the following definition of similarity which produces a normalised coefficient between 0 (entirely different) and 1 (identical):

Definition 4.2.1 (Fingerprint Similarity Coefficient (FSC))**.**

$$\mathrm{FSC}(X,Y) = 1 - \frac{1}{2}\sum_{i=1}^{2} \frac{\mathrm{ED}(x_i, y_i)}{\max\{\mathrm{len}(x_i), \mathrm{len}(y_i)\}}$$

Where $\mathrm{ED}(x, y)$ is the edit distance between chunks x and y.

Standard counting arguments imply that identification of a 48-bit collision, akin to an attack fingerprint with three identical chunks, would require an effort of at least 2^{48}. The remaining 2^{12} work factor could then be applied to maximising the "similarity" of the third and fourth chunks. Hence, a simulation of 2^{12} randomly generated pairs of 16-bit integers was performed to assess the achievable levels of similarity by such a hypothetical attacker. A group of high profile target users was simulated by a distinct set of 100 randomly generated pairs, with a subsequent computation analysing the FSC between each target–attack pair, resulting in the identification of the attack which maximised the FSC for each target. These were subsequently padded with randomly generated 16-bit integers to form 5-chunk fingerprints which could then be encoded as desired.

The final attack set was the reduced to the 15 fingerprint pairs with maximal similarity (see Table C.2). This ensured that participants encountered attacks with the highest possible similarity, whilst ensuring that it was unlikely that participants would encounter the same 5 attacks.

	Key Servers	Simulation	Mann-Whitney U	p-value
Similarity (FSC)	0.12 (0.11)	0.67 (0.04)	510	< 0.01

Table 4.3: Median and IQR FSC for the two attack sets with Mann-Whitney tests of differences between sets.

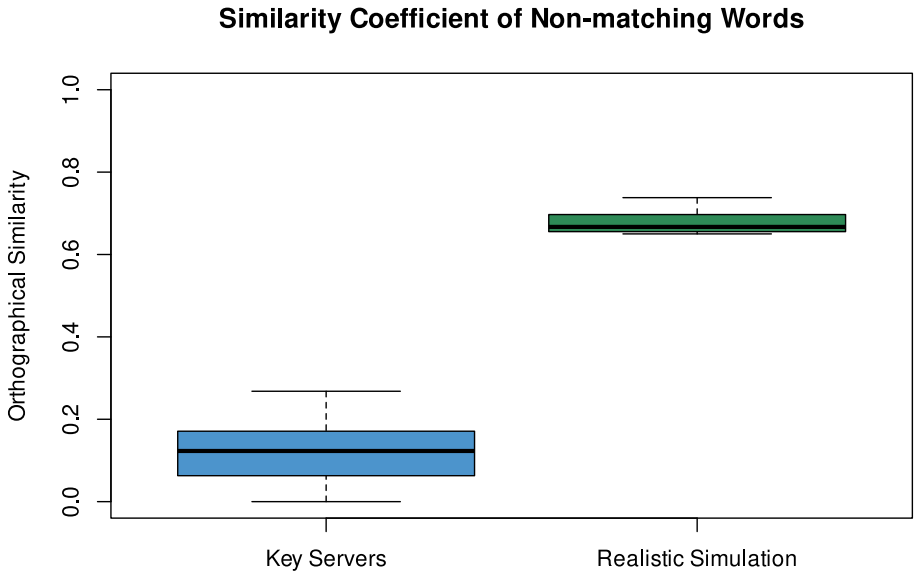


Figure 4.1: Distribution of the FSC for the two attack sets.

4.2.2 Results

Figure 4.1, shows the differences in fingerprint similarity between the simulated attack set and those selected from publicly available sources. The FSC of a fingerprint pair lies in the range $[0, 1]$. A non-parametric Mann-Whitney test was performed to compare between the two sets, with medians and interquartile ranges (IQR) reported as measures of central tendency and spread (see Table 4.3). This demonstrates that the intention to simulate an attack set with greater similarity was achieved.

4.3 Impact of Similarity Metric Upon User Performance

4.3.1 Introduction

The investigation of the optimal levels of similarity achievable by a well resourced and highly motivated attacker (see Section 4.2) generated a related question about measurement of fingerprint similarity. As discussed in the previous section, an important component of the identification of similar words is measurement of their edit distance. However, there are two distinct approaches to measuring the edit distance of two words:

- **Orthographical distance**: applies the metric to the dictionary spelling of the two words.
- **Phonological distance**: applies the metric to the phonological spelling of the two words. In this case the phonological spelling was based upon the International Phonetic Alphabet (IPA).

As an example, "LOOSE" and "JUICE" have orthographical edit distance of 4, but a phonological edit distance of 1.

It was unclear which distance metric would be most appropriate to determine the similarity of two fingerprints, and in turn cause users to make more errors upon attack verifications. It was feasible that one of the two metrics would cause a significant increase in errors within both verification modes. Or the relationship may have been dependant upon the underlying verification mode, with orthographically similar attacks causing more errors within the visual mode with phonologically similar attacks causing more errors within the verbal mode.

The following section describes a human factors study with 41 participants that aimed to establish the importance of this relationship. The study followed a similar design to the exploratory study described in Chapter 3, and sought to identify any significant differences in success rate between attacks with high orthographical or phonological similarity. Ultimately analysis of the data did not identify a significant difference between the two types of similarity, indicating that this is not a significant factor in the number of errors made within the attack condition. Thus, this variation could be controlled within later studies, as discussed in Section 4.3.4.

4.3.2 Method

Design

This study included a within-participants design with two conditions. One included orthographically similar attacks whilst the other included phonologically similar attacks. Each participant was randomly assigned to use either the visual or verbal verification mode throughout the study, and subsequently performed two sets of 25 verifications, with the order of making verifications counterbalanced.

The underlying experimental design was similar to that described in Section 3.2.1 with five modifications:

- To mitigate the impact of floor effects, each set of verifications contained five attack verifications instead of two.
- The set of attack fingerprints used those identified from the simulated brute force pre-computation described in Section 4.2, rather than fingerprints collected from public key servers. These attacks displayed increased similarity which was expected to lead to an increased number of attack errors.
- An additional attention check was included at the end of the study. This aimed to identify any participants who became distracted or fatigued and began to "click through" the study.
- Participants were not asked questions regarding their preferred method to receive information, as this was not a research question that this study intended to investigate.
- Participants were not asked to complete the post-task or post-study questionnaires, which investigate the perceived usability of each mode and prior usage of secure messaging applications.

To generate the phonological attack set, the collection of 2^{12} pairs of 16 bit integers generated in Section 4.2 were mapped the relevant Trustword and then converted to their IPA encoding. A similarity computation was then performed to identify the optimal target–attack pair for each of the 100 members of the target set, which was again reduced to the set of 25 with most similarity. The resulting phonological attack set consisted of pairs with FSC in the range $(0.708, 0.817)$, compared to $(0.65, 0.74)$ for the orthographical attack set (see Tables C.1 and C.2).

The study included a single hypothesis:

H_5 There is a significant difference in the attack success rate between attacks using a phonological measure of similarity compared with those using an orthographical measure.

Ethical principles of no harm and informed consent were followed and formal ethical approval was obtained from the author's departmental ethics committee.

Materials and Task

Participants made verifications using the same web application developed for the exploratory study described in Chapter 3. Only the minimum required changes to implement the modified design were made to retain task consistency.

Procedure

Before running the main study, a pilot study was conducted with three participants from the author's research group. No further improvements were identified. The main study procedure was as follows:

1. An information sheet explained the aims of the study, described the tasks participants would undertake and the data to be collected (see Appendix E). Participants were asked to confirm that they were over 18 and to consent to participation.
2. Participants were asked two screening questions: if they could view an image displayed upon their device and if they could play and hear a sound clip. This ensured that a participants device supported the experimental conditions.
3. Participants were randomly assigned to use either the visual or verbal verification mode throughout the study.
4. Participants were randomly assigned to first complete either the orthographical or phonological condition, and compared the 25 fingerprints in that condition.
5. The above step was then repeated for the other condition.
6. Participants were then asked to answer demographic questions (gender, age, location, education level - see Appendix F), thanked for their participation and provided with a reward of $2.00 (MTurk) or given the chance to enter into a prize draw (Local Networks).

Participants

Age	Count
18–24	1
25–34	8
35–44	11
45–64	20
65 and over	1
Prefer not to say	0

Table 4.4: Age distribution.

Highest Education level	Count
High School education	11
Vocational training	1
Bachelors degree	20
Postgraduate degree	6
Other	3
Prefer not to say	0

Table 4.5: Education background.

Two methods of participant recruitment were used: the author's personal contacts, and through Amazon Mechanical Turk (MTurk). Again, participants recruited from local networks were entered into a prize draw, whilst participants from MTurk were paid $2.00. The initial implementation of this study received data from 60 participants. However, communication errors between the Qualtrics platform and the fingerprint verification application meant that it was impossible to track participants between these two platforms. After resolving these issues, the study was implemented a second time with a distinct set of participants. In total, 53 participants responded to the second study, but data from 12 participants was eliminated: 11 provided only a partial response and one reported an auditory impairment which could have directly impacted their performance. All participants whose data were excluded were still rewarded for their time.

Data from the remaining 41 participants were analysed, 19 men (46.3%) and 22 women (53.6%). The eliminated participants were not evenly distributed across the sample and

of those that remained, 23 (56.1%) had been assigned to the verbal condition and 18 (43.9%) to the visual condition. Participants' ages ranged from 18–24 to over 65, with the majority being in the 45–64 years range (48.7%), see Table 4.4). Educational level ranged from high school education to postgraduate degree, with almost half having a bachelors degree (48.7%, see Table 4.5).

As the experimental task involved reading and listening, participants were again asked about their proficiency in English: 40 rated it good or excellent (97.6%). There were six participants recruited via the author's local network, five located in the UK and one who responded "Prefer not to say". There were 35 participants recruited via MTurk, 34 from the USA and one from the UK. The unequal split between recruitment sources was not intentional, and the causes and impact are discussed in Section 4.3.4. No participants failed the first fully mismatching attention check, but 70.7% (29/41) were observed to fail the second attention check. Given the relatively few errors made within this study, this appears to indicate that participants became fatigued rather than providing poor quality data. This is discussed further in Section 5.2.4.

4.3.3 Results

To remain consistent, non-parametric statistics were again used throughout this analysis with Wilcoxon related samples non-parametric tests used to compare differences between the two similarity metrics.

Effectiveness between similarity metrics

Participants again made generally few errors. There were five attacks in each set of verifications, and so the number of attack errors could range from 0 to 5. Of the 205 total attack verifications within the orthographical similarity condition, only 12 errors were made (5.9%). Only 16 errors were made within the phonological similarity condition (7.8%). Figure 4.2 shows the distribution of attack and non-attack errors within each condition.

	Orthographical	**Phonological**	**Wilcoxon W**	**p-value**
Attack verifications	0 (0.0)	0 (0.0)	10.5	0.2714

Table 4.6: Median errors on correct verifications and SIQR for the orthographical and phonological similarity metrics with Wilcoxon Signed Rank tests of differences between conditions.

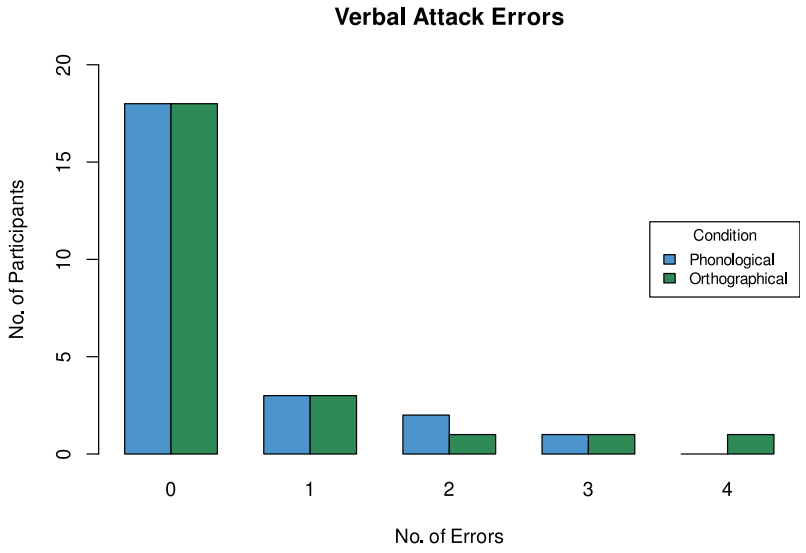


Figure 4.2: Number of errors by each participant on 5 attack verifications

No significant difference was observed in the number of errors made between the two similarity conditions, as tested by Wilcoxon signed-rank tests for related samples (see Table 4.6). Thus H_5, that there will be a difference between in number of attack errors on the two conditions was not supported.

4.3.4 Discussion and Evaluation

This study investigated differences in the number of attack errors that participants make between attacks which possess high orthographical similarity and those that possess high phonological similarity. The results did not identify a significant relationship, which indicates that this is not a significant factor in why users make errors upon word based fingerprint verifications.

The unequal split between recruitment sources was not intentional. The intention had been to recruit all participants using MTurk, but the requirement to repeat the study with a fresh sample severely reduced the pool of available participants. Consequently the study was also shared with the author's personal networks to recruit additional participants. This also proved difficult and after two weeks the decision was made to halt the study and investigate the data for any significant results. If any interesting effects were identified, then the study may have then been repeated using a different platform. The recruitment challenges encountered during this study motivated transition to Prolific during later studies, as discussed within Section 5.2.

4.4 Conclusions

This chapter aimed to develop an improved experimental methodology that is able to resolve the design limitations that were identified during a post–study evaluation described in Section 3.5.

On a micro level, these aims have been accomplished, with the creation of the custom IPP-AV scale enabling assessment of a users auditory–visual information processing preference, and implementation of a simulated brute force effort which has identified attack fingerprints with significantly greater similarity than those available within public PGP key servers.

Though these enhancements have been shown to be effective within isolation it remains to investigate this improved methodology as a whole. The focus of the thesis now shifts from development of the optimal experimental methodology to thorough investigation of the intended research questions. This is the focus of the following two chapters, where the improved experimental methodology is implemented to assess user performance and perceived usability using first a word–based representation followed by a second study that investigates if similar effects are also observed if the fingerprint utilised a numerical representation.

Chapter 5

Word-based Verification Modes and the Effect of Information Processing Preference

5.1 Introduction

Though the study described in Chapter 3 identified a significant usability difference between the visual and verbal verification modes, it represented an exploratory investigation of the phenomena and was unable to determine concrete conclusions to the underlying research questions. Subsequent evaluation identified a number of limitations within the study's design, which motivated the work described in Chapter 4. The results of that chapter provided the following improvements to the overall experimental design:

- Inclusion of attack fingerprints with increased similarity, based on the capabilities of a well resourced and highly motivated attacker. Paired with an increase in the number of attacks, it was predicted that this may enable observation of an interesting difference in effectiveness in the attack condition.
- Inclusion of the custom IPP-AV scale to measure the impact of a participant's information processing preference upon their performance. In Section 4.1.2 participant scores on the IPP-AV were found to align with their pre-existing information processing preference. Thus, it was predicted that if a participant's information processing preference was a significant factor, then this measure should be able to identify it.

This chapter describes the results of an additional within-participants study with 52 participants which investigated differences in security and usability between visual and verbal verification of word-based key fingerprints, and the related impact of a user's own information processing preference. The study implemented the design improvements outlined above, and sought to develop definitive answers to the underlying research questions in the context of a word-based fingerprint. Analysis of the data identified significant differences in the effectiveness and efficiency of the two verification modes, with the visual verification mode found to be more usable and effective. Furthermore, the visual mode was also found to be more usable on three out out six perceived usability dimensions, though two of these were not shared with the findings of the exploratory study. However,

the impact of a participant's information processing preference upon their performance was again found to not be significant.

5.2 Method

The study investigated the following hypotheses:

H_1 There is a significant difference in the number of errors made using the visual and verbal fingerprint verifications.

H_2 There is a significant difference in time to make the correct decision between the visual and verbal fingerprint verifications.

H_3 There is a significant difference in perceived usability ratings between the visual and verbal fingerprint verifications.

H_4 Participants perform significantly better when the verification mode aligns with their preferred method to receive and process information.

Ethical principles of no harm, informed consent and data protection were followed and formal ethical approval was obtained from the author's departmental ethics committee.

The design of this study largely replicated that described in Section 3.2, but incorporated a range of modifications to resolve the design limitations identified in Section 3.4. In addition, these changes enabled investigation of wider aspects of the underlying phenomena. The remainder of this section will describe all differences in study design and includes a justification for each change. For clarity, any aspects of the previous method that are not discussed remain unchanged.

5.2.1 Design

Attack verification rate

A major design flaw of the exploratory study was that the inclusion of a low number of attacks introduced a floor effect which would have made identification of any interesting difference within the attack condition difficult to identify. Consequently, the following changes were made for this study:

- Participants compared 25 pairs of key fingerprints in each condition instead of 20.
- Five verifications of each set represented simulated attacks, instead of only two.
- Three represented full mis-matching attention checks (see Section 5.2.1).
- The remaining 17 verifications simulated identical non-attack verifications.

A side effect of these changes was that the attack verification rate doubled to 20%. This works against the initial intention to include a low-attack rate to prevent participants from becoming accustomed to being attacked, but was deemed to be a necessary trade off when compared with the potential benefits. Moreover, even the 10% attack rate used in the exploratory study is likely to be far greater than that a user would encounter in practise, highlighting an underlying challenge to the ecological validity of the attack related investigations of this research. This is further discussed in Chapter 7.

Inclusion of attention checks

Whilst evaluating the data of the exploratory study, a potential confound was identified related to the level of attention that participants provided whilst they completed their verifications. The exploratory study included only a single fully mismatching attention check at the start of the study, but there existed the potential for some participants to become either distracted or fatigued during later verifications. The following changes attempted to identify such behaviour:

- Each set of verifications included three distinct attention checks in positions 3, 13, and 24.
- The additional two attention checks aimed to identify participants who became distracted or fatigued either during or towards the end of their verification set.

Set of attack fingerprints

The exploratory study utilised an attack set constructed from key fingerprints that were freely available in public key servers. The initial concept of similarity sought to identify pairs of fingerprints within the collected set with the highest number of identical words. The resulting attack set consisted of fingerprint pairs with identical first, second and fifth words, but with no control of the difference in the third and fourth words. Thus, it was possible that attack pairs could still include word differences that participants could easily identify, which is a potential explanation for the low number of attack errors observed in the exploratory study.

Thus, a simulated computation was implemented to identify attack fingerprints with greater similarity. It was hypothesised that such fingerprints would be much harder to identify from a casual verification and may cause participants to make more errors, whilst remaining representative of examples that users may encounter in real attacks. A detailed description of the method to generate these attacks is provided in Section 4.2.

All attacks use a phonological measure of similarity. This was chosen as the study described in Section 4.3 did not identify a significant difference between orthographical and phonological measures of similarity, and theory suggests that users may prefer to process data phonologically [4].

Recruitment source

The exploratory study implemented a dual recruitment strategy, with participants recruited through Amazon's Mechanical Turk and the author's personal networks. In contrast, all participants for this study were recruited via Prolific[1]. Prolific is intended to allow users to participate in research, in contrast to MTurk which has been designed as a work portal with the intention of spreading out work and allowing people to make money. This is a subtle but important difference, and motivated the transition to use of Prolific for this and all subsequent studies.

Furthermore, recent research has found Prolific to produce higher quality data, based upon measures of attention, comprehension, honesty, and reliability, when compared with other online behavioral research platforms, including MTurk [47]. It is important to

[1] `www.prolific.com`

note that of the five authors of this paper, three were employed by Prolific. This could raise some questions regarding potential conflicts of interest, but these associations have been declared within the paper and the authors have made all materials and data freely available via the Open Science Framework. The paper is also published in a well respected journal which instills confidence that this research comes from a reputable source.

5.2.2 Materials and Task

Web Application

The same web application used in the exploratory study was used again. Changes were made to the back end to facilitate the improved design discussed in Section 5.2.1.

An additional change was also made to the application's front-end to enable the fingerprints to be displayed using a vertical alignment rather than upon a single row (see Figures 5.1 and 5.2). This change was implemented to reduce uncontrolled variance within the visual condition, as it was observed that some longer fingerprints strings would introduce a line break within the text message display (see Figure 3.1). If left unchanged, this could have potentially caused users to make a different type of errors unrelated to any challenges with comparing individual words.

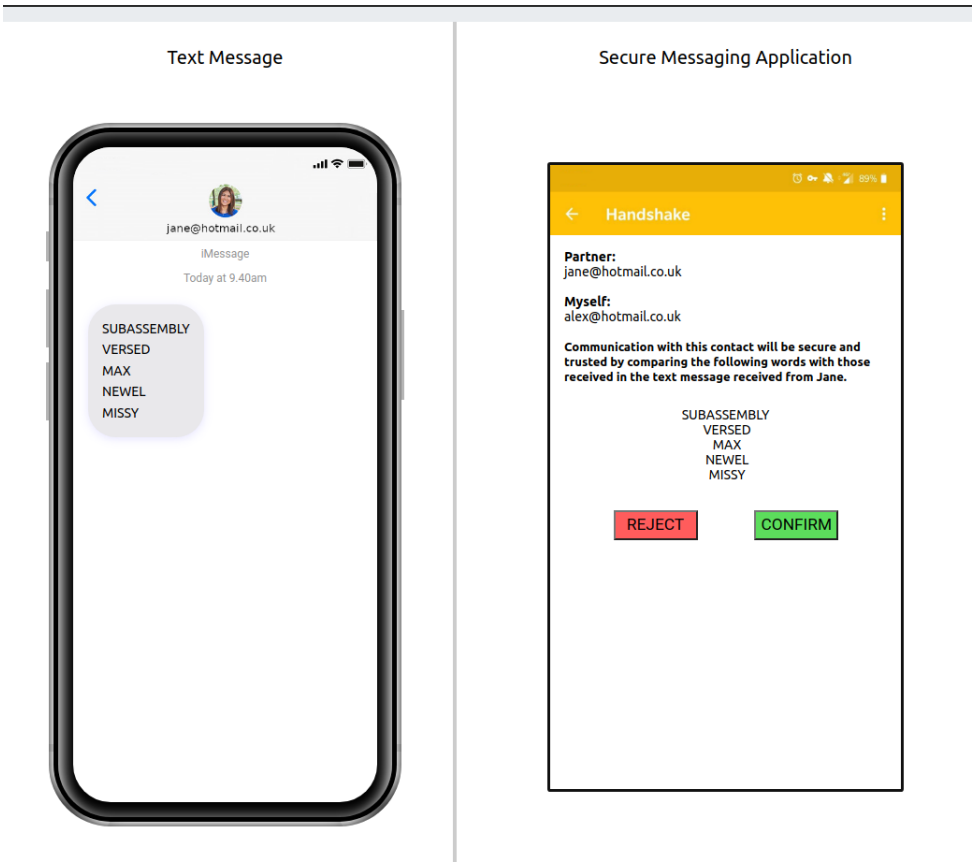


Figure 5.1: Visual verification task interface.

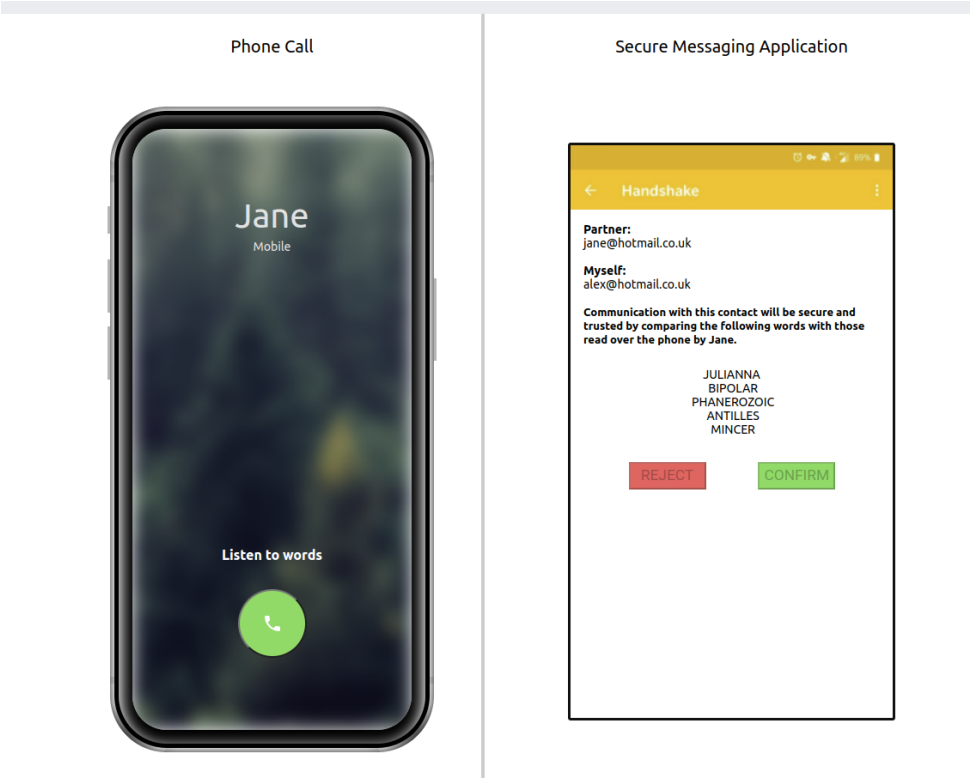


Figure 5.2: Verbal verification task interface.

Classification of a participants information processing preference

The custom IPP-AV scale was developed to measure a participant's information processing preference (see Section 4.1.2). The scale consists of seven 7-level Likert items (with options from "Strongly Disagree" to "Strongly Agree"), so scores could vary between 7–49 with high scores corresponding to an auditory preference to receive information. A principal components analysis found that the scale produced a one factor solution which accounted for 51.4% of the variance.

Post-task and post-study questionnaires

To gain further insight into the perceived usability of the task, Likert items used within the post-task and post-study questionnaires used a 7-level rating scale instead of a 5-level scale. A 7-level scale (with options ranging from "Strongly Disagree" to "Strongly Agree") provides more variation in responses, which in turn makes it easier to observe an effect if one exists.

5.2.3 Procedure

Prior to running the main study, a pilot study was conducted with three participants recruited from within the author's research group. This identified the front-end improvement identified above in Section 5.2.2. The main study procedure was almost identical to that of the exploratory study described in Section 3.2.4, with the only change being the

use of the custom IPP-AV scale instead of the Visual-Verbal subscale of the ILS in step 3.

5.2.4 Participants

All participants were recruited using the Prolific platform and were paid £5.00 upon successful completion of the study. This reward value was specified by Prolific, and represented an increase on that that of the exploratory study. As discussed above, researchers have found prolific to produce higher quality data, which was a key motivation to the transition. Furthermore, personal experience found that Prolific was able to recruit participants much faster than either MTurk or the author's personal networks.

In total, 90 people responded to the study, but data from 38 participants was eliminated: 2 self reported sensory disabilities which may have affected their performance and 36 provided only partial responses. It is believed that the high attrition rate was caused by the custom web application failing to handle the traffic generated by the participant interactions, with some participants reporting that it either crashed or failed to load the second set of verifications. As in the previous studies, all participants whose data was excluded were still rewarded for their time.

Data from 52 participants were analysed, 28 men (54%) and 24 women (46%). Participants' ages ranged from 18–24 to over 65, with the largest group being the 25–34 years range (29%, see Table 5.1). The age distribution largely correlates to that of the UK population, apart from a noticeable lack of participants aged 65 and over [22]. This age group likely requires further attention as they likely face their own challenges (e.g deteriorating eyesight and hearing or lack of confidence using technology). Educational level ranged from high school education to postgraduate degree, with the majority having a bachelors or postgraduate degree (65%, see Table 5.2). This is almost twice that of the UK population, with around 33.8% of people reporting that they had achieved at least a Level 4 qualification during the 2021 census [65]. As the experimental task again involved reading and listening, participants were asked about their proficiency in English; all rated it as good or excellent. There was little variation in participants location, with 49 located in the UK and the remaining three answering 'Prefer not to say'.

Participants were found to display good levels of attention during the study. In addition to relatively few errors overall, no participants failed either of the first two attention checks. However, a large proportion of participants failed the final attention check, with 69.2% (36/52) observed to fail the final verbal check and 69.2% (36/52) found to fail the final visual check. Rather than this being an indication of poor quality data, it instead ap-

Age	**Count**
18–24	12
25–34	15
35–44	10
45–64	12
65 and over	3
Prefer not to say	0

Table 5.1: Age distribution.

Highest Education level	**Count**
High School education	14
Vocational training	4
Bachelors degree	23
Postgraduate degree	11
Other	0
Prefer not to say	0

Table 5.2: Education background.

pears to be an indication that participants became fatigued towards the end of the study. This is not ideal but models human behaviour, and was a necessary trade off to enable a low attack rate (see Chapter 7). As the order in taking verifications was counterbalanced, the fatigue will be distributed across the data which helps to mitigate its impact.

Before beginning the study, participants were again asked about their previous usage of and attitudes towards secure messaging applications. The responses showed 75% (39/52) use a secure messaging application, and 30.7% (16/52) do so every day. These results were much lower than anticipated, particularly when compared with the same responses collected during the exploratory study. In addition, 80.8% (42/52) of participants agree that 'it is important to be able to have private conversations using secure messaging applications', yet 86.5% (40/52) of participants have never performed a fingerprint verification. These responses were comparable with those observed in the exploratory study.

5.3 Results

To remain consistent and enable comparisons with results of the previous studies, non-parametric statistics were again used throughout the following analysis. Medians and semi-interquartile ranges (SIQR) are reported as measures of central tendency and spread, with Wilcoxon related samples non-parametric tests used to compare between conditions. A correlation analysis compared the impact of a participants information processing preference upon their performance.

5.3.1 Performance

Effectiveness

The study included five attack verifications within each condition, and so the number of attack errors could range between 0 and 5. There were also 17 non-attack verifications, and so the number of non-attack errors could range from 0 to 17. Participants again generally made few errors, with 73 errors made across all verbal verifications (5.62%) and only six errors on all visual verifications (0.46%). Figures 5.3 and 5.4 show the distribution of errors in the attack and non-attack conditions respectively.

	Verbal	**Visual**	**Wilcoxon W**	**p-value**
Attack verifications	0 (0.0)	0 (0.0)	13.0	0.0206
Non-attack verifications	1 (1.0)	0 (0.0)	7.0	< 0.01

Table 5.3: Median errors on correct verifications and SIQR for verbal and visual verification conditions with Wilcoxon related samples tests of differences between conditions.

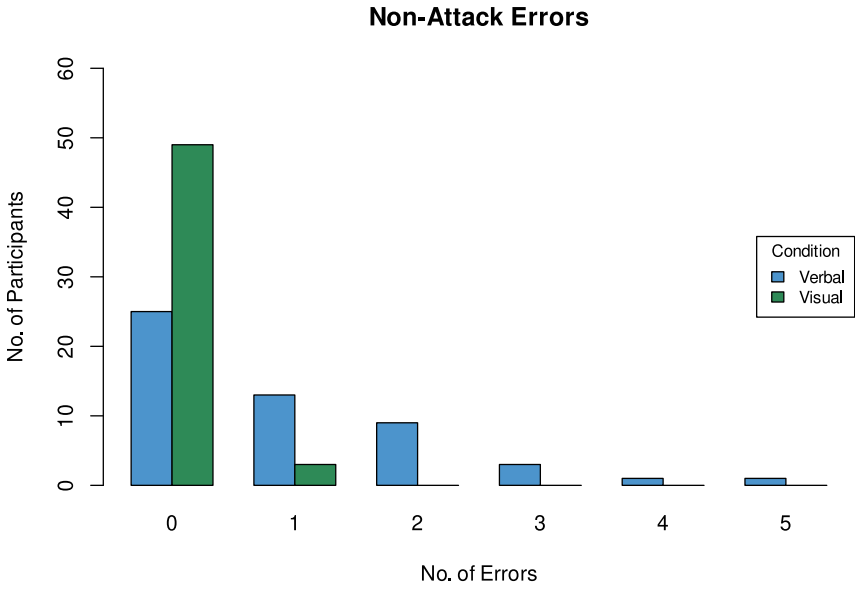


Figure 5.3: Number of errors by each participant on 17 non-attack verifications.

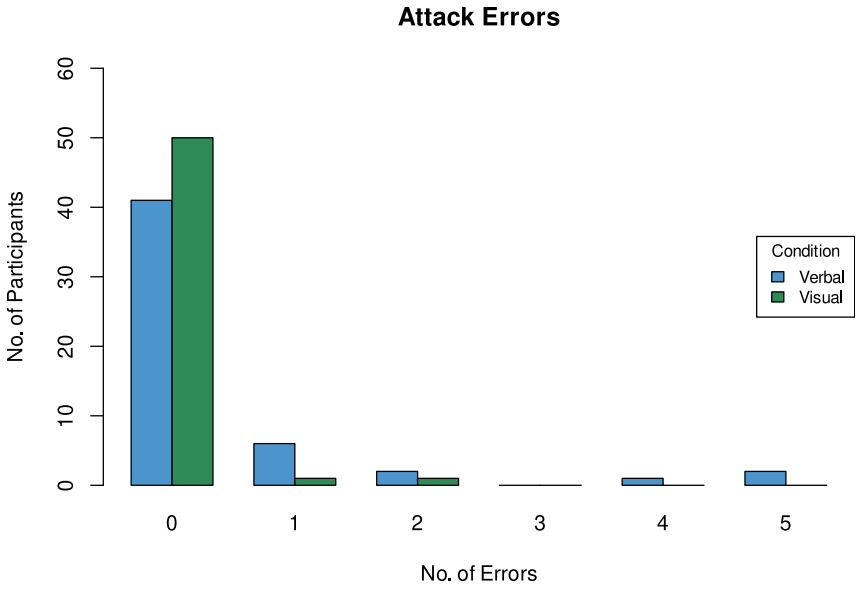


Figure 5.4: Number of errors by each participant on 5 attack verifications.

There was a significant difference in errors between the two conditions, with participants found to make more errors when using the verbal verification mode for both the attack and non-attack conditions. Table 5.3 reports the results of Wilcoxon related samples non-parametric tests for each case. Thus H_1, that there will be a difference between the errors on the two conditions, was supported.

Efficiency

The difference in time to complete correct verifications between the two conditions was also found to be significant for both the attack and non-attack conditions with participants found to be significantly faster in making the correct decision when using the visual verification mode. Figures 5.5 and 5.6 show the timing distribution for both the attack and non-attack verifications, with results of the Wilcoxon signed-rank tests for related samples provided in Table 5.4. Thus H_2, that there will be a difference in average correct verification time between the two modes, was supported.

	Verbal	**Visual**	**Wilcoxon W**	**p-value**
Attack verifications	10.61 (2.79)	8.22 (4.09)	229.0	< 0.01
Non-attack verifications	11.66 (1.81)	10.40 (2.79)	406.0	< 0.01

Table 5.4: Median times (seconds) and SIQR on correct verifications for verbal and visual conditions with Wilcoxon signed rank tests of differences between conditions.

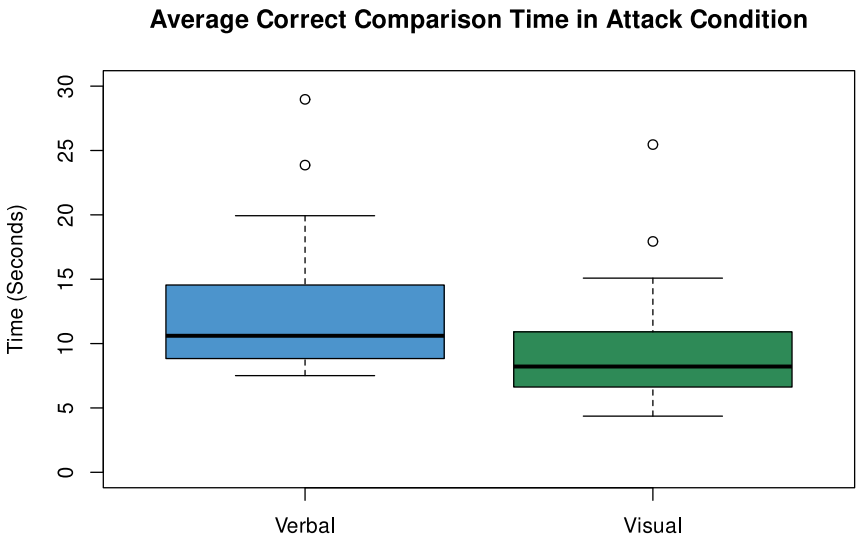


Figure 5.5: Distribution of mean correct attack verification times by condition.

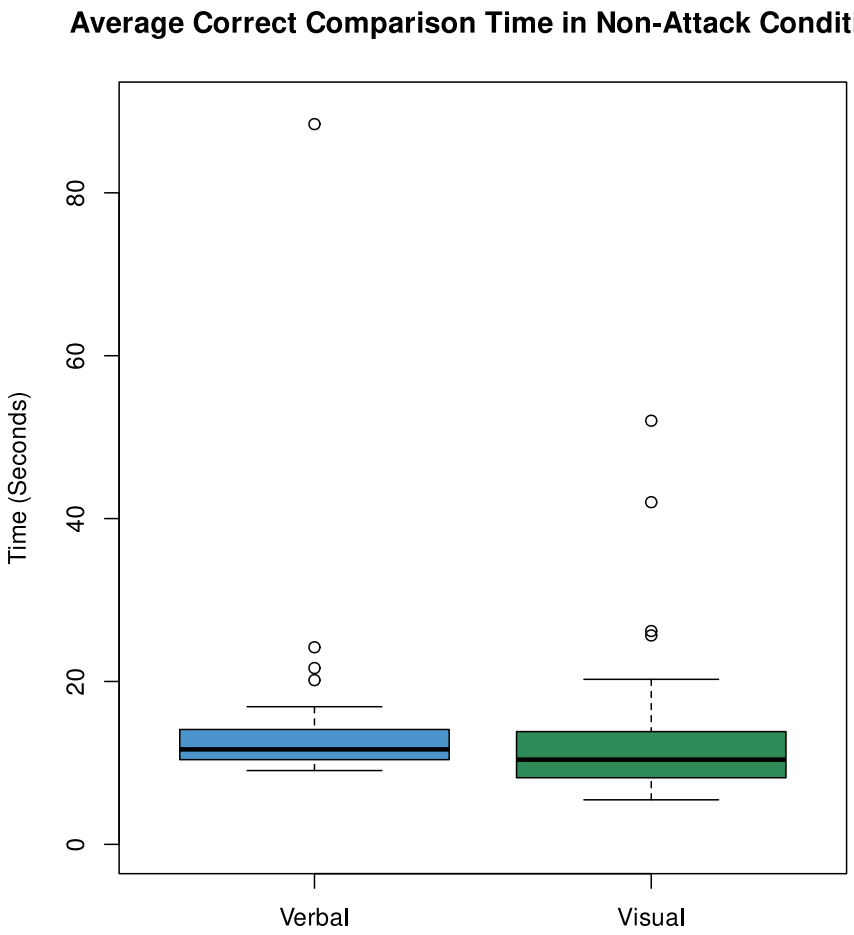


Figure 5.6: Distribution of the mean correct non-attack verification times by condition.

5.3.2 Perceived Usability and Related Concepts

Perceived usability was assessed using the same questions described in Section 3.2.1, with all answers reported using the 7-level Likert items. Table 5.5 shows participants median ratings for each dimension on both the verbal and visual condition. There was a significant difference upon three of the six dimensions. Participants reported that they would be more likely to re-use the visual verification mode ($p = 0.057$), and perceived it to instill greater trust ($p < 0.01$) and confidence ($p = 0.041$). Furthermore, at the end of the study participants were asked to report their preferred verification mode. A clear majority preferred the visual verification mode, with 63.4% (33/52) reporting it as their favourite. Though this preference was not significant, there was a strong trend towards significance ($\chi^2 = 3.76$, $p = 0.052$). These results provide partial support for H_3, that there is a difference in the perceived usability of the two conditions, but this may be a complex relationship (see Section 5.4).

Dimension	Verbal	Visual	Wilcoxon W	p-value
Efficiency	5.00 (1.00)	5.00 (1.00)	582.5	0.96
Ease of use	5.75 (1.0)	5.75 (0.81)	460.5	0.68
Low mental workload	4.00 (1.25)	3.00 (1.06)	501.5	0.37
Confidence	5.50 (0.88)	6.00 (0.75)	259.0	0.041
Repeat Use	4.50 (1.00)	5.00 (1.00)	367.5	0.057
Trust	4.50 (1.0)	5.25 (1.00)	188.5	<0.01

Table 5.5: Median ratings (with SIQR) of the perceived usability dimensions for verbal and visual conditions and Wilcoxon Signed Rank tests of differences between conditions

5.3.3 Effect of Information Processing Style: Auditory vs Visual

Participants scores on the custom IPP-AV scale were distributed across the full range, with a slight skew towards participants with a visual preference (see Figure 5.7). A correlation analysis was performed to investigate if there were any significant effects between a participants preferred method to process new information and their performance when comparing word-based key fingerprints.

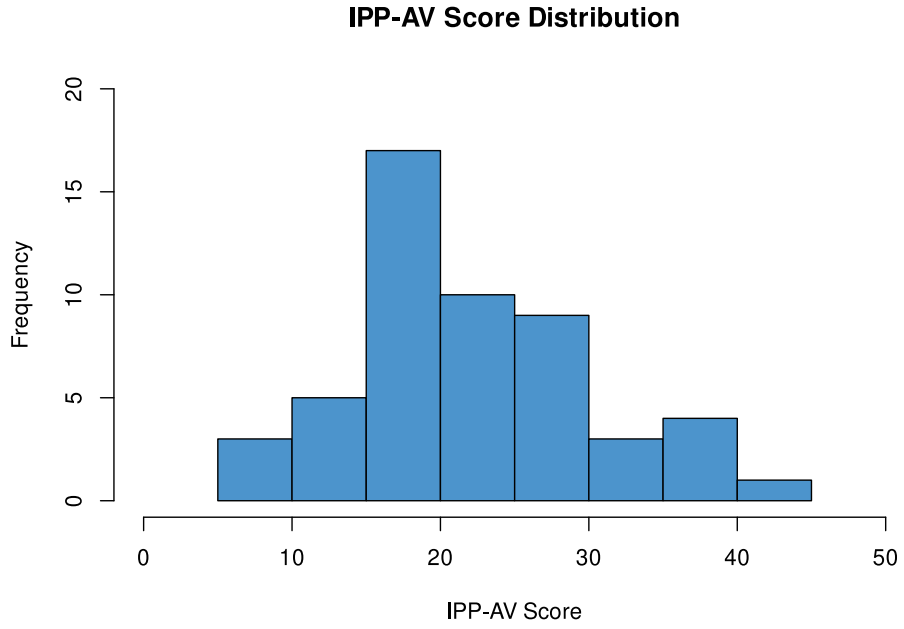


Figure 5.7: Participants scores from the 7 question IPP-AV scale.

Effectiveness

There was no significant correlation between participants scores on the custom IPP-AV scale and the number of errors that they made (see Figures 5.8 and 5.9). Table 5.6 reports the Spearman's rank correlation coefficients for both attack and non-attack verifications using each mode.

	Verbal	Visual
Attack verifications	0.169	-0.176
Non-attack verifications	-0.095	-0.195

Table 5.6: The Spearman's rank correlation coefficients for the error data.

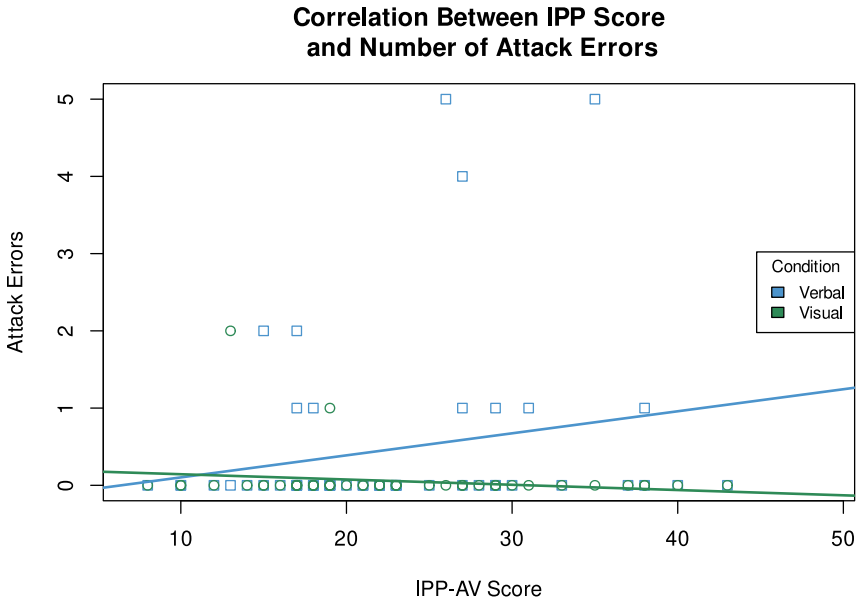


Figure 5.8: Distribution of number of attack errors against IPP-AV score.

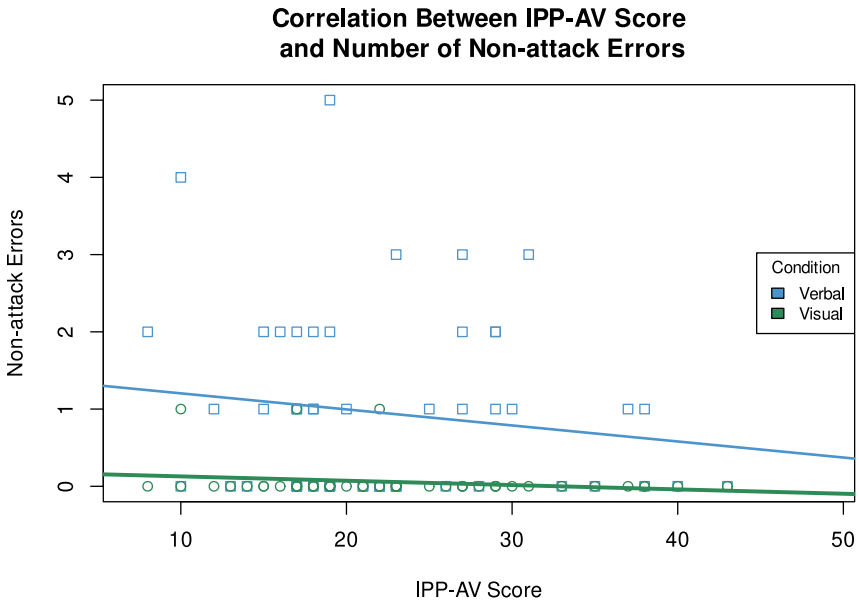


Figure 5.9: Distribution of number of non-attack errors against IPP-AV score.

Efficiency

There was also no significant correlation between IPP-AV score and the time to make the correct verification (see Figures 5.10 and 5.11). Table 5.7 reports the Spearman's rank correlation coefficients for both attack and non-attack verifications in each verification mode. Thus, these results do not support H_4, that their would be a difference in performance when the verification mode is aligned to a participants preference to receive information.

	Verbal	**Visual**
Attack verifications	-0.206	-0.119
Non-attack verifications	-0.191	-0.042

Table 5.7: The Spearman's rank correlation coefficients for timing data.

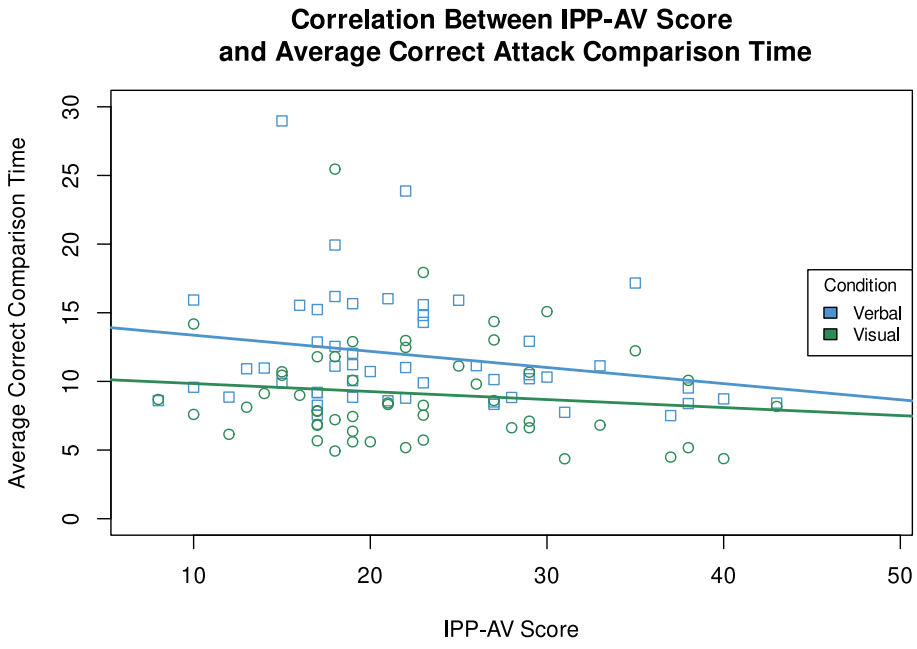


Figure 5.10: Distribution of average correct attack verification time against IPP-AV score.

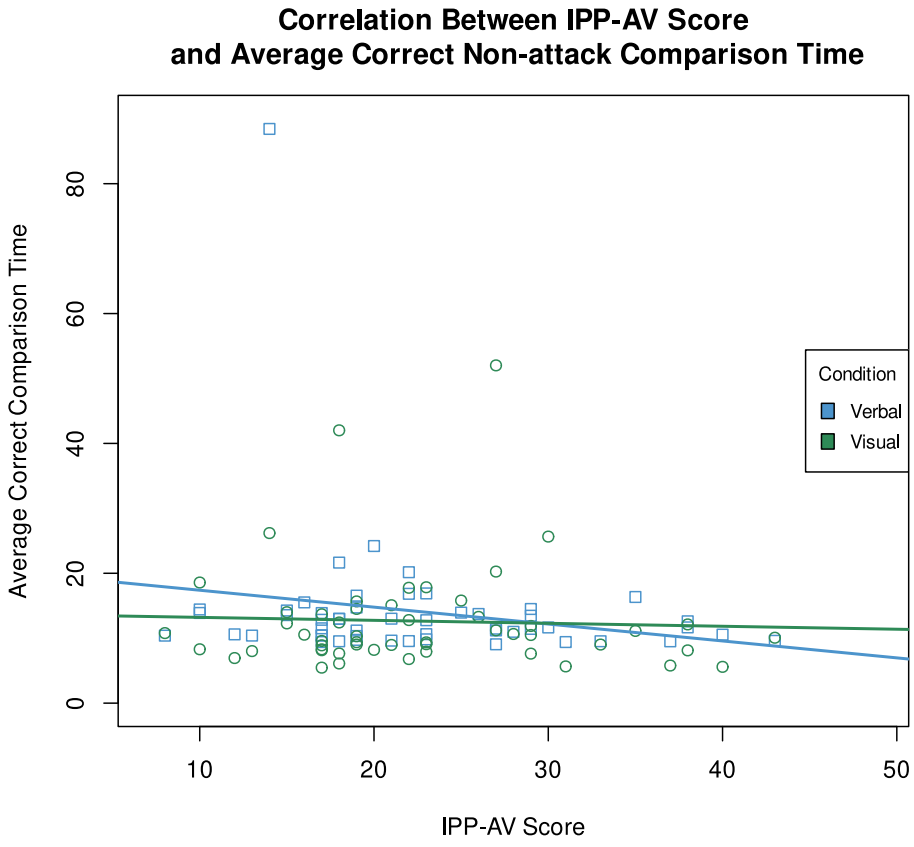


Figure 5.11: Distribution of average correct non-attack verification time against IPP-AV score.

5.4 Discussion

This chapter reported the results of a further investigation of the security and perceived usability of word based key fingerprints, and a first investigation of the use of the custom IPP-AV scale described in Chapter 4 to measure the impact of a participants perceived preference to receive information.

Participants made more errors when using the verbal verification mode, in both the attack and non-attack conditions. The observed non-attack effect was also observed in the exploratory study described in Chapter 3, and indicates that this may be a robust effect. It is interesting that this effect was found to also occur in the attack condition, which was not the case in the exploratory study. This indicates that including attack fingerprints with increased similarity, based on the work described in Section 4.2, was successful in causing participants to make more security errors. This is an important finding as it demonstrates that these are not only hypothetical attacks, but could find practical significance if an attacker is able to generate a malicious fingerprint with sufficient similarity.

There was also a significant difference in the time taken to make a correct verification between the two conditions, with verifications found to be 2.39 and 1.26 seconds faster in the attack and non-attack conditions (resp.) when using the visual verification mode.

This was a surprising result, as the effect was not observed within the exploratory study and it was not predicted that the experimental manipulation would lead to an increase in time to make verbal verifications. For example participants were not required to click any extra buttons within the application. A potential explanation is that inclusion of attacks with increased similarity had a greater impact upon the average verification time for verbal verifications, as it was harder to identify differences between the words when they were read out.

The visual verification mode was again found to instill greater confidence, a result that was previously seen in the exploratory study and may indicate that this is a robust effect. Participants also reported that they would be more likely to re-use the visual verification mode and had more trust in its usage. However, the latter effects were not observed in the exploratory study. There was also no significant difference upon the "Low Mental workload" or "Ease of use" dimensions in this study, unlike the earlier study. It is worth noting that in this study the questions were answered using 7-level Likert items, in contrast with 5-level items in the exploratory study. This makes a direct verification of the results between the two studies difficult.

These results indicate that perceived usability is a complex phenomena to assess. Some potential explanations for the differences observed in this study are:

- This study was more challenging for participants. It included more attacks which were also much possessed much more similarity.
- Participants were recruited from a single source and almost all were located within the UK. This is in contrast to the exploratory study which used two distinct recruitment sources and a roughly even split between those located in the USA and UK. It is feasible that cultural differences could have impacted the perceived usability of the verifications, with one example being the inclusion of a text-to-speech application that utilised an English accent which some American participants found to be annoying.

However, it may be the case that participants perceptions of usability vary based upon the complexity of task or based on the underlying characteristics of the sample, such as demographics. Or it may be that the selected questions are not a durable measure. Further investigation with similar samples is required to determine if these questions can be used to determine a generalised assessment of the perceived usability of a fingerprint verification.

No significant differences were identified between a participant's auditory-visual information processing preference and their performance. It may be that the custom IPP-AV scale still does not accurately measure the underlying phenomena of interest, or the effect of verification mode may dominate any potential effect.

5.5 Evaluation and Conclusions

The results of this study provide interesting insight into the differences between visual and verbal verifications, and provide support for both H_1 and H_2. However, it is not yet possible to definitively conclude that these effects are a fundamental property of a fingerprint verification task, as the data only supports this conclusion for word-based fingerprints.

An alternative explanation of these results is that the cause of errors stems from an unfamiliarity with the words included within the fingerprint string. All fingerprints of this study are encoded using the Pretty Easy privacy (PEP) representation which included a large number of unfamiliar words. Consequently, it is possible that some participants using the verbal mode may reject fingerprints that include unfamiliar words due to a belief that the word displayed upon their device sounds different to that which has been read out to them. If true, this would explain the increase number of non-attack errors observed when participants use the verbal mode, and conflict with the conclusion that the difference is caused by inherent differences between the two verification modes. Hence, Chapter 6 will attempt to answer this question by determining if the same usability effect is observed when verifications use a numerical representation instead.

Chapter 6

Numerical Verification Modes and the Effect of Information Processing Preference

6.1 Introduction

This chapter reports the results of a within participants study with 51 participants which investigated the differences in security and usability between visual and verbal verifications of numerical key fingerprints. The study aimed to assess if the previously identified effects of this research extended to non-word based representations, which would provide support for the claim that the previously identified differences relate to inherent properties of the two modes rather than the chosen fingerprint representation. A numerical representation was chosen as they have been shown to possess good levels of security and usability within previous work [16,58], and as they are widely used in modern messaging applications (e.g. WhatsApp).

In addition, the study again recorded data about the perceived usability of the two modes. Though the earlier studies of this thesis had found the visual mode to be perceived to be more usable, the results did not always describe a clear picture with different usability dimension achieving significance. Since the design and sample shared many similarities to the study described in Chapter 5, it was of interest to see if the same usability effects could be identified within the data of this study. The study also investigated the impact of a participants preferred method to receive information upon their performance. Though the previous studies were yet to identify a significant result related to this research question, it remained possible that this may be observed within this study.

An analysis of the data again identified a significant difference in effectiveness within the non-attack condition, with participants again found to make less usability errors. This supports the assessment that this is a fundamental difference between visual and verbal verification of key fingerprints. However, a similar effect was not observed within the attack condition. This indicates that the alternative explanation, that the previously observed effect is a property of the underlying word base, may hold true for attack verifications.

The visual verification mode was again perceived to be more usable on two out of six usability dimensions, but these again failed to fully agree with the results observed

in earlier chapters. It is also possible that the experimental manipulation had a greater effect upon the usability of the verbal verification mode, which may have exaggerated the underlying effect. The impact of a users preferred method to receive information was again found to not be significant.

6.2 Method

This study investigated the following hypotheses:

H_1 There is a significant difference in the number of errors made using the visual and verbal fingerprint verifications.

H_2 There is a significant difference in time to make the correct decision between the visual and verbal fingerprint verifications.

H_3 There is a significant difference in perceived usability ratings between the visual and verbal fingerprint verifications.

H_4 Participants perform significantly better when the verification mode aligns with their preferred method to receive and process information.

Ethical principles of no harm, informed consent and data protection were followed and formal ethical approval was obtained from the author's departmental ethics committee.

6.2.1 Design

The design of this study was almost identical to that described in Chapter 6. The only modification was the transition to use of a numerical fingerprint to facilitate the intended manipulation of the independent variable.

Care was taken to ensure that the numerical attack fingerprints used in this study were comparable to those of Chapter 5. All attacks were again based upon the simulated pre-computation discussed in Section 4.2, but the underlying hash values were encoded into five chunks of five digits instead of five words (see Table 6.1). In practise this was achieved by terminating PEP fingerprint generation protocol prior to the final trustwords encoding step. The full set of attack fingerprints used within this study can be found in Table D.1.

It is important to note that the hash values that produce highly similar numerical fingerprints may not be the same as those that produced highly similar word fingerprints. Thus, the similarity computation was repeated to identify attack fingerprints with minimal digit edit distance. Repetition of this step is realistic and models the expected behaviour of a real attacker whilst not introducing a bottleneck to the attack identification procedure.

Original fingerprint	11226 25536 43511 59432 44815
Attack fingerprint	11226 25536 42511 55432 44815

Table 6.1: Example of an attack fingerprint pair used within this study.

Numerical similarity was measured using the orthographical edit distance, in contrast to words which used the phonetic distance. The complexity of task is quite different when using a numerical representation. Different digits possess a good phonetic difference, and orthographical challenge are more relevant, for example some users may struggle to differentiate between similar looking digits (e.g between 1 and 7).

6.2.2 Materials and Task

Web Application

The study again utilised the web application introduced in Chapter 3. The only modification was to display all fingerprints using a numerical representation, with an example task for each verification mode shown in Figures 6.1 and 6.2. The source code for the numerical version of the application is available at `https://github.com/lli90/NumbersExperimentApp`.

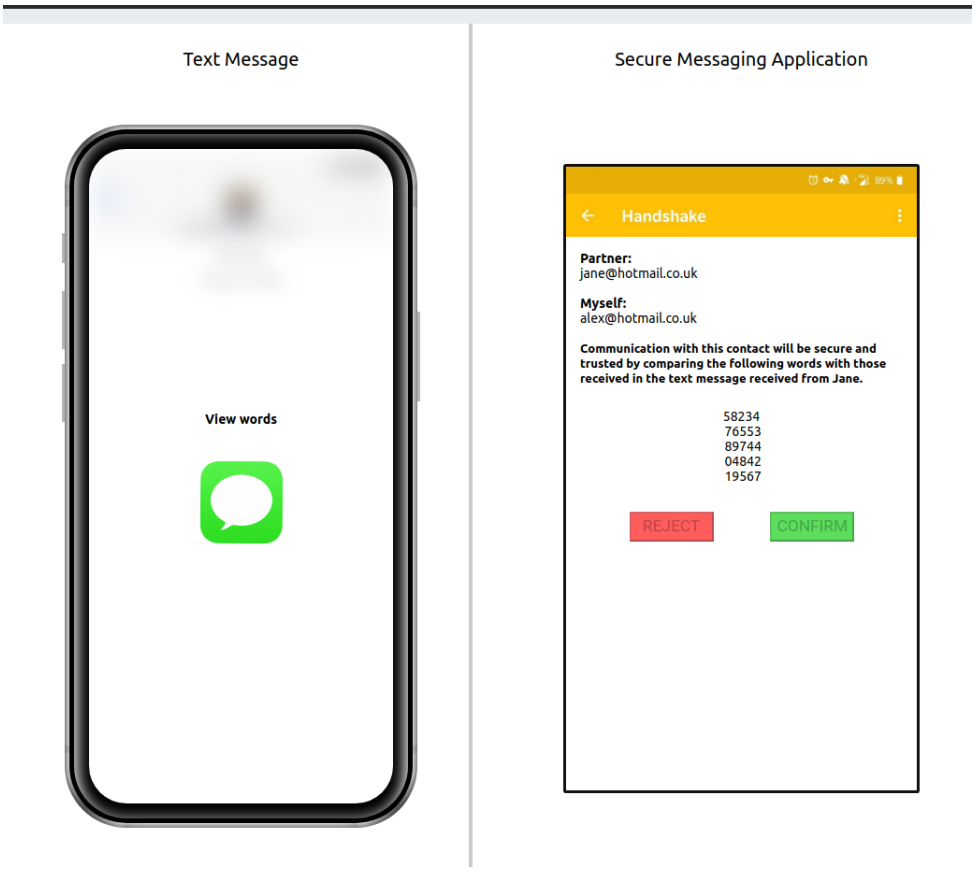


Figure 6.1: Visual verification task interface.

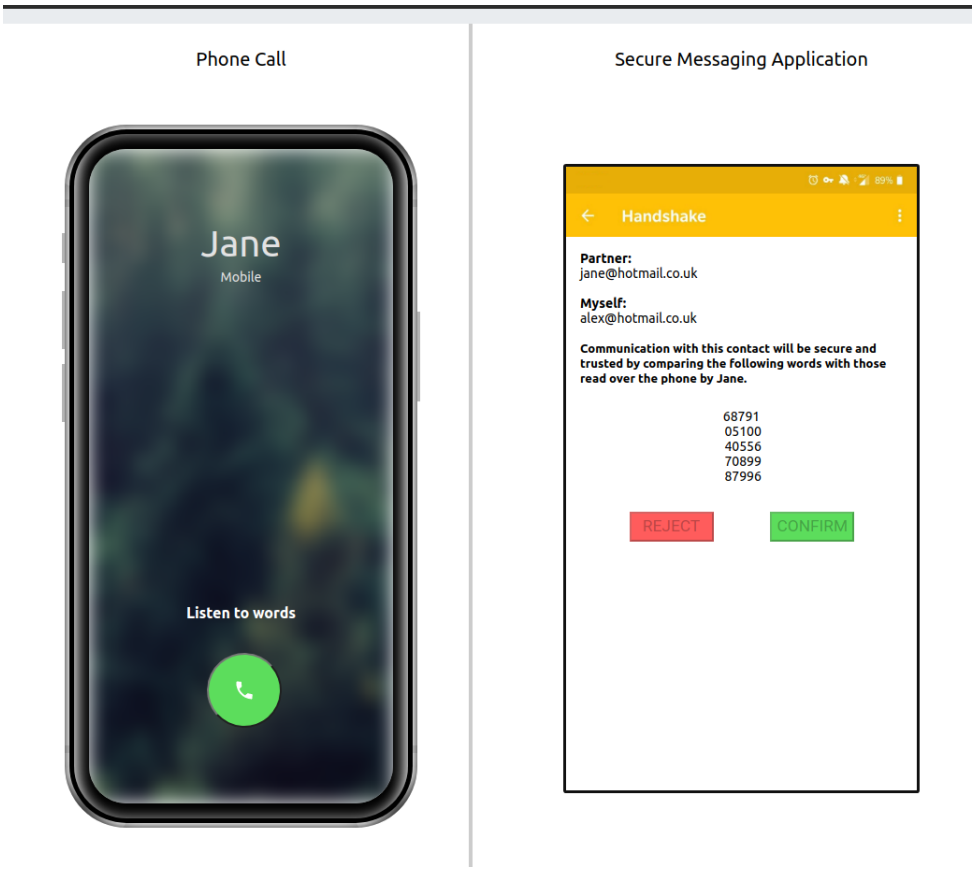


Figure 6.2: Verbal verification task interface.

Post-study questionnaire

Participants were also asked to report if they encounter challenges interacting with numerical information. This intended to identify any participants whose performance may have been impacted by a dissatisfaction with the numerical representation, rather than on the challenges of the actual fingerprint verification.

6.2.3 Procedure

The procedure was identical to that described in Section 5.2.3.

6.2.4 Participants

All participants were again recruited via Prolific and paid £5.00 upon successful completion. In total, 77 people responded to the study, but data from 26 participants were eliminated as they only provided partial responses. Data form 51 participants were analysed, 24 men (47.1%), 26 women (51.0%) and one who identified as non-binary. Participants ages ranged from 18-24 to over 65, with the 25-34 range again largest (29.4%, see Table 6.2). Education level again ranged from high school level to postgraduate degree, with a majority holding a bachelors or postgraduate degree (62.7%, see Table 6.3).

Age	Count
18–24	12
25–34	15
35–44	12
45–64	11
65 and over	1
Prefer not to say	0

Table 6.2: Age distribution.

Highest Education level	Count
High School education	13
Vocational training	5
Bachelors degree	19
Postgraduate degree	13
Other	0
Prefer not to say	1

Table 6.3: Education background.

Almost all participants reported that they were located within the UK, the only differing response was from a single participant who answered "Prefer not to Say". As the experimental task involved interaction with numerical data, participants were asked if they encounter challenges with these types of tasks. None reported any.

It is noteworthy that although the two sets of participants were distinct, they both possessed almost identical demographics. The same observations regarding the correlation with the characteristics of the general population outlined in Section 5.2.4 also apply here, i.e that the sample lacks participants aged over 65 and participants have increased academic attainment compared to the general population.

Three participants reported that they were dyslexic and two have been diagnosed with ADHD. These five participants were not eliminated as these disabilities were not deemed to be incompatible with the experimental manipulation of verification mode. In fact, of these five participants only two made errors which were all made when comparing fingerprints verbally. This indicates that their disability did not directly affect their performance.

The attention checks followed a similar pattern to those of the study described in Chapter 5. No participants failed the first or second attention checks when using either the visual or verbal verification mode. However, 31 participants (60.7%) were observed to fail the final attention check, which again indicates that participants became fatigued during the second half of their verifications.

Before beginning the study, participants were again asked about their previous usage of and attitudes towards secure messaging applications. The responses showed 74.5% (38/51) use a secure messaging application, and 23.7% (12/52) do so every day. Additionally, 78.4% (40/51) of participants agree that 'it is important to be able to have private conversations using secure messaging applications', yet 52.9% (27/51) of participants have never performed a fingerprint verification. These results are roughly comparable with those observed in Chapter 5, apart from an noticeable increase in the proportion of participants who had preciously completed a fingerprint verification task.

6.3 Results

Non-parametric statistics were again used in the following analysis. Medians and semi-interquartile ranges (SIQR) are reported as measures of central tendency and spread, with Wilcoxon related samples non-parametric tests used to compare between conditions. A correlation analysis was performed to investigate the effect of a participants preferred method to receive formation upon their performance.

6.3.1 Performance

Effectiveness

The study included five attack verifications in each condition, and so the number of attack errors could range between 0 and 5. There were also 17 non-attack verifications, and so the number of non-attack errors could range from 0 to 17. Like the previous studies of this thesis, participants made few errors. Across the 1275 verifications for each condition, 55 errors (4.3%) occurred using the verbal verification mode and 13 (1.02%) on the visual verification mode. Figures 6.3 and 6.4 show the distribution of errors in the attack and non-attack conditions respectively.

	Verbal	**Visual**	**Wilcoxon W**	**p-value**
Attack verifications	0 (0.0)	0 (0.0)	15.5	0.7193
Non-attack verifications	1 (0.5)	0 (0.0)	34.5	< 0.01

Table 6.4: Median errors on correct verifications and SIQR for verbal and visual verification conditions with Wilcoxon Signed Rank tests of differences between conditions.

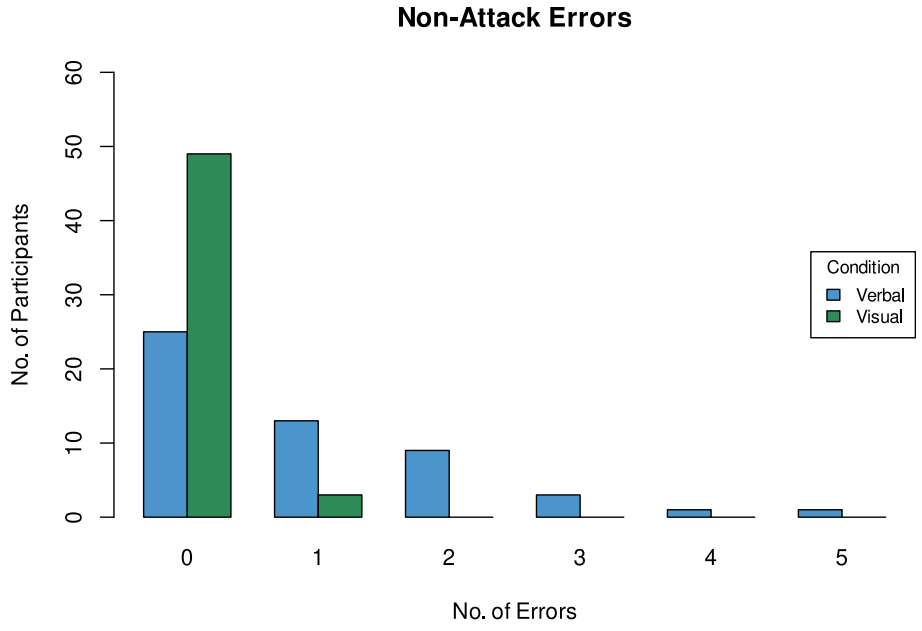


Figure 6.3: Number of errors by each participant on 17 non-attack verifications.

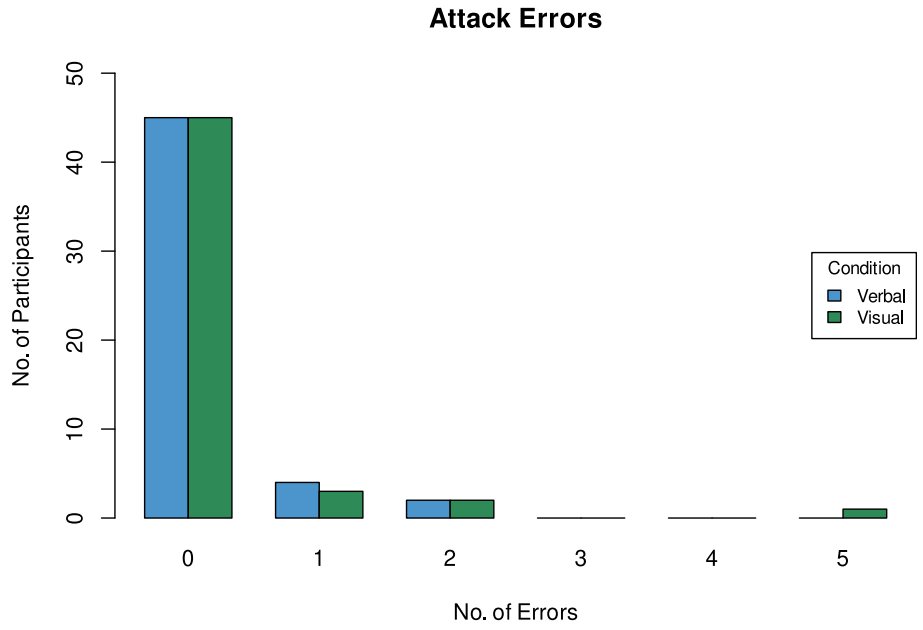


Figure 6.4: Number of errors by each participant on 5 attack verifications.

There was a difference in effectiveness between the two verification modes within the non-attack condition, with participants making significantly more errors when using the verbal mode. However, there was no significant differences between the two modes within the attack condition. Table 6.4 reports the results of Wilcoxon related samples tests for each case. Thus, H_1, that there is a difference in errors between the two conditions was supported, but only for the non-attack condition.

Efficiency

There was a significant difference in time to complete correct verifications between the two verification modes for both the attack and non-attack conditions, with participants significantly faster in making the correct decision using the visual mode. Figures 6.5 and 6.6 show the timing distribution for both the attack and non-attack conditions, with results of the Wilcoxon related samples tests provided in Table 6.5. Thus H_2, that there will be a difference in average correct verification time between the two modes, was supported.

	Verbal	**Visual**	**Wilcoxon W**	**p-value**
Attack verifications	15.84 (2.10)	8.48 (1.64)	122.0	< 0.01
Non-attack verifications	20.66 (1.96)	12.71 (1.82)	53.0	< 0.01

Table 6.5: Median times (seconds) and SIQR on correct verifications for verbal and visual conditions with Wilcoxon signed rank tests of differences between conditions.

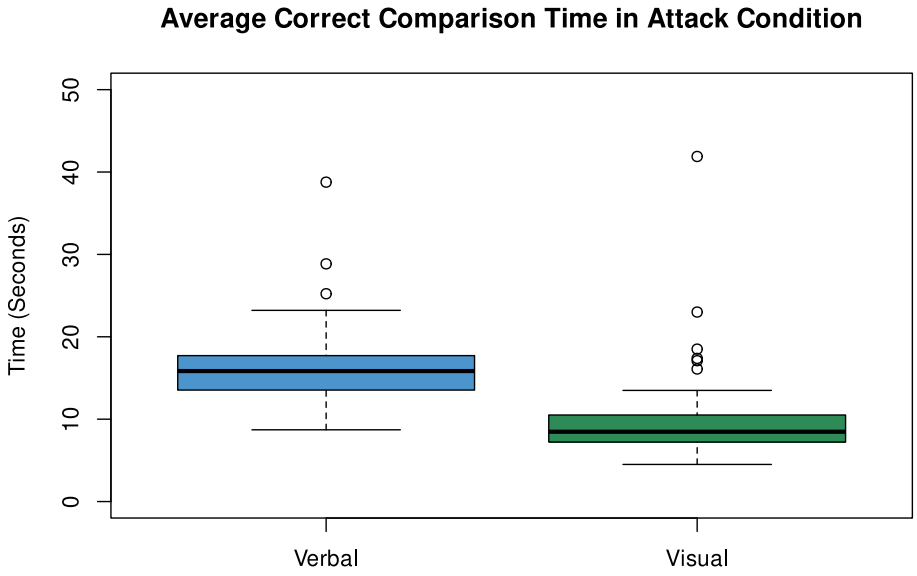


Figure 6.5: Distribution of mean correct attack verification times by condition.

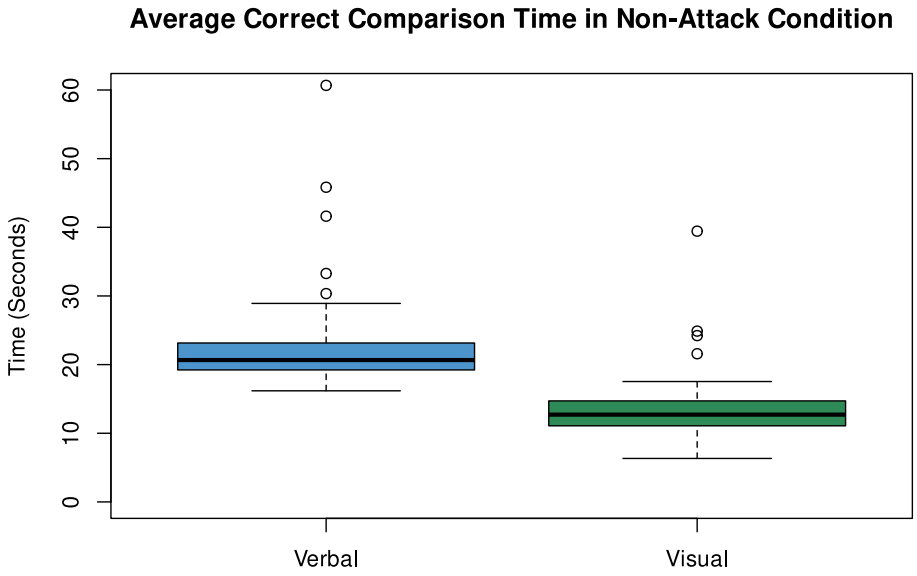


Figure 6.6: Distribution of the mean correct non-attack verification times by condition.

6.3.2 Perceived Usability and Related Concepts

Perceived usability was assessed using the same questions described in Section 3.2.1, with all questions answered using 7-level Likert items. Table 6.6 shows participants median ratings for each dimension on both the verbal and visual verification modes. The results provide partial support for H_3, that there is a difference in perceived usability between the two verification modes, with significant differences identified on two of the six dimensions. Participants perceived the visual verification mode to be more efficient and repeatable. In addition, at the end of the study, participants were again asked which verification mode they would prefer to use, verbal or visual. There was a significant preference ($\chi^2 = 4.41$, p = 0.036), with 64.7% (33/51) of participants preferring the visual verification mode.

Dimension	Verbal	Visual	Wilcoxon W	p-value
Efficiency	4.5 (1.38)	5.50 (1.00)	277.5	0.011
Ease of use	6.0 (0.86)	6.0 (1.00)	322.5	0.660
Low mental workload	4.0 (1.00)	4.0 (1.00)	411.0	0.800
Confidence	6.5 (0.88)	6.0 (0.63)	318.5	0.818
Repeat use	4.5 (1.25)	4.5 (1.0)	296.0	0.020
Trust	5.0 (0.75)	5.0 (0.63)	308.5	0.171

Table 6.6: Median ratings (with SIQR) of the perceived usability dimensions for verbal and visual conditions and Wilcoxon Signed Rank tests of differences between conditions

6.3.3 Effect of Information Processing Style: Auditory vs Visual

Participants scores were again distributed across the full range, but there was a notably large number of participants who reported a strong visual preference (see Figure 6.7). Consequently, a correlation analysis was performed to investigate if there were any significant differences between a participants preferred method to process new information and their performance.

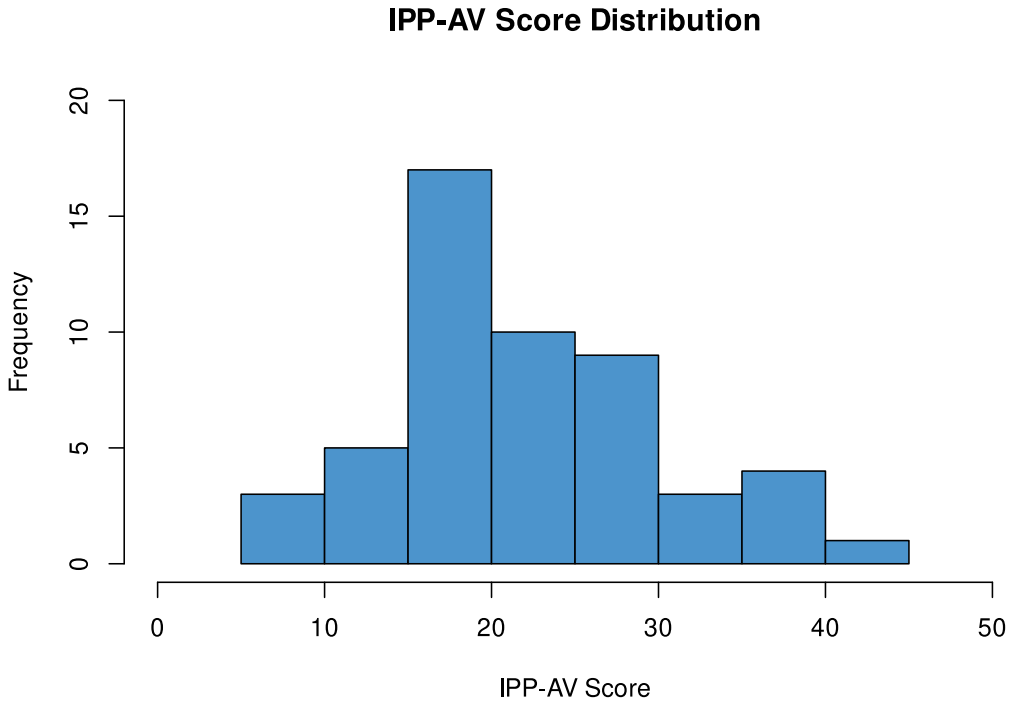


Figure 6.7: Participants scores from the 7 question custom IPP-AV scale.

Effectiveness

Again, no significant correlation was observed between a participants score on the custom IPP-AV scale and the number of errors that they made (see Figures 6.8 and 6.9. Table 6.7 reports the Spearman's rank correlation coefficients for both the attack and non-attack conditions for each verification mode.

	Verbal	**Visual**
Attack verifications	0.006	0.170
Non-attack verifications	-0.068	-0.074

Table 6.7: The Spearman's rank correlation coefficients for error data.

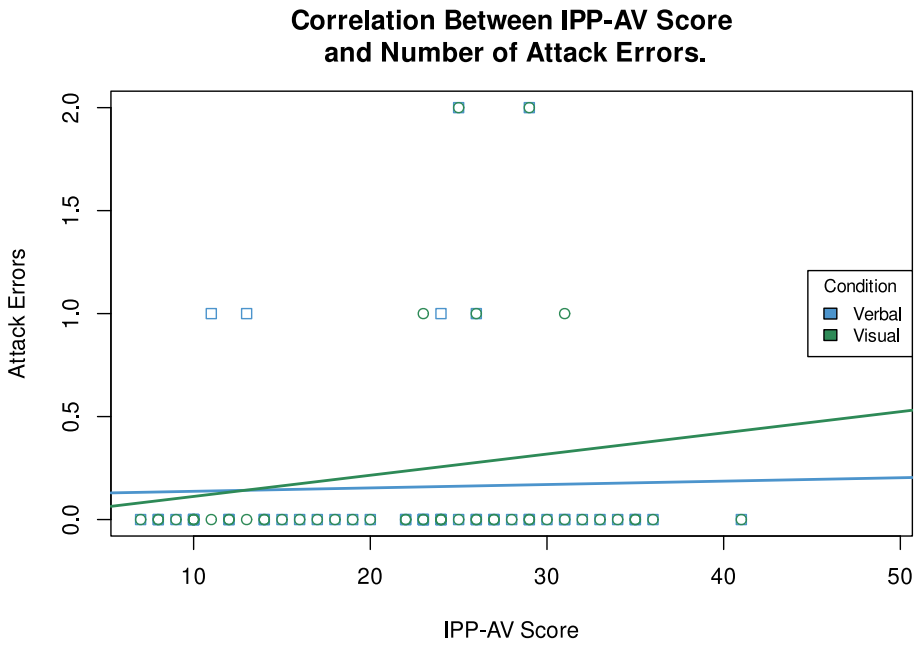


Figure 6.8: Distribution of number of attack errors against IPP-AV score.

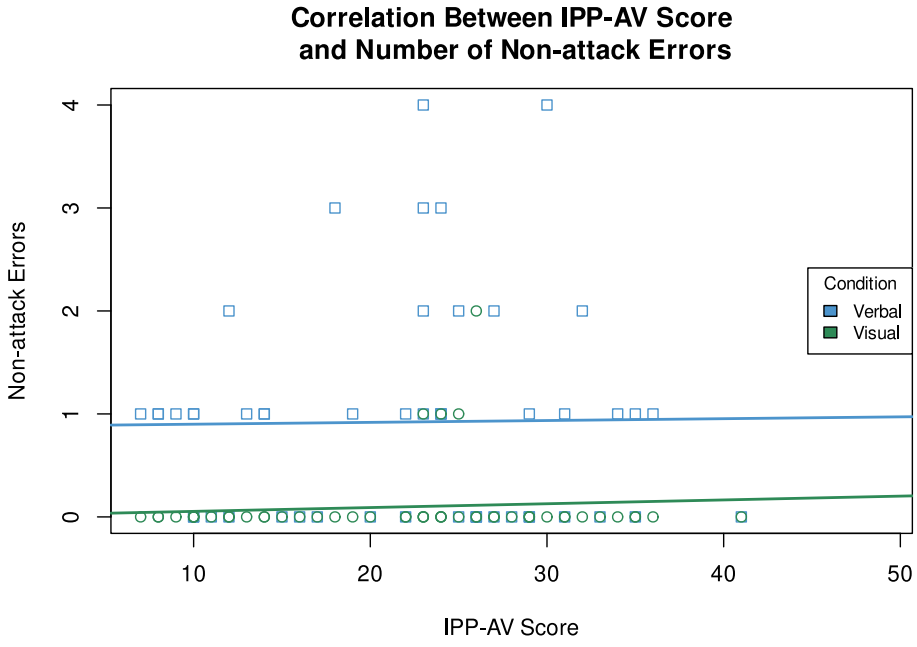


Figure 6.9: Distribution of number of non-attack errors against IPP-AV score.

Efficiency

Table 6.8 reports Spearman's rank correlation coefficients between a participant's IPP-AV score and the average time to make correct verifications. There was no significant correlation for attack verifications or verbal non-attacks (see Figures 6.10 and 6.11). A significant correlation was observed between a participant's IPP-AV score and the time to correctly verify visual non-attacks ($\rho = 0.31$, $p = 0.027$), in agreement with the underlying hypothesis H_4. However, given the few number of overall non-attack errors and given the lack of a full pattern, this is considered weak evidence.

	Verbal	**Visual**
Attack verifications	0.181	0.210
Non-attack verifications	0.002	0.310

Table 6.8: The Spearman's rank correlation coefficients for timing data.

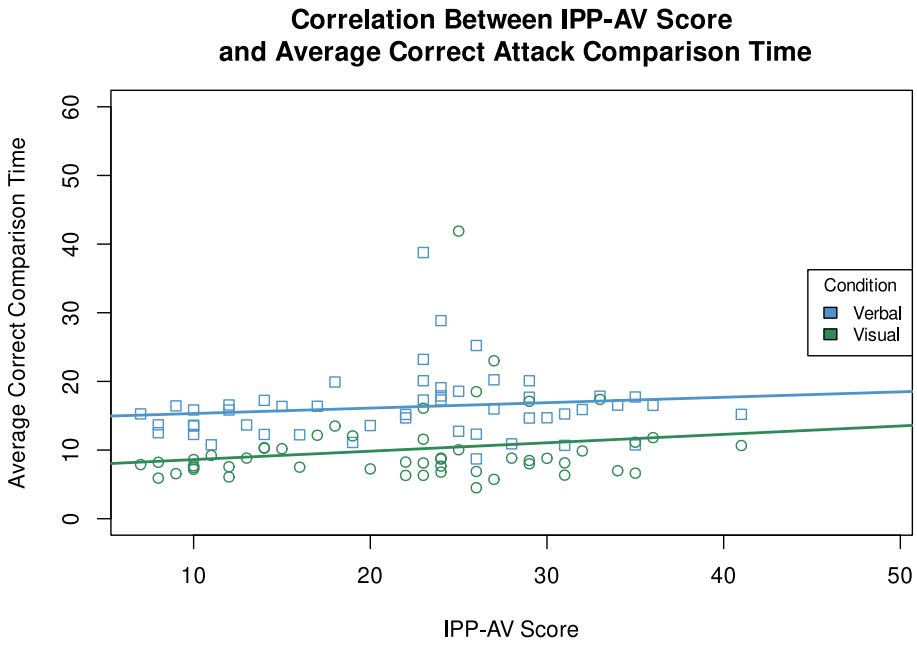


Figure 6.10: Distribution of average correct attack verification time against IPP-AV score.

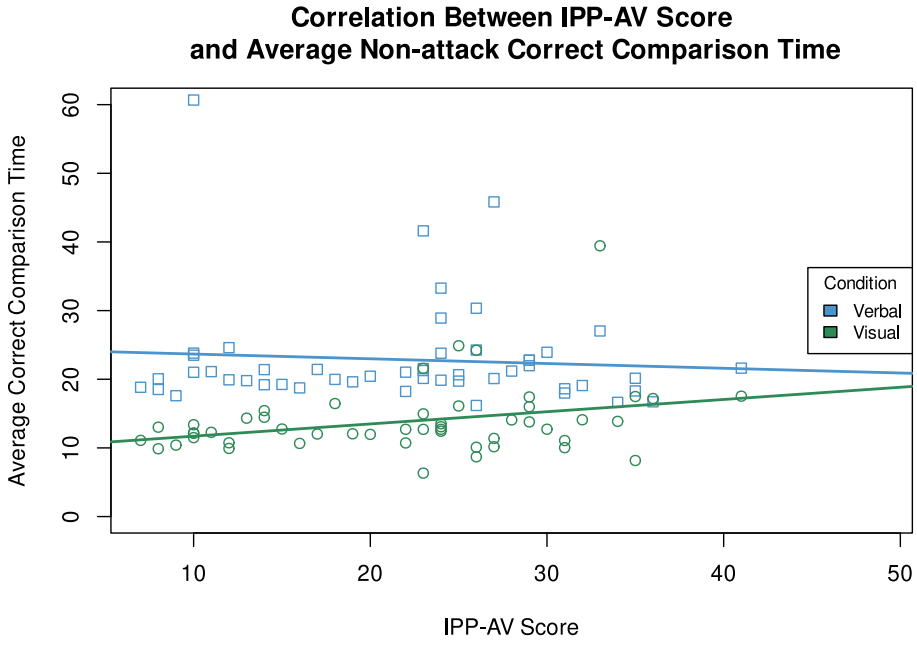


Figure 6.11: Distribution of average correct non-attack verification time against IPP-AV score.

6.4 Discussion

This chapter reported the results of an investigation of the security and perceived usability of numerical key fingerprints, and aimed to determine if the effects identified within earlier sections of this research could be extended to non-word based representations. The study also included further investigation of the impact of a users information processing preference upon their performance, mirroring the analysis performed in Section 5.3.

Participants were observed to make significantly more non-attack errors using the verbal verification mode. This is an important observation which was also observed within the studies described in Chapters 3 and 5, and indicates that this may be a robust effect. The observation also provides evidence against the alternative explanation proposed in Section 5.5 that the increase number of non-attack errors for the verbal mode was a property of the chosen word base. If this were the case, then it would be expected that this study would not identify a significant effect related to the number of non-attack errors. Instead the observed result supports the hypotheses of this research; that the identified differences are an inherent property of the two different verification modes.

In contrast to Chapter 5, no significant differences were observed on the attack verifications. This was surprising as the attack set of this study used the same raw data and was able to identify pairs of attack fingerprints with similar levels of similarity. Thus, there does not seem to be a robust effect related to the effectiveness within the attack condition, and instead any effects are largely dependant upon the chosen representations. A possible explanation for this difference is that participants may have found the attacks easier to identify when encoded using a numerical representation, in line with the alternative explanation. This suggests an interesting contrast between the attack and non-attack

conditions, and that the errors in each condition do not share the same root cause.

Participants were again found to be significantly faster at making the correct decision when using the visual verification mode, with verifications on average 7.36 and 7.95 seconds faster in the attack and non-attack conditions (resp.). The same effect was also observed in Chapter 5, but the difference more than doubled in this study. This raises the question of whether there were other factors in play that caused the increased time, in addition to any differences between the two verification modes. One potential explanation is that the transition to a numerical representation had a greater impact on the amount of time to make verbal verifications, as the fingerprints were read out digit by digit. Consequently, rather than consisting of five easily verifiable chunks, the verbal fingerprints instead form a string of 25 short words. This is a limitation which is not replicated in the visual verification mode, as participants can simultaneously view all digits, enabling a chunk by chunk verification. Thus, though there is some support for a robust difference in the time taken to make the correct decision between the two verification modes, care much be taken regarding the size of the effect as some variation may be caused by other uncontrolled sources.

The visual verification mode was again perceived to be more usable, with significant differences observed upon two of the six dimensions. Visual verification was perceived to be more efficient and participants were more likely to be re-use this mode. Given the efficiency observation discussed above, it is unsurprising that this was replicated in the perceived usability data. However, the study did not observe a significant difference upon the "Confidence" dimension, in contrast to the results of Chapters 3 and 5. Thus, though there appears to be support for a robust effect related to the perceived usability of the two modes, this effect cannot be definitively attributed to the individual usability dimensions. Different factors appear to be of greater importance depending on the specific context of the task.

Chapter 7

Overall Discussion and Conclusions

This thesis aimed to determine the impact of verification mode upon the performance and perceived usability of public key fingerprint verifications. The research included four related human factor studies which systematically investigated a range of aspects that were anticipated to be significant factors of both performance and usability, including scenarios implementing different fingerprint representations and others which considered different measures of attack similarity. The thesis also determined to identify the impact of a users preferred method to receive information upon their performance, measured first by the visual verbal subscale of the ILS and subsequently by a custom IPP-AV scale.

The research was able to develop answers to all of its initial questions, though some were unexpected and not in line with the initial hypotheses. The following section provides a review of the initial research questions and provides answers based upon the results observed.

7.1 Summary of Results

What is the optimal level of fingerprint similarity achievable from a computational search performed by a well resourced and highly motivated attacker?

The results of a computational simulation modelling the fingerprint similarity achievable by a well resourced and highly motivated attacker were reported in Section 4.2. The simulation assumed that the attacker could compute and store 2^{60} random public keys and efficiently identify the related key fingerprint with maximal similarity to that of given high value target. The assumed upper bound of 2^{60} was chosen as it is likely feasible for the envisaged attacker, for example a nation state actor. Furthermore, though difficult and expensive to implement, there are examples of such large computational efforts that have been performed within the academic community, for example the efforts to factor RSA-768 [30]. It is reasonable to suggest that there are attackers willing to implement a similar effort to facilitate an attack against a targets of sufficiently high value.

Care was taken to ensure that the level of similarity of implemented attacks was not artificially inflated by use of an unrealistic model that assumed control in excess of that which is likely to be achieved from a brute force search, for example assuming control over additional components of the central section of the fingerprint. This was a limitation of previously reported work, which simulated attack pairs with unrealistically high levels of

similarity that may have impacted their ecological validity [16].

This research defined a quantitative measure of the similarity of non-matching fingerprint sections, with the simulated brute force attack identifying a set of attacks with phonological similarity coefficient of at least 0.7. These were subsequently used in the studies described in Chapters 5 and 6, causing an increased number of attack errors compared to the exploratory study described in Chapter 3.

Is there a significant difference in user performance and perceived usability between visual and verbal key fingerprint verifications?

The initial premise of this research aimed to identify an interesting effect related to the security of a fingerprint verification. It was hypothesised that an experimental design which included attacks with high similarity but still achievable by a well resourced and highly motivated attacker would be difficult for users to distinguish, and that there may be a significant difference in performance between verifications performed visually and verbally. However, the research found that this was not a clear relationship, with users typically making similar number of errors using either the verbal or visual mode. The outlier to this narrative are the results of the study described in Chapter 5, which did identify a significant difference for word-based fingerprints, with users observed to make more errors using the verbal verification mode. Yet subsequent investigation of fingerprints which employed a numerical representation to the same underlying attack set was unable to find a similar effect (see Chapter 6). This suggests that the effect observed in Chapter 5 may be a property of the chosen word base rather than the chosen verification mode.

Greater success was observed from investigation of differences in usability between the two verification modes. All of the studies described in Chapters 3, 5 and 6 identified a significant difference in the number of non-attack errors between modes, with users observed to make less errors when performing a visual verification. Given the initial premise of this research, this result was somewhat surprising but the evidence suggests that this may be a robust effect. A potential explanation is that it is easier to mishear a word or number than to misread it, which may lead to an increase in errors within verbal verifications.

Significant efficiency effects were also observed within the studies described in Chapters 5 and 6, with participants found to be significantly faster at making correct verifications when using the visual mode.

The research also identified a significant difference in perceived usability between verification modes. As it was of interest to determine the usability of this specific task, a custom set of questions were developed to investigate aspects of usability that were predicted to be of importance, including factors such as confidence, efficiency and trust. The studies described in Chapters 5 and 6 both found significant effects upon a subset of the questions six dimensions, with the visual verification mode perceived to be more usable in each case, though the specific dimensions varied across the studies. Hence though there appears to be some evidence that participants view visual verifications as more usable, there is not sufficient evidence to make conclusions about the individual dimensions.

How can a user's auditory–visual information processing preference be measured?

The measurement of a participants auditory–visual information processing preference was not trivial. Previous instruments included within the literature tended to incorporate the visualiser-verbalisier hypothesis, which groups together the processing of both spoken and written text within the verbal dimension. In addition, these instruments also routinely included additional factors, for example how users prefer to share information themselves, which may have distracted from the objective of assessing information processing.

Thus, a custom seven question IPP-AV scale was developed, based upon previous questions observed within the literature that included auditory–visual information processing scenarios. A preliminary investigation found that this measure aligns with a users perceived preference to receive information, and was subsequently used to investigate the influence of a participant's information processing preference in Chapters 5 and 6. This measure may be of benefit to future researchers who wish to investigate the impact of a user's information processing preference in different scenarios, and such investigations may help to provide additional evidence as to its correctness and validity.

What is the impact of a user's auditory–visual information processing preference upon a key fingerprint verification?

The results of this thesis conclude that a user's auditory–visual information processing preference does not play a significant role in a key fingerprint verification. This is a surprising result, as it is widely accepted that users possess a personal preference for receiving information either auditorily or visually, and it was expected that this may play a role in their performance. However, this was found not to be the case, with the results of Chapters 5 and 6 demonstrating a distinct lack of correlation between a participants IPP-AV score and both effectiveness and efficiency.

7.2 Validity

Human factors research often requires design trade-offs which may affect some aspects of validity. An example of such a trade off within this research was the inclusion of a low attack rate within the exploratory study described in Chapter 3. This sought to maximise ecological validity, but created a floor effect on the number of errors that users could make, compromising the internal validity.

The remainder of this section highlights additional design trade offs which may have affected aspects of validity, with a justification for why each was deemed necessary.

7.2.1 Fingerprint verification is a secondary task

It is important to note that in practise a fingerprint verification is a secondary task to the main goal of enabling users to communicate. This is something that this research did not attempt to model due to the associated increase in complexity of such an experimental

design. Thus, participants may have taken more care with their verifications than real users, and consequently displayed improved performance in verification to the general population.

It appears that participants did provide sufficient attention during these studies, as none failed either of the first two attention checks. Further evidence can be drawn from the fact that participants were significantly faster when completing of attack verification (see Tables 5.4 and 6.5). This is understandable, particularly as attack verifications enable participants to "short-circuit" an attack verification as soon as they identify a difference. This cannot be replicated for non-attacks; there exists the potential for difference to occur anywhere in the fingerprint, and so participants must perform a full verification to ensure that they are identical.

This may be atypical of real user behaviour, and be possible symptoms of the Hawthorne effect. However, even if this is a factor, participants were still observed to display significant differences in performance between modes. It may well be that real users make even more errors due to making less attentive verifications, and so the observed non-attack error effect may be even more pronounced in practise.

7.2.2 Fatigue

A further challenge to ecological validity is that real users will not perform multiple verifications during a single session. A fingerprint verification represents an irregular task performed in very specific scenarios, typically when a user seeks to ensure the confidentiality of their communication with a new correspondent. Repetition is only typically required if a user installs the secure messaging application upon a new device.

However, to facilitate investigation of the difference in user performance between verification modes, it was required to task participants to complete a large number of verifications during a single session. This may have caused fatigue, as evidenced by the high failure rates of the the final attention checks within Chapters 5 and 6 (see Table 7.1). Though this is far from ideal, and could be a potential alternative explanation for some of the errors on the attack and non-attack verifications, it was a necessary trade off. Reducing the number of verifications would would have a consequential impact on the chosen attack rate.

The implemented attack rate of each study aimed to prevent participants becoming accustomed to attack, as attacks are infrequent in practise. The exploratory study of Chapter 3 employed a 10% attack rate, but allowed participants to make at most 2 errors. To mitigate potential floor effects, all subsequent studies included a 20% attack rate. However, both attack rates are likely far higher than would be encountered by a typical user, and this may have caused participants to become overly accustomed to attacks, and display behaviour that fails to be representative of the general population.

	Attention Check 1	**Attention Check 2**	**Attention Check 3**
Chapter 5	0%	0%	69.2%
Chapter 6	0%	0%	60.7%

Table 7.1: The percentage of participants who failed the final attention check in the final two studies.

7.2.3 Is the IPP-AV a good measure?

Identification of a suitable validated instrument to measure a participants auditory–visual information processing preference proved difficult. Consequently a custom scale of seven questions was developed to measure this preference, yet even with this improved measure of information processing preference, this research was still unable to identify a significant difference in performance and perceived usability related to a participants IPP-AV score. The accepted interpretation of this thesis is that a users information processing preference does not play a significant role, but an alternative explanation is that the custom IPP-AV scale also fails to accurately measure the phenomena of interest.

Evidence against the alternative interpretation includes an initial investigation finding that participants IPP-AV scores were aligned to their self reported preference to receive information (see Chapter 4) and that each of the seven questions were carefully selected to include scenarios that were directly related to the phenomena of interest. Consequently, this provides support that the IPP-AV is a good measure of a users auditory–visual information processing preference, at least in the specific scenario of this research. Further research across a range of different information processing tasks is required to develop evidence for the generalisability and reliability of the measure.

7.3 Future Work

There consists considerable potential for further development of this research within future work, both in resolving limitations of the existing work and also developing on the current results. Possible directions include:

- **Including verifications between pairs of participants.**
  An advantage of the verbal verification mode is that it enables a fully synchronous verification between a pair of users, with the the option to compare and confirm with their partner or ask them to repeat specific sections (e.g. What was the third word? How is the last word spelt?). This is a characteristic that is not reflected in the verbal verification mode of this research. Instead the recipient is completely passive, and if a participant seeks clarification they must repeat the full fingerprint. Synchronous verifications would be possible within a laboratory environment, and it is conceivable that such studies may identify different behaviour. Participants may find it easier to agree whether ambiguous words are actually identical, thus reducing the observed error rate within the verbal verification mode.

- **Inclusion of toggle functionality**
  This research simulated a verification task upon two devices are placed side by side, but this is unlikely to be the case in practise. Instead, it is likely that use of visual verifications would be performed on a single device, and that users would be required to toggle between two screens during the verification. Inclusion of this behaviour would help improve ecological validity, and would appear to introduce significant additional challenge as participants would be required to remember the words whilst toggling. This could cause participants to make more errors whilst using the visual verification mode, and so should be further investigated.

- **Impact of fingerprint structure**
  As discussed in Chapter 2 the WhatsApp verification interface implements a tabular fingerprint structure which is different to the typical interface structures which tend to display fingerprints upon a line, with line breaks based upon the devices screen dimension. The tabular structure is an interesting design, and it would likely effect the sections of the fingerprint that an attacker would seek to control.
  An eye-gaze study would help to empirically determine the sections of the tabular fingerprint that are most important in a manual verification, and may assist in the generation of improved attacks. It would be of interest to determine the pattern that participants tend to follow during such a verification. Do they compare by row or column, from left or right? Cultural differences may also play a factor, with participants whose first language is read from right to left applying this approach to a fingerprint verification.

- **Improvement of accessibility for users with individual differences**
  There exist sections of the population who will likely find the fingerprint verification tasks implemented within current secure messaging applications to be particularly challenging, for example user's with sensory impairments (e.g. deafness or blindness), who are dyslexic or are aged over 65. Users with these individual differences may display different behaviour during a fingerprint verification than that of the general population, who formed the intended target audience of this research. This motivated the choice to exclude such users from the studies described in Chapters 3 and 5, as there inclusion would have introduced a degree of uncontrolled variance which may have impacted the observed results.

  However, though user's who possess individual differences were not the main focus of this research, this does not discount the need to develop fingerprint verification tasks with improved accessibility that is suited to their individual needs. Without this there may be specific types of users who are unable to effectively verify a received fingerprint which forces them to face an increased risk to the security of their messages when compared to to global population.

  An interesting area of future work would be investigation of user performance and perceived usability of the current fingerprint verification tasks within these types of users. For dyslexic users it would be interesting to determine if their performance and the perceived usability increased if they were given the option to choose a non word-based representation scheme, for example a numerical representation. Furthermore, it would be interesting to determine if users who are restricted to a particular verification mode are able to effectively utilise it, and if not if any modifications can be made to increase the accessibility of the task. Take for example a blind user who is restricted to usage of a verbal verification mode, that has been shown to cause an increase in the number of non-attack errors. It would be interesting to determine if a blind user's reliance upon verbal processing would facilitate an improved ability to correctly identify identical fingerprint pairs when compared with users with no visual impairment, or if the task would need to be completely redesigned to facilitate their individual needs. This could lead to the formulation of specific guidance for a range of groups with specific disabilities which could improve their ability to communicate securely.

7.4 Overall Conclusions

This thesis aimed to investigate the differences in the performance and perceived usability between visual and verbal fingerprint verification, and the related impact of a users own auditory–visual information processing preference.

The results provide evidence that visual verifications provide improved performance and increased usability, particularly in terms of effectiveness within non-attack scenarios which was deemed to be a robust effect. If participants are required to use a verbal verification, then usability effect may prevent users from achieving their primary goal of communication, and in turn lead them to avoid performing future fingerprint verifications. This may then increase the risk to the confidentiality of their communications if an attacker were to subsequently attempt a MitM attack.

The research also developed a custom scale to determine a user's preference to receive information upon an auditory–visual scale, the IPP-AV, which may benefit future research which investigates the impact of this preference within other fields. Though a participant's information processing preference was not found to be a significant factor within their performance during fingerprint verifications, the participants themselves possessed a clear preference for the mode that they perceived they will perform best in and it is likely that they will show greater levels of satisfaction if provided with the option to use their preferred verification mode.

These findings are likely to be of interest to the developers of secure messaging applications, as they provide guidance and recommendations around how to design fingerprint verification tasks that are both secure and usable. The main recommendation is that applications should provide support and integration for visual verifications within their task interfaces, in addition to the commonly implemented verbal verification mode. This provides an additional outcome for this research, with a specific developer guidance document provided as Appendix A.

This research provided insight into the significant factors of a key fingerprint verification task. It was clear that a visual verification mode is more usable and leads to significantly less non-attack errors. This may lead to improvements in the usability of the fingerprint verification task if the recommendations of this research are integrated within the design of future application releases.

Bibliography

[1] R. Abu-Salma, M. A. Sasse, J. Bonneau, A. Danilova, A. Naiakshina, and M. Smith. Obstacles to the adoption of secure communication tools. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 137–153, May 2017.

[2] A. Anastasi and S. Urbina. *Psychological testing, 7th ed*, volume 7. Prentice Hall & Pearson Education, 1997.

[3] A. Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[4] A. D. Baddeley. Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Q. J. Exp. Psychol.*, 18(4):362–365, Nov. 1966.

[5] J. Borger, J. Ball, and G. Greenwald. Revealed: how US and UK spy agencies defeat internet privacy and security. *The Guardian*, Sept. 2013.

[6] T. Brewster. Warning: Zoom makes encryption keys in china (sometimes). *Forbes Magazine*, Apr. 2020.

[7] J. Brooke. SUS: a 'quick and dirty' usability scale. *Usability Evaluation in Industry*, 1996.

[8] J. Chandler, P. Mueller, and G. Paolacci. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behav. Res. Methods*, 46(1):112–130, Mar. 2014.

[9] T. L. Childers, M. J. Houston, and S. E. Heckler. Measurement of individual differences in visual versus verbal information processing. *Journal of Consumer Research*, 12(2):125–134, 1985.

[10] J. M. Clark and A. Paivio. Dual coding theory and education. *Educ. Psychol. Rev.*, 3(3):149–210, Sept. 1991.

[11] F. Coffield, D. Moseley, E. Hall, and K. Ecclestone. Should we be using learning styles? what research has to say to practice. Technical report, London: Learning and Skills Research Centre, 2004.

[12] J. Cuevas. Is learning styles-based instruction effective? a comprehensive analysis of recent research on learning styles. *Educ. Res. Eval.*, 13(3):308–333, Nov. 2015.

[13] D. Curry. Signal revenue & usage statistics (2023). `https://www.businessofapps.com/data/signal-statistics/`, Jan. 2021. Accessed: 2023-9-3.

[14] A. Das. 9 best secure and encrypted messaging apps for android & iOS — 2020 edition. https://fossbytes.com/best-secure-encrypted-messaging-apps/, Dec. 2019. Accessed: Apr 2023.

[15] S. Dechand, A. Naiakshina, A. Danilova, and M. Smith. In encryption we don't trust: The effect of end-to-end encryption to the masses on user perception. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–415. IEEE, June 2019.

[16] S. Dechand, D. Schürmann, K. Busse, Y. Acar, S. Fahl, and M. Smith. An empirical study of textual key-fingerprint representations. In *USENIX Security*, pages 193–208, 2016.

[17] R. Dunn, K. Dunn, and G. Price. Productivity environmental preference survey. In *Handbook of Measurements for Marriage and Family Therapy*, pages 134–140. Routledge, 2013.

[18] R. S. Dunn, K. J. Dunn, and G. E. Price. *Learning style inventory.* Price Systems Lawrence, KS, 1981.

[19] R. M. Felder, L. K. Silverman, et al. Learning and teaching styles in engineering education. *Engineering education*, 78(7):674–681, 1988.

[20] R. M. Felder and J. Spurlin. Applications, reliability and validity of the index of learning styles. *Int'l Iournal of Engineering Education*, 21(1):103–112, 2005.

[21] N. D. Fleming and C. Mills. Not another inventory, rather a catalyst for reflection. *To Improve Acad.*, 11(1):137–155, June 1992.

[22] O. for National Statistics. UK population pyramid interactive. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/ukpopulationpyramidinteractive/2020-01-08, Jan. 2020. Accessed: 2023-9-30.

[23] M. T. Goodrich, M. Sirivianos, J. Solis, G. Tsudik, and E. Uzun. Loud and clear: Human-Verifiable authentication based on audio. In *26th IEEE International Conference on Distributed Computing Systems (ICDCS'06)*, pages 10–10, July 2006.

[24] Google. WhatsApp messenger. https://play.google.com/store/apps/details?id=com.whatsapp&hl=en_GB&gl=US&pli=1. Accessed: 2023-9-3.

[25] N. Heninger, Z. Durumeric, E. Wustrow, and J. A. Halderman. Mining your p's and q's: Detection of widespread weak keys in network devices. In *Presented as part of the 21st USENIX Security Symposium (USENIX Security 12)*, pages 205–220. usenix.org, 2012.

[26] A. Herzberg and H. Leibowitz. Can Johnny finally encrypt? evaluating E2E-encryption in popular IM applications. In *STAST '16*, pages 17–28. ACM, 2016.

[27] R. Kainda, I. Flechais, and A. W. Roscoe. Usability and security of out-of-band channels in secure device pairing protocols. In *SOUPS*, page 11. ACM, July 2009.

[28] J. R. Kirby, P. J. Moore, and N. J. Schofield. Verbal and visual learning styles. *Contemporary educational psychology*, 13(2):169–184, 1988.

[29] S. Kirchgaessner. Saudis behind NSO spyware attack on jamal khashoggi's family, leak suggests. *The Guardian*, July 2021.

[30] T. Kleinjung, K. Aoki, J. Franke, A. K. Lenstra, E. Thomé, J. W. Bos, P. Gaudry, A. Kruppa, P. L. Montgomery, D. A. Osvik, H. te Riele, A. Timofeev, and P. Zimmermann. Factorization of a 768-bit RSA modulus. In *Advances in Cryptology – CRYPTO 2010*, pages 333–350. Springer Berlin Heidelberg, 2010.

[31] 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*, pages 1–12. ACM, 2009.

[32] W. H. Kruskal and W. A. Wallis. Use of ranks in One-Criterion variance analysis. *J. Am. Stat. Assoc.*, 47(260):583–621, Dec. 1952.

[33] W. L. Leite, M. Svinicki, and Y. Shi. Attempted validation of the scores of the VARK: Learning styles inventory with Multitrait–Multimethod confirmatory factor analysis models. *Educ. Psychol. Meas.*, 70(2):323–339, Apr. 2010.

[34] A. Lenstra, J. P. Hughes, M. Augier, J. W. Bos, T. Kleinjung, and C. Wachter. Ron was wrong, whit is right. Technical report, IACR, 2012.

[35] L. Livsey, H. Petrie, S. F. Shahandashti, and A. Fray. Performance and usability of visual and verbal verification of word-based key fingerprints. In *Human Aspects of Information Security and Assurance: 15th IFIP WG 11.12 International Symposium, HAISA 2021*, pages 199–210. Springer, 2021.

[36] B. Marczak and J. Scott-Railton. Move fast and roll your own crypto: A quick look at the confidentiality of zoom meetings. `https://citizenlab.ca/2020/04/move-fast-roll-your-own-crypto-a-quick-look-at-the-confidentiality-of-zoom-meetings/`, Apr. 2020. Accessed: Apr 2023.

[37] H. Marques and B. Hoeneisen. Pretty Easy Privacy (PEP): Contact and Channel Authentication through Handshake. IETF Network Working Group, Draft, 2020.

[38] R. E. Mayer. Multimedia learning. *Psychol. Learn. Motiv.*, 2002.

[39] J. M. McCune, A. Perrig, and M. K. Reiter. Seeing-is-believing: using camera phones for human-verifiable authentication. In *2005 IEEE Symposium on Security and Privacy (S P'05)*, pages 110–124, May 2005.

[40] S. Messick and others. *Individuality in learning*. Jossey-Bass, Oxford, England, 1976.

[41] Meta. Two billion users — connecting the world privately. `https://about.fb.com/news/2020/02/two-billion-users/`, 2020. Accessed: 2023-9-3.

[42] A. Meylan, M. Cherubini, B. Chapuis, M. Humbert, I. Bilogrevic, and K. Huguenin. A study on the use of checksums for integrity verification of web downloads. *ACM Trans. Priv. Secur.*, 24(1):1–36, Sept. 2020.

[43] B. M. Murphy, B. J. Titcomb, J. Titcomb, B. M. Murphy, B. J. Cook, and B. O. Rudgard. NHS doctors told not to use zoom for video calls with patients. *The Daily Telegraph*, Apr. 2020.

[44] A. Paivio. Dual coding theory: Retrospect and current status. *Can. J. Exp. Psychol.*, 45(3):255–287, Sept. 1991.

[45] A. Paivio and R. Harshman. Factor analysis of a questionnaire on imagery and verbal habits and skills. *Can. J. Exp. Psychol.*, 37(4):461–483, Dec. 1983.

[46] H. Pashler, M. McDaniel, D. Rohrer, and R. Bjork. Learning styles: Concepts and evidence. *Psychol. Sci. Public Interest*, 9(3):105–119, Dec. 2008.

[47] E. Peer, D. Rothschild, A. Gordon, Z. Evernden, and E. Damer. Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods*, 54(4):1643–1662, Aug. 2022.

[48] C. G. Penney. Modality effects and the structure of short-term verbal memory. *Mem. Cognit.*, 17(4):398–422, July 1989.

[49] R. Dunn and K. Burke. The clue to you. `https://webs.um.es/rhervas/miwiki/lib/exe/fetch.php%3Fmedia=lscy_rimanual_v1.pdf`. Accessed: Apr 2023.

[50] E. Rescorla. The transport layer security (TLS) protocol version 1.3. Technical Report RFC 8446, Internet Engineering Task Force (IETF), Aug. 2018.

[51] A. Richardson. Verbalizer-visualizer: A cognitive style dimension. *Journal of Mental Imagery*, 1(1):109–125, 1977.

[52] M. Rosenberg, N. Confessore, and C. Cadwalladr. How trump consultants exploited the facebook data of millions. *The New York Times*, Mar. 2018.

[53] N. J. Salkind. *Encyclopedia of Research Design*. SAGE Publications, Inc., Aug. 2010.

[54] J. Schaad, B. Ramsdell, and S. Turner. Secure/multipurpose internet mail extensions (S/MIME) version 4.0 message specification. Technical Report RFC 8551, Internet Engineering Task Force (IETF), 2019.

[55] S. Schröder, M. Huber, D. Wind, and C. Rottermanner. When SIGNAL hits the fan: On the usability and security of State-of-the-Art secure mobile messaging. In *1st European Workshop on Usable Security*. Internet Society, 2016.

[56] M. Shirvanian, N. Saxena, and J. J. George. On the pitfalls of End-to-End encrypted communications: A study of remote Key-Fingerprint verification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC 2017, pages 499–511, New York, NY, USA, Dec. 2017. Association for Computing Machinery.

[57] Statista. Viber: number of registered users 2020. `https://www.statista.com/statistics/316414/viber-messenger-registered-users/`. Accessed: 2023-9-3.

[58] J. Tan, L. Bauer, J. Bonneau, L. F. Cranor, J. Thomas, and B. Ur. Can unicorns help users compare crypto key fingerprints? In *CHI'17*, pages 3787–3798, 2017.

[59] K. A. Thomas and S. Clifford. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Comput. Human Behav.*, 77:184–197, Dec. 2017.

[60] Threatpost. Final report on DigiNotar hack shows total compromise of CA servers. `https://threatpost.com/final-report-diginotar-hack-shows-total-compromise-ca-servers-103112/77170/`. Accessed: Apr 2023.

[61] D. Turner, S. F. Shahandashti, and H. Petrie. The effect of length on key fingerprint verification security and usability. In *ARES*, pages 23:1–23:11. ACM, 2023.

[62] A. Ulrich, R. Holz, P. Hauck, and G. Carle. Investigating the OpenPGP web of trust. In *Computer Security – ESORICS 2011*, pages 489–507. Springer Berlin Heidelberg, 2011.

[63] E. Vaziripour, J. Wu, M. O'Neill, J. Whitehead, S. Heidbrink, K. Seamons, and D. Zappala. Is that you, Alice? a usability study of the authentication ceremony of secure messaging applications. In *SOUPS 2017*, pages 29–47, 2017.

[64] M. Vengattil and J. Roulette. Elon musk's SpaceX bans zoom over privacy concerns - memo. *Reuters*, Apr. 2020.

[65] B. Waddington. Education, england and wales - office for national statistics. `https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/bulletins/educationenglandandwales/census2021`, Jan. 2023. Accessed: 2023-9-30.

[66] WhatsApp Security. `https://www.whatsapp.com/security/`. Accessed: Apr 2023.

[67] A. Whitten and J. D. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *USENIX Security*, volume 348. USENIX, 1999.

[68] F. Wilcoxon. Individual comparisons by ranking methods. In S. Kotz and N. L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer New York, New York, NY, 1992.

[69] D. T. Willingham, E. M. Hughes, and D. G. Dobolyi. The Scientific Status of Learning Styles Theories. *Teach. Psychol.*, 42(3):266–271, July 2015.

Appendix A

Developer Guidance on the Design of Secure and Usable Key Fingerprint Verification Methods

Between October 2018 and January 2023, researchers led by Lee Livsey at the University of York investigated the security and perceived usability of public key fingerprint verification tasks implemented within modern secure messaging applications (e.g Whatsapp, PGP). Such verifications can provide an extra layer of security for users, as they can provide assurance that they are not the target of a man-in-the-middle (MitM) attack.

Previous related research has concentrated on the usable security of the representation used to encode the fingerprint, with textual representations found to provide a significant improvement over traditional machine encodings, e.g Hexadecimal. However, these investigations investigated only visual verifications with the fingerprint displayed on an intended recipient's business card.

Modern guidance has increasingly encouraged a verbal verification via a voice call, which appears to possess clear benefits of reduced latency and user identification, but this variation was yet to be fully investigated within the academic literature. A verbal verification poses a very different task for users, and it was reasonable to suggest that users may display significantly different behaviour. This research sought to address this gap through investigation of the differences in performance and perceived usability between visual and verbal verification modes. Of particular interest was investigation of:

- Acceptance rates of similar near-matching fingerprints generated by a well resourced and highly motivated attacker (an attack error).
- Rejection rates of identical fingerprints which users perceived to be different (a non-attack error).
- Investigation of the impact of a user's own auditory–visual information processing preference upon their performance.

A collection of related human factors studies identified a robust usability effect between the two verification modes. In all studies, participants were observed to make significantly less non-attack errors when using the visual verification mode. The interpretation is that it is easier to mis-hear than mis-read information, and applications that fail to include support for visual verifications may be less usable. This in turn may discourage users

from completing the fingerprint verification task leaving them vulnerable to a man-in-the-middle (MitM) attack.

Regarding the attack verifications, when fingerprints were displayed using a word-based representation, participants were observed to make significantly more attack errors when using the verbal verification mode. However, an analogous study that used a numerical representation to display the fingerprints was unable to identify a significant difference between the two verification modes. The conclusion is that the verification mode is not a significant factor in an attackers success rate but that the chosen representation is, with a numerical representation more effective at enabling users to identify attacks. A potential explanation of this effect is that word-based representations may include an range of unfamiliar words which may be difficult for users to distinguish between. This effect would likely to be exacerbated for any non-native users or users whose intended recipient possesses a particularly strong accent. These challenges are not shared by the numerical representation, as digits possess significant phonetic differences reducing the potential confusion.

A surprising result was the lack of an effect between a user's auditory-visual information processing preference and their performance. This research developed a custom scale to measure such a preference, named the IPP-AV scale. It was hypothesised that participants with a strong visual preference would display improved performance when using the visual verification mode, and vice-versa. However, this was not shown to be the case, with the data indicating that participants showed equivalent performance regardless of their inherent preference. However, participants were found to possess a clear personal preference for use of one of the verification modes. It is reasonable to conclude that providing users with the option to use their preferred mode may increase their satisfaction and motivate them to continue to perform the fingerprint verification in future.

Another observation was that participants were unfamiliar with the fingerprint verification task and unaware of the protections that it can provide. This result has also been observed within previous research and it is something that should be addressed so that users are able to develop accurate mental models of both message encryption and message integrity.

The results of this research cam be summarised to the following high level recommendations:

- Applications should support both visual and verbal verification modes within their key fingerprint verification tasks.
- Numerical representations should be used to reduce the likelihood of an attacker implementing a successful MitM attack.
- Efforts should be made to improve the user base's mental models of both message encryption and message integrity to help motivate the usage of secure messaging applications and facilitate sufficient protection against MitM attacks. Such efforts will require a determined and co-ordinated education programme, which is likely to be most effective if integrated within the school curriculum to prevent an inaccurate mental model from taking root.

Appendix B

The Attack Set Used in Chapter 3

No.	Authentic fingerprint (top, 5 words) Attack fingerprint (bottom, 5 words)	Similarity Coefficient
1	ALIKEE FLATTERER CURD ROAN CALCOMP ALIKEE FLATTERER COMPLECTED FEELINGLY CALCOMP	0.156
2	BASIS BLEVINS WIRELESS NANETTE BLINDFOLD BASIS BLEVINS PENICILLIN THREADY BLINDFOLD	0.171
3	BUCK HANDEDNESS HOPE JOSEE AUDRA BUCK HANDEDNESS HORROR BRYNNER AUDRA	0.238
4	BUICK EXPEL COATING TIMEWORN RESTORE BUICK EXPEL SCHMALTZ VOLE RESTORE	0.125
5	CHARMION BRIGHT IMITATIVE HOBARD STINK CHARMION BRIGHT EFFUSION BELCH STINK	0.056
6	CHRISTIANS SCROD JEANINE ERMANNO FLYSPECK CHRISTIANS SCROD CERULEAN AINDREA FLYSPECK	0.063
7	CUTTHROAT BARN ROSETTA ORLON DICTUM CUTTHROAT BARN FANIA ULRICA DICTUM	0.155
8	DIGESTIFS WAVELET MANTEGNA COMMENT OUTGOES DIGESTIFS WAVELET ENDEARMENT FLANNEL OUTGOES	0.100
9	FEATHERTOP OVERLONG CONVIVIALITY SETTER DIAGNOSTIC FEATHERTOP OVERLONG JUGGED HISTOLOGIST DIAGNOSTIC	0.045
10	FLATT AZORES NICHROME ACUTENESS CRANDALL FLATT AZORES DARIN SUPPRESSANT CRANDALL	0.199
11	GARGLE KARLOTTE CONGENIAL SUMMABLE DRUB GARGLE KARLOTTE WASSERMANN ENLARGER DRUB	0.163
12	HYGIENIST CENSURER SHAKING ORDERER CHECHEN HYGIENIST CENSURER CONFECTIONER TIEBOUT CHECHEN	0.083
13	LEAN STICKINESS FONSIE DETERRENT REVERTER LEAN STICKINESS SHIPPED MATTHUS REVERTER	0.127
14	LYMPHS SINDHI DEEP WHINING ASSUAGE LYMPHS SINDHI DECISIVENESS UNIT ASSUAGE	0.268
15	MELLICENT TUBER COPYIST ANIMIST BERNOULLI MELLICENT TUBER CLAIROL BRAILLE BERNOULLI	0.071

16	MOISTNESS RADIOLOGY ARACHNID BLUEBUSH SNAP MOISTNESS RADIOLOGY JETH KATHIE SNAP	0.063
17	NATTY VARY CONGRUENCY MENINGITIS CHIMPANZEE NATTY VARY CRAP FORECASTER CHIMPANZEE	0.150
18	NONPAREIL CIVICS YORE GIRLHOOD BONDY NONPAREIL CIVICS PADDED DAEL BONDY	0.146
19	ODDMENT CESSATION ROADWORTHY BYTE RACCOON ODDMENT CESSATION JOELLE MODICUM RACCOON	0.050
20	OFFICER DISCRETION SETTABLE BIRK LOURDES OFFICER DISCRETION SMOOTHEN ESTELL LOURDES	0.063
21	PACIFISM ZED YAHWEH MEEK TRANSFORMER PACIFISM ZED PHILATELIST UNMOBILIZED TRANSFORMER	0.136
22	PAIL BRONTOSAURUS LINGUIST DEFT BOOTPRINTS PAIL BRONTOSAURUS ASHTON THAMES BOOTPRINTS	0.000
23	RECEPTIVITY WASHOUT DURST DADDY ROTUND RECEPTIVITY WASHOUT CIRRI CARPEL ROTUND	0.183
24	RHODOLITE WADI MUGGED BOY JODY RHODOLITE WADI DUMBWAITER CALISTHENICS JODY	0.100
25	SWEDEN COURIER MOCK MYRLENE SCARCE SWEDEN COURIER CROWNER ADROITNESS SCARCE	0.221
26	TREACHERY SEMANTICAL BIBLICISTS TOBIAS RUMMEL TREACHERY SEMANTICAL APPROVAL FIZZLE RUMMEL	0.000
27	ADIPOSE MALRAUX COZUMEL TAXATION MACK ADIPOSE MALRAUX FEUD BOLSHEVIST MACK	0.121
28	ANSELMO BLUEBUSH SEMPSTRESS ABSTENTION AFIELD ANSELMO BLUEBUSH HUMVEE JAUNTY AFIELD	0.200
29	BOB CRAGGY FATAL GOLDI ENDURABLE BOB CRAGGY BOGGLING DEMIMONDAINE ENDURABLE	0.188
30	BRYNN BLUEGILL FERREIRA TONGUING BEAUMONT BRYNN BLUEGILL BLACKSNAKE OKEECHOBEE BEAUMONT	0.050
31	CAPISTRANO ECUMENIST BIONIC TEDDY JAM CAPISTRANO ECUMENIST MUKLUK BONGO JAM	0.000
32	COLLETE LEVIER ARCH RESOLUBLE STAID COLLETE LEVIER REGROUP JAVA STAID	0.071
33	COSILY NOSTALGIC CHEST NATIVIDAD RHETTA COSILY NOSTALGIC HURON CRUSTACEAN RHETTA	0.050
34	COX PUPPETEER GRATER AUNT IMPLEMENTOR COX PUPPETEER PAGE CONDENSE IMPLEMENTOR	0.229

Table B.1: The set of attack fingerprints used in Chapter 5.

Appendix C

The Attack Set Used in Chapter 5

C.1 Phonological Attack Set

No.	Authentic fingerprint (top, 5 words) Attack fingerprint (bottom, 5 words)	Similarity Coefficient
1	MADAM REASON BOOM WRIGGLY LOCKSTEP MADAM REASON BOOTHE GIGGLY LOCKSTEP	0.817
2	PEACETIME CADET CYNDI TRUTHS MINIFY PEACETIME CADET CYNDIE ABSTRUSE MINIFY	0.778
3	CARPATHIAN SOPHISTIC FOURIER NESSA EUROPA CARPATHIAN SOPHISTIC ORELIA NESTLE EUROPA	0.774
4	HULK SERVE TORI CHRISTYNA EWELL HULK SERVE MAURY CATINA EWELL	0.767
5	WELDER AGRONOMY DENNEY APHELIA REWIND WELDER AGRONOMY DENNI HOLIER REWIND	0.750
6	SLUMMY OMERO ALYSON MAGGI LIFELONG SLUMMY OMERO ALISON CRANNY LIFELONG	0.750
7	RENSSELAER GOLLY KEVON IMPLICANT PHLOX RENSSELAER GOLLY SEVEN IMPLEMENTER PHLOX	0.742
8	GYMNASIUM NATO BECKON COAT CALORIMETRY GYMNASIUM NATO DECO SHOAT CALORIMETRY	0.733
9	CHERI INVINCIBLE ZOOM CORKS ALEXANDER CHERI INVINCIBLE NOON CAUCUS ALEXANDER	0.729
10	SHELBY COMPOSED CRISTIANO ACCORDANCE MOISTNESS SHELBY COMPOSED CHRISTIANO KOONTZ MOISTNESS	0.722
11	DESTRUCTION NUTRITIOUS CONQUER DISSENTER ESSENTIALLY DESTRUCTION NUTRITIOUS WRONSKIAN DESCENDER ESSENTIALLY	0.715
12	POLYGON DEVIATED ANT VITAL QUIRING POLYGON DEVIATED AUNT BINDS QUIRING	0.714
13	PROLIX ASHER BARHOP JOELLY POLICEMEN PROLIX ASHER BERTI JOLIE POLICEMEN	0.714

14	PHARAOHS SUNNING TERRILL EWELL BRIEFED PHARAOHS SUNNING ATTESTER EUELL BRIEFED	0.714
15	CESSPIT BOPPED FRASER SERRATE OUTLIVE CESSPIT BOPPED ALTERATION CERATE OUTLIVE	0.708

Table C.1: The set of attack fingerprints used in Section 4.3 and Chapter 5.

C.2 Orthagraphical Attack Set

No.	Authentic fingerprint (top, 5 words) Attack fingerprint (bottom, 5 words)	Similarity Coefficient
1	KNUCKLEHEAD ALICA BEE MATILDA MAC KNUCKLEHEAD ALICA BETTE MATHILDA MAC	0.738
2	DECLINER NEEDY CALOOCAN ANOINTMENT PLAINSPOKEN DECLINER NEEDY CLOACA APPOINTMENT PLAINSPOKEN	0.722
3	KOHINOOR AROMATICITY SWAM HOLMAN SAXONY KOHINOOR AROMATICITY SWIM BOWMAN SAXONY	0.708
4	DESPONDENCY ADSORBATE BARR COLLIMATES ACETYLENE DESPONDENCY ADSORBATE BALD COLLIMATED ACETYLENE	0.700
5	SHAKABLY GREY ROONEY OVEREATER HATCHERY SHAKABLY GREY COINED OVEREAGER HATCHERY	0.694
6	DOVISH ASSEMBLAGE BETTE MECHANIZE COLORIZATION DOVISH ASSEMBLAGE BEALE MECHANIST COLORIZATION	0.689
7	APHELIA ANNULMENT DAREEN CONTRIVE MOTTO APHELIA ANNULMENT BARMEN CONTRIBUTE MOTTO	0.683
8	ANTHRACES POETICAL ERADICABLE RABBLE KRONE ANTHRACES POETICAL ORDINAL RUBBLE KRONE	0.667
9	EMBARKATION ANTIFREEZE EQUIVOCATOR PERISH GRENADINES EMBARKATION ANTIFREEZE EQUIVOCAL IMPOVERISH GRENADINES	0.664
10	CONGO NEGLIGIBLE BENEFACTOR DEMITTED MAGGOTY CONGO NEGLIGIBLE BENEFACTION AMITIE MAGGOTY	0.659
11	AINSLEE ASTIGMATISM SHANI APPROPRIATED SPOKESMAN AINSLEE ASTIGMATISM HAS APPROPRIATE SPOKESMAN	0.658
12	SECRETARY KITCHENETTE SHURLOCKE CARTOGRAPHER RISKER SECRETARY KITCHENETTE SHERLOCKE STRAFER RISKER	0.653
13	GUMPTION THWART BOBINETTE CARING BRUXELLES GUMPTION THWART BONED CARLING BRUXELLES	0.651
14	ASYMPTOTE LOST BOBS BINDS BOREALIS ASYMPTOTE LOST BOBCAT BONDS BOREALIS	0.650
15	JUNG TIRADE BEARABLE EXTRACTION HELIPORT JUNG TIRADE EQUABLY TRACTION HELIPORT	0.650

Table C.2: The set of attack fingerprints used in Section 4.3.

Appendix D

The Attack Set Used in Chapter 6

No.	Authentic fingerprint (top, 5 chunks) Attack fingerprint (bottom, 5 chunks)	Similarity Coefficient
1	11226 25536 43511 59432 44815 11226 25536 42511 55432 44815	0.8
2	19349 51455 36531 23326 21594 19349 51455 33531 23356 21594	0.8
3	16716 03182 49270 09682 00989 16716 03182 49230 09782 00989	0.8
4	56611 58026 03985 51861 19525 56611 58026 02985 51061 19525	0.8
5	44893 25013 34325 20103 13555 44893 25013 34315 20143 13555	0.8
6	57300 15033 51599 52706 11156 57300 15033 41599 05276 11156	0.7
7	52024 40079 01668 57575 08836 52024 40079 16638 59575 08836	0.7
8	53597 35217 52617 03092 64297 53597 35217 52613 02099 64297	0.7
9	04844 60368 60113 59401 59734 04844 60368 65013 59601 59734	0.7
10	61689 21366 39302 47441 39223 61689 21366 39522 47440 39223	0.7
11	48515 62602 11438 03227 16394 48515 62602 10431 03207 16394	0.7
12	00230 61404 20582 34746 60047 00230 61404 20412 04746 60047	0.7
13	36930 08372 14451 18101 03545 36930 08372 14437 18103 03545	0.7
14	63964 35993 02795 04788 01822 63964 35993 24795 04488 01822	0.7
15	52051 60956 24079 55827 50424 52051 60956 24379 05582 50424	0.7

Table D.1: The set of attack fingerprints used in Chapter 6.

Appendix E

Information Sheet

You are invited to take part in a study investigating the performance of word-based security checks, an optional extra step commonly found in many secure messaging applications. These security checks provide extra assurance that your messages remain secure and cannot be read by a malicious third party.

The security check in this study requires users to compare a five-word combination displayed on their screen with their intended recipient. If the word combination is identical on both devices, then they can be confident that no one else can read their messages. If any of the words are different, this may indicate that an unauthorised party is listening in on their messages. Some comparisons may include words that are not part of the English dictionary and seem made up, but they are acceptable so long as they are identical.

There are a variety of methods to exchange the five word combination. We are particularly interested in investigating differences between comparisons that use a text message with those that use a phone call. We are also investigating the impact that a user's preferred method to receive and process information has upon their performance.

Who is conducting the research?

This study is part of a research project by Lee Livsey (lwl501@york.ac.uk) a PhD student in the Department of Computer Science at The University of York.

Lee is advised by Prof. Helen Petrie and Dr. Siamak F. Shahandashti, and his PhD study is supported by a scholarship from The Engineering and Physical Sciences Research Council.

What will I be asked to do?

At the beginning of the study you will be asked to answer 10 multiple choice questions that enable identification of your preferred method to receive and process new information.

You will then compare 50 pairs of word combinations split into two groups of 25. One group will show both word combinations on screen to simulate an exchange by text message, and the other will include a pre-recorded voice clip to simulate an exchange using a phone call.

At the end of the study we will ask some questions about your general usage of secure messaging applications followed by some demographic questions.

The survey should take around 25 minutes to complete.

Your participation in this study is voluntary and you may withdraw up to two weeks after participation. If you wish to withdraw your responses after completion of the study, please send an email to lwl501@york.ac.uk and quote your unique participant identifier. Please make a note of your identifier now:

Your ID: *[...REDACTED...]*

Are there any risks associated with completing this study?

The study does not pose any foreseen risks and does not include any offensive or inappropriate material, though some of the words that are included in the word combinations may be unusual or crass.

This study follows ethical principles and has been approved by our University ethics board.

What data will be collected?

This study is anonymous. Participants are only identifiable via their random identifier, and we do not ask for any personally identifiable information such as your name.

There will be limited demographics data collected including gender, age, education level, location, English proficiency, and any disability affecting the tasks in the study. You can opt not to share this information if you wish. These are collected only to have a statistical overview of the composition of our sample and are not analysed on an individual basis.

Further data collected concern preferences on receiving and processing information, accuracy and time taken to complete each of the comparisons.

We will only collect data that is relevant to our study and will ensure that it remains confidential and is stored securely.

This data will be analysed by the team working with Lee and may be used in future academic publications, where all individual responses are anonymised and any personal identifying data removed.

The data may be stored for up to 10 years following the University standards.

Are there any special requirements for participation?

Due to screen size requirements, the study should not be completed using a mobile device. Instead we ask that you use a desktop or laptop device.

To complete this study you must be over the age of 18.

We ask that you find a quiet and well-lit space to complete the study, where you are unlikely to be disturbed and can clearly hear spoken words and read displayed words.

What can be gained by completing this study? Through completion of this study you will gain a greater appreciation for the additional security provided by word-based security codes and contribute to original academic research.

What's the legal basis for this research?

The University processes personal data for research purposes under Article 6 (1) (e) of the GDPR: Processing is necessary for the performance of a task carried out in the public interest.

Special category data, e.g. those included in demographics data, is processed under Article 9 (2) (j): Processing is necessary for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes.

The "data controller" for this study is the University of York. If you have any data protection questions, comments, or complaints, you can contact the University's Data Protection Officer (dataprotection@york.ac.uk).

- I have read and understood this participant information sheet, and I consent to participate in this study.
- I do not wish to take part in this study.

What is your age?

- I am over the age of 18
- I am aged 17 or younger

Appendix F

Demographics Questions

1. What is your gender?
    - Male
    - Female
    - Non-Binary
    - Prefer to self identify
    - Prefer not to say
2. What is your age?
    - 18-24
    - 25-34
    - 34-44
    - 45-64
    - 65 and over
    - Prefer not to say
3. What is the highest level of education that you have completed?
    - High-school education or equivalent
    - Bachelors degree (e.g., BSc, BA)
    - Postgraduate degree (e.g., MSc, MA, MBA, PhD)
    - Vocational training (e.g., NVQ, HNC, HND)
    - Other (please specify)
    - Prefer not to say
4. In which country do you currently reside?

5. How do you rate your English proficiency?
    - Poor
    - Fair
    - Average
    - Good
    - Excellent
6. Do you consider yourself to have a disability that affected your interaction with this study?
    - Yes - Dyslexia
    - Yes - Visual Impairment
    - Yes - Auditory Impairment
    - Yes - Other (please specify)
    - No
    - Prefer not to say
7. Do you have any problems dealing with numbers, e.g entering PIN numbers or credit card numbers in the wrong order?[1]
    - Yes
    - No
8. Please briefly describe the problems you encounter when handling numbers.

[1]The final two questions were only asked in the study described in Chapter 6.