Designing for Numerical Transcription Typing: Frequent Numbers Matter

CANDIDATE

Sarah EM Wiseman

A dissertation submitted in partial fulfilment of the requirements for the degree of:

Doctor of Philosophy of **University College London**

UCL Interaction Centre,
Department of Psychology and Language Sciences,
University College London

DECLARATION

I, Sarah Wiseman, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

In the text entry domain, the task of number entry is often overlooked despite the prevalence of number entry tasks in the real world. Number entry often occurs in safety critical contexts, such as the medical domain, where errors can lead to patient death. In order to prevent errors from happening, it is important to design devices that help the user in their number entry task, and guard against error. To do this effectively, more needs to be known about the task of number transcription so that appropriate design interventions can be created.

Current research commonly uses randomly generated numbers in the evaluation of number entry interfaces. However, it is not clear that random numbers are appropriate in this context. The first half of the thesis builds on research that shows that the familiarity of a number can affect how it is read, and investigates how this finding impacts upon transcription of familiar numbers. This is investigated by replicating seminal transcription typing studies using both words and numbers. The results of these experiments suggest that familiar numbers are represented more strongly than non-familiar numbers in memory, and as a result familiar numbers are significantly faster to transcribe.

This novel finding then motivates a series of studies that aim to reduce errors in a medical number entry task. First, a log analysis of hospital devices shows that there are clear patterns in the numbers used, providing evidence that medical workers are likely to be more familiar with some numbers rather than others. The knowledge of these frequently used numbers is then utilised in three novel approaches to number entry interface design. First, knowledge of the landscape of frequent numbers in this context is used to create a set of heuristics for the design of number entry interfaces. Second, an experiment shows that adapting the interface specifically for frequent number entry can speed up interaction. Finally an experiment explores how an understanding of the numbers used to program devices can be used to check for and prevent number transcription errors.

This thesis highlights the importance of understanding the frequency and familiarity of numbers used in specific contexts. It explores how this knowledge can improve both evaluation and design of number entry interfaces.

Acknowledgements

During my PhD I have been surrounded by people who have made the experience not only bearable but enjoyable. I would like to thank a small set of those people here. Unfortunately, there is no space to thank all of my friends who have been there for me when I needed them, but their support has been invaluable.

I must thank my PhD supervisors Anna Cox and Duncan Brumby. As a supervisory team they are perfectly balanced and have encouraged me to think big, but still get things done (eventually). Their guidance, at all hours of the day, has helped me produce a piece of work I can feel proud of.

Being a part of UCLIC has been an incredibly rewarding experience. I feel very lucky to be able to work in a place where I can call so many people friends. Discussions with many of my colleagues, both in and out of the office, have helped shape this thesis. I would especially like to thank Sandy for chatting throughout the years, and for sharing with me his newly acquired knowledge of the PhD submission process.

I was also able to be part of the CHI+MED team. I am thankful to everyone on the team for their useful discussions on my work, and for continuing to be an inspiration by working to make medical devices safer.

I would like to thank Anna BD for not only being a proof-reader, but a cheer leader too; both were equally necessary jobs. And thanks too, to Mum for reading the entire thesis and significantly reducing the number of typos and grammar errors.

I would finally like to thank my parents and sister for being incredibly supportive of me throughout my academic career. They have always encouraged me in my various pursuits and have inspired me to work hard to achieve my goals. It is thanks to them that I have had the great privilege to be able to do something I love for the last 4 years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Since the invention of the typewriter, researchers have been investigating the way we transcribe words. The interest of these studies is often, but not always, to gain an insight into the way we type. For instance, in 1916, Wells studied the categories of errors made when copy typing on a typewriter, and aimed to understand how typists could become more efficient.

However, typing studies can also be used as a means to understand other cognitive processes. Swift published a study of typewriting with the primary aim of understanding the process of skill-learning (Swift, 1904). During the study, Swift closely followed a typist as he learnt to use a typewriter, hoping to document his improving typing skill through typing speeds. Since this point, typing studies have been used to understand underlying cognitive processes.

One of the key milestones in this field in 1986 was the paper "Perceptual, Cognitive and Motoric Aspects of Transcription Typing" (Salthouse, 1986). This review article collated a series of previous typing studies into one list of typing "phenomena". These phenomena consisted of replicated typing effects that had been found in many studies over multiple participants. The 29 phenomena listed by Salthouse cover various aspects of typing, such as speed of typing, likely errors, and how typing is affected by multitasking.

These typing phenomena have become very important in the study of transcription typing. They have also taught us what cognitive processes occur during transcription typing, such as how our perceptual and motor processes work in parallel, and how much information is stored in working memory. Through typing, we have learnt how we read and think about words and non-words. This is a key aspect of the Salthouse phenomena: that words and non-words elicit different typing behaviours when transcribed. In addition, the typing phenomena have become a set of benchmarks for cognitive models of transcription typing, and it is widely held that they should be replicated in models of typing (John, 1996; Wu & Liu, 2008).

Typing studies have not only improved our knowledge about the cognitive processes involved in transcription typing, they have also improved our day-to-day experiences with keyboard interfaces. One initial study into the words often typed in the English language led to the rearranging of the QWERTY layout that many people had grown used to. In 1936, the researcher Dvorak proposed a new layout for the letters on the keyboard (Dvorak et al., 1936). This new layout would improve the user's ability to type common digrams (pairs of letters) by separating them and making them easier to type with alternate hands. This was in addition to other improvements based on the ergonomics of hand movements and letter frequencies in the English language. This new layout did not appeal to typists and we still type on the original QWERTY layout, despite its apparent inefficiencies. However, the QWERTY keyboard has been adapted for non-English speaking countries. In France, for example, the AZERTY keyboard is used, as it better matches the requirements for typing words in French. The design of keyboards since this work has not remained static; much research is still conducted into understanding the optimal functionality of the keyboard.

A modern example of this comes from the introduction of the touchscreen. Touchscreens have presented a new challenge with regard to understanding how text entry systems should function, but they have also allowed for more creativity with design. For instance, keyboard buttons can now be moved around dynamically, changing the area on screen that relates to each key, without changing the visual look of the key (Rudchenko, Paek, & Badger, 2011). This helps compensate for poor targeting of keys, without confusing the user. This can be done due to context information gathered from the words being typed. Similarly, prediction software can also help make the typing experience easier for users; by either prompting words for typists, or correcting likely errors.

One commonality between all typing studies is the importance of understanding which words should be used when testing the interfaces. The Salthouse phenomena highlight the fact that aspects of a typist's behaviour (speed, accuracy, etc) are different when typing words and non-words, and for this reason it is important that, when designing a text entry system, users test it with real words, not randomised character strings to ensure that the test is ecologically valid (Brunswik, 1955). Beyond this, most text entry evaluations also strive to accurately simulate common typing tasks, often using the same word set to ensure fair and accurate comparisons between designs (MacKenzie & Soukoreff, 2003).

The research field for text entry is one of the oldest in the domain of HCI, and is still active and necessary today as the technology on which we enter text changes. However, despite this level of innovation, much of this research into "text" entry has focused solely on alphabetic text entry, despite the fact that text entry can involve both alphabetic and numeric characters.

Numeric text transcription, although not as prevalent as alphabetic text transcription, is a task that occurs often. A specific key area for numeric transcription is in the medical domain, where numbers are needed to specify key patient information, as well as drug dosage information. In this context, it is incredibly important that number entry is done correctly, as an incorrect dose or piece of patient information can lead to serious harm and even death.

In one such case in 2010, a patient was given incorrectly sized lungs during a transplant, as two numbers in the donor's height had been transposed: the donor was recorded as being 156 cm tall, when correct height was 165 cm. This error meant that the lungs were too large for the 151 cm tall recipient, and ultimately led to the patient's death (Adam, 2011). This, unfortunately, is not a lone case and many instances of incorrect number entry leading to deaths in the medical domain have been recorded (Vicente et al., 2003). Beyond just number entry errors, it is estimated in the US alone that up to 98,000 deaths per year in the medical system are due to medical error (Kohn et al., 2000).

It is not just the medical domain in which accurate number entry is important, it is also a key issue in the financial domain where one incorrect number can cause a company to collapse. In numerous cases stocks have been bought or sold for 10 to 100 times their intended price, causing giant fluctuations in the stock market (McCurry, 2005). A more general, and everyday example of a number entry task in the financial domain is that of entering credit card details during online shopping. This involves copying (or recalling) a series of numbers from the credit card to be typed into a device. When an error occurs during this task, the payment will not go through and the transaction may be aborted causing frustration to the user.

When considering the importance of accurate number entry in these domains alone, it is surprising that it has attracted relatively little attention. To date, number entry research has tended to focus on understanding the errors that occur during number entry, and how they might be prevented (Thimbleby & Cairns, 2010; Wiseman et al., 2011) as well as what effects the number entry interface could have on error rates and typing speeds (Cauchi et al., 2012; Oladimeji et al., 2011). However, there has been very little work on understanding the domain specific aspects of numbers, and how that can affect both the user and design of number entry interfaces.

There is currently no numerical equivalent to Salthouse's seminal paper on typing phenomena that can apply to numbers. What if there are similar discoveries to be made with regard to numerical typing? If we are to design safer, easier to user number entry interfaces, firstly we must know this information. A list of typing phenomena relating to the task of number entry is required.

Additionally, there are currently no number sets that are analogous to the standard word sets commonly used in text entry studies (MacKenzie & Soukoreff, 2003). Without knowledge about the numbers being used, there is no clear way to adapt the interfaces to better suit the task of number entry. The features that can aid users in the alphabetic text entry domain cannot currently be replicated for number entry. In addition to understanding how users transcribe numbers, it is also important to know exactly what numbers they are transcribing.

The aim of this thesis, therefore, will be to explore these ideas and to begin working towards both an understanding of the typing phenomena of number transcription, and of the numbers that are used in applied domains. This thesis will contribute to both the theory of number reading and typing, and also to the design and evaluation of number entry interfaces.

1.2 Thesis Structure

The literature review highlights the importance of words compared with non-words and ecological validity in the field of text entry research. A case is made that similar care should be paid to the numbers entered during numeric transcription. This point is strengthened by case studies from aphasia research, in which patients are presented who retain the ability to read aloud "meaningful" numbers, yet lose the ability to read "meaningless" numbers. Comparisons are drawn between this phenomena and similar cases of patients losing the ability to read and identify non-words. This evidence supports the theory that there may be different cognitive processes occurring during the transcription of familiar versus non-familiar numbers, much as there are when transcribing words and non-words. These differences may be accounted for by representation of words and numbers in memory.

This thesis consists of ten studies. The first empirical chapter (Chapter 3) presents (Study 1), which aims to elicit a set of familiar and non-familiar numbers from a general population of English-speaking internet users for use in future typing experiments. This study also provides the first result to support the hypothesis that familiar and non-familiar numbers are read differently: an analysis of the response times during the study shows that participants were able to decide that a number was familiar significantly quicker than deciding it was non-familiar. This suggests that the familiar numbers are stored in long term memory and are easily accessed, whereas non-familiar numbers are not. This distinction is similar to the difference between words and non-words in equivalent studies (Meyer & Schvaneveldt, 1971).

A second study in Chapter 3 aims to establish what makes the numbers gathered in the previous study familiar. Study 2 presents the results of a card-sorting task performed with participants. The participants generated a series of possible categories that could provide reasons for the numbers' familiarity; these categories include important years, phone numbers, and prime numbers for example. The categories are grouped, and a short set of rules are produced that can aid in future experiments which may require familiar number lists for different populations.

Chapter 4 makes use of the familiar and non-familiar number corpuses generated in Study 1 through a series of experiments aimed at exploring possible Salthouse typing phenomena for numbers. Four experiments (Studies 3-6) investigate the effects of number familiarity on four of the typing phenomena. These experiments show that there are indeed significant differences in the way that familiar and non-familiar numbers are typed (in terms of speed of transcription). This provides further evidence that familiar numbers have a stronger representation in long term memory. This result suggests that familiar and non-familiar numbers should not be treated equally in typing studies. However, familiarity does not appear to affect typing spans, suggesting the distinction between familiar and non-familiar numbers may not be measured as easily as that between words and non-words. The experiments also highlight how similar non-familiar numbers are to non-words in terms of how they are stored in memory.

In Chapter 5, the results of the previous studies and experiments motivate investigations of number entry in the medical domain. Many studies conducted with the aim of improving number entry in the medical domain will ask users to enter randomly generated numbers (Oladimeji et al., 2011; Thimbleby & Cairns, 2010; Wiseman et al., 2011). If there is a possibility that the numbers entered in particular tasks are not random, then this method is not ecologically valid. In Study 7 the logs from medical devices used in a hospital are collected, and the numbers entered are analysed. This analysis shows that there are novel patterns in the numbers used when programming infusion pumps. This suggests that there are a frequent set of numbers that medical workers are familiar with. Evidence from the previous experiments suggests that medical workers may be entering these numbers more quickly than numbers they are less familiar with. This suggests that random number generation is not an ecologically valid way of testing the number entry task.

The finding that numbers used in particular domains have patterns does not only aid in the evaluation of number entry devices, but it can also inform design. Study 8 summarises the log analysis results into a short set of heuristics that can be used when designing or evaluating medical device interfaces. Study 9 uses these heuristics, and a knowledge of the landscape of numbers used when programming infusion pumps, to adapt existing number entry interfaces. This adaptation allows the most frequent numbers to be entered more quickly, and with fewer keypresses. These tests show that speed can be increased without incurring more errors. This approach however, does not successfully reduce the error rate. In Study 10, another aspect of the number patterns is used; by incorporating the natural checksum within medical numbers into a new interface design, it is possible to check for number entry errors made by the user. An experimental study provides evidence that this approach to error prevention is successful at reducing the errors made. Through these experiments, Chapter 5 highlights the importance of understanding the numbers used in a particular context. With this knowledge, interfaces can be adapted to better suit the particular number entry tasks they are used for, and can increase speed and reduce error rates.

1.3 Contributions

Through these experiments, this thesis has made multiple contributions to both the theoretical and applied domains; the contributions are listed here.

1.3.1 Contribution to Theory

This thesis suggests that number transcription is a different task to word transcription. Results from the studies within Chapter 4 presented evidence that words are typed significantly faster than numbers, and that the eye-hand span for words is larger than the span for numbers. This suggests that words have a stronger representation in memory than numbers. This result also has implications for transcription modelling; applying alphabetic transcription metrics directly to the task of number transcription may not accurately simulate the process.

In addition, this thesis also highlights the effect of familiarity on number transcription. Familiar numbers are typed faster than non-familiar numbers and, in a decision task, familiar numbers are recognised significantly faster. These results suggest that familiar numbers are represented in memory whereas non-familiar numbers are not. This result has implications for number transcription studies in which a participant is required to transcribe random numbers. Random numbers are not likely to be familiar to the participant, and so they will not be stored in memory and will be typed slower than familiar numbers. This is an important distinction in an experiment evaluating number entry within a context where users are familiar with the numbers they are entering.

1.3.2 Contribution to Methodology

Study 7 in this thesis highlighted that the numbers used during a number entry task in an applied domain do not adhere to a random distribution. This study showed that, in particular domains, novel patterns of numbers may be used during transcription tasks. This means that users in this situations are likely more familiar with the numbers they use regularly. This result verifies the importance of the finding that familiarity of a number affects how it is typed. It highlights the necessity of considering the numbers that will be used in a task when evaluating a number entry interface. This has implications for all number entry studies that involve participants transcribing random numbers.

Additionally, the sets of familiar and non-familiar numbers generated in this thesis can be used in future experiments where number familiarity is key. Whereas previous experiments using such numbers have not shared their resources (Cohen et al., 1994), the lists generated within this thesis can be used in further experiments. Despite the familiar number list being generated largely by a British population, the lessons learnt from the subsequent card-sorting activity have highlighted a set of "reasons for familiarity" that could be used to generate or seed future population-specific familiar number sets if required.

1.3.3 Contribution to Design

This thesis has shown how an understanding of number patterns used in a number entry task can affect the design of the interface. By using the knowledge of the numbers used when programming infusion pumps, two different novel interface adaptations were designed and tested in this thesis. One tailored the interface to allow for the most frequent numbers to be entered more quickly, whilst the other utilised a feature of the number set to check for number entry errors made by the user. These design interventions were based upon an understanding of the domain-specific number patterns, and have shown how beneficial that knowledge can be when applied to the design of number entry interfaces.

1.4 Summary

This thesis has highlighted the importance of understanding the familiarity and frequencies of numbers when considered in a transcription task and has shown how this can affect both the user and design of number entry interfaces. It is not enough to model number entry using randomly generated number strings: typists will not have representations of those numbers in memory and will, therefore, not type them as fast. Number entry in the certain contexts cannot be validly modelled using this technique. Number entry design and evaluation can also benefit from an understanding of the numbers entered during a number entry task in a specific domain. Incorporating the findings from this thesis can improve number entry evaluation and design by taking into account familiarity effects during number transcription, and adapting interface design to use information about the context-specific number patterns.

This thesis has established the first steps towards understanding number entry to the same extent that we do text entry, and provides a pathway for future work in the area.

Chapter 2

Motivation

2.1 Introduction

Typing numbers is a task that many of us perform every day, from entering a phone number to performing a calculation on a computer or calculator. Number transcription is also a key part of many people's jobs, from a cashier typing in the barcode of a product, to nurses reading from prescriptions to program infusion pumps. In most cases these tasks are completed without incident, but occasionally errors are made. Many errors are inconsequential and are either noticed quickly and corrected, or have no serious impact, such as dialling the wrong phone number. However, some similarly small errors can have more serious consequences. Unfortunately, number entry errors that occur in the medical domain can be severe; some of which result in the death of the patient (Institute for Safe Medication Practices Canada, 2007). In fact, it has been estimated that medical error accounts for between 44,000 to 98,000 deaths per year (Kohn et al., 2000). This statistic suggests that the interfaces that people are using, and the tasks that they are being asked to do, are not perfectly designed. The problems within number entry tasks can partially be attributed to our current lack of understanding of how numbers are transcribed. Some intervention is needed to understand more about how number transcription occurs and what can be done to prevent unnecessary errors from occurring.

This task is not trivial. Before an error can be prevented, a deep knowledge about the causes of that error is required; this means more needs to be understood about the process of number transcription itself. Previously, a large amount of research has been completed in the text transcription domain, resulting in changes being made to the way we enter text (for example Dunlop & Levine (2012); Wobbrock et al. (2003); Zhai & Kristensson (2003)). This work aims to make these interfaces faster to use and less error prone. This area attracts a lot of research attention as the technology we use changes and becomes smaller (as with the introduction of mobile phone technology, and more recently smart watches (Dunlop et al., 2014)), more easily adaptable (as with the introduction of touchscreens on devices (Kim et al., 2013)), and necessary on a diverse range of technology (for example modern, accelerometer-based gaming consoles (Jones et al., 2010)). It is currently unclear as to whether the same knowledge can be applied to the process of entering numbers. There are a number of factors that make reading and transcribing *numbers* a different task to that of reading and transcribing *words*.

Looking at occupations that involve number entry can tell us more about why the two tasks might be different. One common number entry related job is that of a data entry clerk. For jobs such as this, adverts often require the applicant to have a certain level of typing ability. This is similar to positions that will involve a large amount of alphabetic text transcription, which also ask that applicants can type at certain speeds. Reading these job advertisements provides the first suggestion that number transcription and text transcription can be treated as two separate tasks, as there appear to be differences in the ability of people to type words and numbers. Typically, it seems as though words per minute (wpm) rates of between 60-75 are requested for text transcription jobs[1], whereas the required rate for typing numbers (keys per hour or kph) is generally between 6000-10000[2]. Taking the standard of an average of 5 characters per word, and a conservative 50 minutes per hour of typing, this places the requirements for typing alphabetic characters at 15000-18750 kph, around twice that required for number entry. If text transcription and number transcription were comparable tasks, one could reasonably expect that jobs requiring a similar skill for both would ask potential employees for the same level of typing ability.

Research is required to investigate the process of number transcription. This research will explore the possibility that number transcription is a unique text entry task, one that cannot be suitably explained by the more general text entry literature. Such information would impact the design and evaluation of number entry tasks and interfaces in comparison with text entry tasks and interfaces. It is hoped that with a better knowledge of the number transcription process, actions can be taken to make the task easier and less error prone, through better understanding and design intervention.

In this chapter, some of the current work in the field of text transcription will be explored. This will relate to both number transcription and alphabetic text transcription. Prior to this review, the chapter will aim to establish firstly that numbers and words are not comparable in a typing task, and will then explore what aspects of numbers could affect how they are

[1] As taken from a search for “typist” jobs on the recruitment website www.indeed.co.uk

[2] As taken from a search for “data entry” jobs on the recruitment website www.moster.com

transcribed. This will specifically be done from a number reading perspective. The ultimate aim of this chapter is to establish what can be learnt from the alphabetic text entry domain and applied to the task of number entry, and what possible discrepancies there could be between the two and how that could inform number entry research and design.

2.2 Terminology

Before continuing, definitions of key terminology used throughout this thesis will be established. Where necessary, similar existing terminology will be highlighted.

Transcription Transcription typing is the task of copying information from one source and entering it into a device, and can apply to both alphabetic and numeric text. Transcription can apply to the process of hearing or reading the information. This thesis will focus solely on the task of reading-based transcription.

An example of a transcription task is copying a credit card number into an online shopping checkout. The task of transcription typing does not involve entering information that has previously been memorised. An example of such a task is recalling and entering a PIN at an ATM. This distinction is made, as research has shown that performance during recollection and transcription tasks varies (Kristensson & Vertanen, 2012).

Text Throughout this thesis, a distinction will be made between alphanumeric, alphabetic and numeric text. Previous research however often refers simply to “text”, which often means just alphabetic text. If no clarification is used when referring to previous research, alphabetic text should be assumed.

Language The assumed language throughout this thesis is English. This is important when the phrase non-word is used, this refers to a non-word in the English language.

Digit A digit is a basic building block for a number. The numbers 0 - 9 are digits. During this thesis, the decimal point is also considered a digit, as it is often typed when copying numbers.

Number A number is a whole entity, and may be comprised of one or many digits. For instance, the number 543 is comprised of three digits: 5, 4 and 3. Unless specified otherwise, it is assumed that the number is well-formed and adheres to rules such as numbers do not contain more than one decimal place. This rule is broken in some contexts (such as IP addresses for instance) but in this thesis this situation is not considered.

Unless otherwise stated, a number is assumed to be an “Arabic Number”, which are the number characters used in English-speaking countries. There are alternative numbering systems in use in other languages. Another numbering system which may be familiar to the reader is the Roman numeral system, which represents amounts as letters (I for 1, V for 5 and so on). This system requires different parsing to the arabic numeral system and is therefore not considered in this thesis.

Letter A letter is a basic component of a word. In the English language, there are 26 possible letters (the alphabet). Each letter in the English language has two forms: lower case and upper case. The case of letters in this thesis will not be important so it can be assumed that there are 26 possible letter choices.

Word A word is comprised of one or many letters. For instance the word “a” is both a word and a letter. The word “cat” in comprised of three letters: ‘c’, ‘a’ and ‘t’.

Legal/Illegal Non-word A non-word is a word that is not recognised as belonging officially to a language. A legal non-word is a non-word that still adheres to the orthographic rules of that language. For instance, the word “phar” is a legal non-word in the English langauge: each of the letters within it are placed in an orthographically correct order. This results in legal non-words being pronounceable.

An illegal non-word does not adhere to the orthographic rules of a language. For instance, the word “kjre” is an illegal non-word: it is not possible for a word to begin ‘kj-’ in the English language. Often illegal non-words are non-pronounceable.

Character This describes both digits and letters. It denotes a basic building block that can be any form of alphanumeric character. This definition of character includes the decimal point.

2.3 Reading numbers and words

The transcription typing process involves three key steps: reading, memorising and motor action. There is currently little research looking specifically at the process of number transcription, however, there is more research within the sub-tasks of transcription, namely addressing task of number reading. Any aspects of numbers that affect how they are read, will also affect how they are typed in a transcription task. This section aims to establish the similarities and differences between number reading and word reading.

2.3.1 Studies with aphasic patients

Most healthy adults have no problems when reading text and numbers. However, people affected with aphasia will often struggle with language. Aphasia is a neurological condition caused by brain damage resulting in impairments relating to language understanding and can manifest itself in a lack of ability to speak or a reduced ability to read (Damasio, 1992).

The symptoms of aphasia for patients can vary in both severity and the linguistic abilities that they effect. Investigating how linguistic ability breaks down in aphasic patients can highlight the different cognitive pathways used when reading.

Delazer & Bartha conducted a review of case studies with aphasic patients who experienced problems with numerical tasks such as reading aloud and performing calculations (Delazer & Bartha, 2001). From the review, differences between word and number reading strategies can be deduced. Firstly, some aphasic patients had problems with reading numbers aloud, whilst maintaining their ability to read words aloud (Cipolotti & Butterworth, 1995), whilst others were more accurate at reading aloud numbers compared to words (Thioux, Pillon, Samson, de Partz, Noël, & Seron, 1998). Thioux et al. suggest that this differing ability to read words and numbers suggests that the two are processed in different areas of the brain. It is outside the scope of this thesis to investigate brain regions associated with language processing, but these case studies do suggest that number reading and word reading are separate cognitive tasks, as one can be impaired without affecting the other. The implication of this theory on numerical transcription typing is that it is not necessarily suitable to apply research from the alphabetic text transcription domain directly to numbers, without first ensuring that this different reading process has no effect.

The aphasic research does also show that there are similarities between the way that words and numbers are read. This is specifically the case for words and non-words. Some cases of aphasia prevent the patient from being able to read non-words, whilst maintaining their ability to read words (Patterson & Marcel, 1977; Noel & Seron, 1993). This impairment highlights the importance of long term memory representations of words during word reading tasks. Without a representation in memory, non-words cannot be read by aphasic patients who have impairments in phonological production (converting the graphical representation of a word into the sound).

Initially, it may not seem that there is a parallel to this finding with regard to number reading: there is no “non-number” equivalent to a non-word. However, case studies with other patients have shown a distinction between “meaningful” or “familiar” numbers and “meaningless” or “non-familiar numbers”. It is reported in Cohen et al. (1994) that one patient was able to read meaningful numbers, but lost the ability to read meaningless numbers. This refers to the ability to read numbers that had some important significance (for instance 1789, the year of the French Revolution, a significant date for a French patient) was preserved in the patient who otherwise was unable to read numerals. This was similarly the case for the patient when asked to read words and non-words: the patient was able to read words but could not read non-words.

Such a finding suggests that it might be possible to think of non-familiar numbers in a similar way to non-words in terms of how they are read. Delazer & Girelli (1997) tested the patient further and presented him with a list of semantically familiar numbers. The numbers had both autobiographical and encyclopaedic meaning. When presented as a list of numbers, the patient read only 33% correctly. Later, when the patient was told that each number had a meaning, and was encouraged to consider that meaning before reading the number aloud, the patient's reading performance increased to 93%. Just as words were facilitated by their representation in long term memory, it appears a similar effect is noted with familiar numbers: when the patient tried to access the number in memory, they were able to read the numbers that previously they could not. This would not be possible for non-familiar numbers, as there would be no stored representation in memory.

These studies have firstly made a case for exploring the differences between word and number transcription, the two may not be directly comparable. Secondly, they highlight potential reading differences within numbers with regard to familiarity, which suggests that careful attention should be paid to the types of numbers used during number transcription tasks and number entry evaluation.

2.3.2 Models of number reading

Cohen et al. (1994) aim to explain this difference in reading familiar and non-familiar numbers by proposing that there are two different ways to process numbers: a semantic and a non-semantic route (Figure 2.1). Their model of number reading suggests that numbers can be read and recited via the visual-phonological route as long as the number is well-formed. This route does not involve any understanding of the number's magnitude or meaning. This route uses lexical rules to translate the digits to the words they are associated with; for instance reading ‘1945’ as “one thousand nine hundred and forty five”.

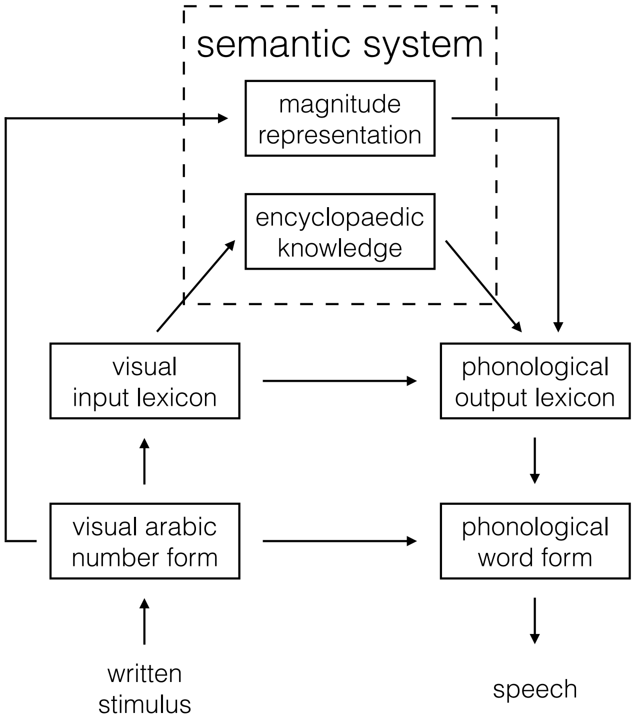


Figure 2.1: Cohen et al's model of arabic number reading. The multiple routes of number processing can be seen, with the semantic route at the top of the model and non-semantic in the middle. (Cohen et al., 1994)

It is possible for a reader to process a number on one semantic level, which involves understanding the magnitude of the number, for instance reading '1945' and knowing it is odd, or that it is larger than 1000 and smaller than 2000. This involves a different route, with the number passing from visual input, then being converted to a magnitude representation, and then to the phonological output.

A second form of semantic reading is also possible. This involves an encyclopaedic look-up of the number being read, which returns information about the number's meaning beyond the magnitude information. Using the current example that would mean reading the number '1945' and knowing that this could represent a year, and that the year relates to the ending of the Second World War. This involves a different reading route, this time via encyclopaedic knowledge. Most numbers would not have any representation in memory, and would therefore not be able to be read using this semantic route.

This model has built upon previous models of number reading that omitted any semantic or encyclopaedic routes of processing, a review of these models can be found in Cohen et al. (1994). These models, however, could not account for aphasic patients, such as the one in the Delazer & Girelli (1997), who were able to read meaningful numbers and not others. This model suggests that these patients lost their ability to transfer numbers directly between the visual input and the phonological output, without facilitation from a long term memory representation of the number. Numbers that could successfully be found during an encyclopaedic look-up were able to avoid the faulty connection between visual input and phonological output.

2.3.3 Models of word reading

Models of number reading are largely based upon studies with aphasic patients. Research into the effects of familiarity of number on reading are limited to such studies, and are difficult to test empirically. In the word reading domain however, there is a larger body of work which is based on experiments with healthy participants. It might be possible to apply results from the word research to our understanding of numbers by comparing how words and non-words are read to familiar and non-familiar numbers.

Words and non-words, unlike familiarity of a number, are objectively defined concepts. A word is contained within a language's dictionary, and a non-word is not (with some exceptions, such as places or names for instance). To inform a model of word reading, many empirical tests have been performed using lexical decision tasks (Meyer & Schvaneveldt, 1971).

This task is chosen as it is a simple experiment, and it is possible to draw comparisons between how words are read within the task, to how numbers might be read. The basic lexical decision task requires a participant to respond to a target stimulus. The participant must then decide whether that target is a word or non-word and then respond accordingly. Early research shows that participants are faster when classifying words than non-words (Meyer & Schvaneveldt, 1971). There are a number of theories about what causes this effect, they are explored in Ratcliff, Gomez, & McKoon (2004). Here, a selection of these theories will be presented, and their applicability to number reading will be considered.

One of the earliest explanations of this finding is that when a word is read, a serial search is done through an internal "dictionary", or lexicon, of known words, stored in long term memory (Rubenstein et al., 1971). The internal lexicon of words is scanned and the input word is compared to entries that are similar orthographically. When a match is found, a word response can be given. Only after the search has been exhausted, and no match found, can a non-word response be given. The non-word response relies on a search failure and so causes this decision to be slower than a successful word response.

The activation-level account for this decision task however, relies upon representations of the words in long term memory (Paap et al., 1982). As the stimuli are read, aspects of the input will activate similar words in memory. Similarity might refer to shared letter patterns or similar phonological representations. These activation levels then need to reach a threshold, where a 'word' decision can be made; the highest levels of activation would occur when there was a match between the input word and a word stored in long term memory. A non-word decision would need to wait a designated period of time to ensure that the threshold activation would not be reached. The necessity for this cut-off point explains the slower response times for non-words.

These two accounts rely upon internal representation of words within memory. This has already been established as a potential cause for different reading abilities for familiar and non-familiar numbers in aphasic patients. Familiar numbers have some internal representation, whereas non-familiar numbers do not. In a decision task for familiar and non-familiar numbers, it might be expected that, as with words, numbers stored in memory (familiar) would elicit faster response times than those not in memory (non-familiar).

The final model discussed however, does not rely upon internal representation of words. The diffusion model suggested by Ratcliff et al. (2004) suggests that when a stimuli is read, the decision is based upon a series of pieces of information from this input. These small pieces of information affect the drift towards two boundaries, representing the two possible decisions within a task: word or non-word. Once enough information has accumulated, the decision drift will have reached a boundary, and a response can be made. To explain the slower response rate to non-word stimuli compared to word stimuli, the diffusion model suggests that there is confounding evidence that a non-word is a word. When a non-word is read, some of the information it provides suggest that it could be a word: it is made up of letters, some of the letters might be ordered in orthographically legal ways. This confounding evidence causes the drift to incorrectly move towards a 'word' decision. Eventually, enough evidence is gathered to move the drift towards the non-word boundary, and the correct response is given. This initial misinformation causes the decision to take a longer time.

This model relies upon features of a target stimuli that may give it word-like features. These features include adherence to well-defined orthographical rules. There are no such rules for numbers, where any arrangement of digits can constitute a legal number. It is difficult therefore, to understand how a number decision task could be explained by the diffusion model. This model would predict that both familiar and non-familiar number responses would take a similar amount of time.

2.3.4 Summary

This section has highlighted that words and numbers can be read using different routes. As reading is a step within the transcription task, this finding could impact on the way that numbers are typed, and suggests that text entry cannot be generalised to numbers without further research and justification. Additionally, different reading routes within numbers were highlighted, suggesting that familiarity of a number may also have an effect upon how it is typed. This is not something that current number entry studies consider, as most rely upon randomly generated numbers during transcription (Oladimeji et al., 2011; Thimbleby & Cairns, 2010; Wiseman et al., 2011). This section has suggested it is not enough to understand that numbers and words may be transcribed differently, but that number entry tasks may also be affected by the familiarity of number.

2.4 Text Transcription

With potential differences between the reading routes of words and numbers established, and further suggestions that familiarity of a number may also affect typing, the implication on current text transcription research is explored. This section will firstly investigate what research there is currently in the domain of number transcription. After this, the more well-researched domain of alphabetic text transcription is explored, and its applicability to number transcription evaluated.

2.4.1 Number Transcription

There is currently very little research aimed at understanding number entry from a theoretical, rather than applied approach. The current work in this area has focussed on understanding the process of transcribing number from an auditory source, rather than a written source, as is the focus of this thesis.

Lin & Wu (2011) aim to understand the effects that speed of auditory input can have upon the number transcription task. In their study, the researchers conducted a laboratory experiment aiming to understand the factors affecting numerical hear-and-type tasks. In these tasks, participants were told to listen to and type sets of 9 digit numbers that were read aloud to them. The factors manipulated in this task were the speed of the numbers being read out, the urgency (whether or not they were penalised by a slow or wrong response) and the typing strategy (whether they used one or two fingers). The authors point out that speed has an effect upon the accuracy of performing a task, and that this effect has not yet been investigated in a numerical typing task. The results of the study report the following: errors are increased by urgency, using multiple fingers to type and from hearing a fast reading pace. Additionally, the fast paced presentation of numbers resulted in a longer reaction time when typing.

These findings are important for consideration in future number entry studies. However, there are issues with the methodology when applying them more broadly to the task of number entry. Firstly, these results are not generalisable to other number transcription tasks, namely those that involve reading numbers. Speed of input is not often a consideration when the typist reads numbers, as they can control the speed that numbers are presented to them.

Another query about the generalisability of this study is the fact that not all numbers are read out in this way: as a series of digits. Depending upon the number's meaning, the way that it is read out can vary. Telephone numbers for example, are numbers used as unique character identifiers, and no more than that and for this reason, these numbers are read out as a string of digits. However, numbers that represent something, such as a monetary value, are more likely read as words. This experiment does not allow for these two different strategies to be used.

A final concern that has not been addressed in this work is the importance of familiarity of number. This study was conducted with random numbers, but this is not necessarily representative of number transcription tasks as a whole. As has been discussed in the previous section, the meaning behind a number can affect the way it is read. By presenting participants in these studies with random numbers, the possible effects of familiarity are not explored. If one is aiming to fully understand the number transcription process, this phenomenon must first be investigated and then applied to the model.

This study represents the current research within the field of number transcription. The findings have proven interesting, but not generalisable. Further research is required into understanding number transcription in a reading-transcription task.

2.4.2 Alphabetic Text Transcription

The research into numerical transcription typing is, at present, minimal. It may be beneficial to look at alphabetic text transcription research to learn more about the process. The studies involving aphasic patients showed that, although word and number reading involve different processes, the two are comparable in terms of how long term memory representations can affect reading. It is possible that words may be similar to familiar numbers, and non-words similar to non-familiar numbers in terms of memory representation. In this section, current alphabetic text transcription research is reported and possible translations to number transcription are explored.

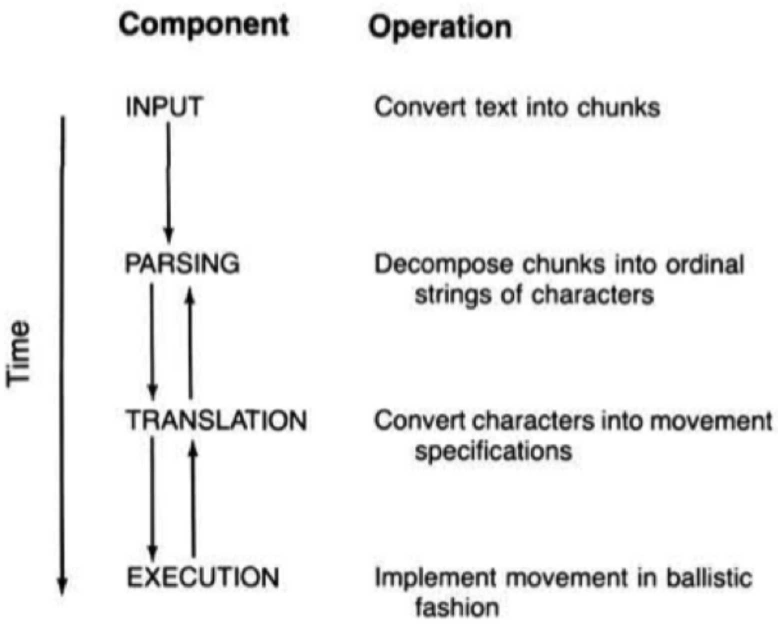

Component
Operation
Time
INPUT
Convert text into chunks
PARSING
Decompose chunks into ordinal strings of characters
TRANSLATION
Convert characters into movement specifications
EXECUTION
Implement movement in ballistic fashion

Figure 2.2: The four components of transcription typing (Salthouse, 1986)

2.4.3 The process of transcription typing

One of the seminal papers in the field of transcription typing research is “Perceptual, Cognitive and Motoric Aspects of Transcription Typing” (Salthouse, 1986). In this review article, Salthouse gathers and summarises the text typing research in the field forming a set of typing ‘phenomena’. These phenomena are now important when attempting to model the process of text transcription and are accepted as key benchmarks in transcription typing (Wu & Liu, 2008; John, 1996).

Salthouse defined a basic model to describe the process of transcription typing. The components of Salthouse’s model (Salthouse, 1984) consist of 4 categories of processing: input, parsing, translation and execution. Figure 2.2 shows the interaction between each of the processes and an example operation from each component. This four-stage model of text transcription is not unlike the Model Human Processor (MHP), which separates actions into 3 stages: the perceptual, cognitive and motor system (Card, Newell, & Moran, 1983).

Salthouse’s input component relates to the perceptual task of reading the input material and can be likened to the MHP’s perceptual stage. The parsing stage in Salthouse’s typing model incorporates the cognitive elements of transcription typing; that is memorising the text that will be typed, which is similar to the cognitive stage of the MHP. Salthouse’s translation stage is the main module that is not represented in the MHP. Translation relates to the cognitive process of planning where the fingers should move to on the keyboard. The necessity of the ‘Translation’ step in the Salthouse model has been questioned by some (Rieger, 2004), who believe that automatic responses are possible directly between the ‘Parsing’ and ‘Execution’ modules, without the need for mediation in the ‘Translation’ module. This may be especially true now that typing is an activity that many in the population are highly skilled at; it seems likely that automatic connections are possible between letters and their position on the keyboard without the need for explicit translation. For this reason, we might group both translation and execution together and liken it to the MHP by calling this the motor module. In this way, Salthouse’s model can be discussed in terms of the more widely used 3-stage MHP. Despite its construction in the 1980s, Salthouse’s model of transcription typing has remained a reliable model that is still used to describe transcription typing. Specific elements of the model have been replicated in experiments (Wu & Liu, 2008; John, 1996; Bosman, 1993).

From an understanding of this description of the model, it is possible that the steps will also relate to the task of number transcription: a user performing number transcription must look at the target numbers, remember them, and then type them as described in Salthouse’s model. This, therefore, makes the model a suitable starting point for understanding more about number transcription. It provides a framework with which to structure the comparison between text and number transcription.

As we have seen in the previous section, there are clear and documented differences in the way in which we read words and numbers. Reading as a step in the transcription process occurs in the perception module. For this reason, this module of the text transcription model will be explored, deciding how it may apply to number entry.

Firstly, it is important to understand how the modules work together. Although the modules within the model appear separated, Salthouse highlights the parallelism between the modules. It is not the case that one must first read some text, memorise it and then type it as a serial process, it is possible to both read and type simultaneously. The relationship between the components of the model has been explored through a number of different ‘spans’ that are notable when someone performs transcription typing (Salthouse & Saults, 1987). A span represents an amount of information, measured in characters, that a typist is able to process.

The four spans that will be considered in this thesis are as follows: detection span, copy span, eye-hand span, and replacement span. All spans are measured in terms of the number of characters used. These spans are concerned with the perception or parsing module, and thus are concerned with the way that input text is read and stored in memory. A fifth span, the stopping span, is also reported by Salthouse & Saults. This span will not be considered further as it does not relate to the perception or parsing modules and thus does not relate

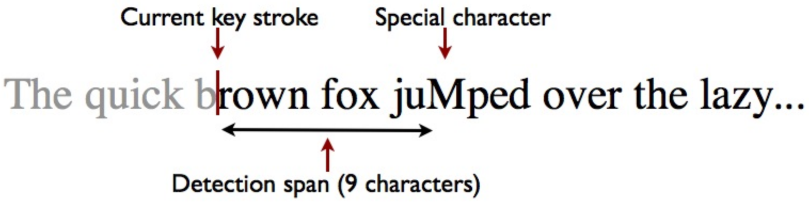

Current key stroke
Special character
The quick brown fox juMped over the lazy...
Detection span (9 characters)

Figure 2.3: Illustration of the detection span. The typist is about to type the character 'r' and has just noticed the special character. Meaning this typist has a detection span of 9.

to the way that numbers are read. This is an important distinction as the motivation for this work relies upon the differences between how words and numbers are read. The spans are separated depending upon which module from the Salthouse transcription typing model they occur. They will be discussed in more detail in Chapter 4.

Salthouse Input Module

Two spans are determined by the input module, the detection span and the copy span. The detection span is defined as the point at which the typist is able to detect a special character in the input text. This shows how far ahead the eyes look when typing. This is tested by asking the typists to signal when they have seen the special character in the text that they are typing from. The detection span is the number of characters between the character being typed and the placement of the special character. See Figure 2.3 for an example. Although typists will be focusing on the characters that require typing immediately, it is still possible for the typist to look ahead at the rest of the text; the detection span illustrates just how far ahead this distance is. The detection span for words is reported to be around eight characters (John, 1996).

The second span associated with the input module is the copy span. This is the measure of how many characters the typist is able to type when the source material is unexpectedly removed. This span shows how much of the source material the typist stores in memory when typing. The copy span ranges from two to eight words or about seven to 40 characters.

Although this span is undoubtedly affected by the typist's working memory, it is not a direct measure of it. This typing task does not test the typist's ability to memorise a list of words, but instead tests how many words the typist chooses to memorise whilst typing. This distinction is the difference between the typist actively trying to memorise the words and the typist naturally storing words in memory as they type. It is likely that, if the typists were asked specifically to memorise the words, they would have been able to store more in working memory.

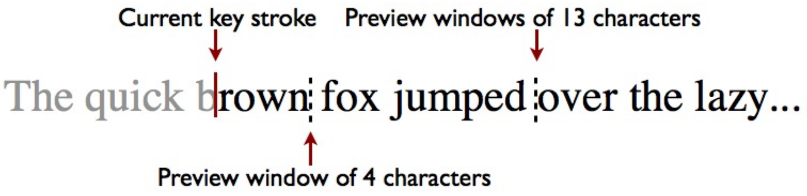

Current key stroke
Preview windows of 13 characters
The quick brown fox jumped over the lazy...
Preview window of 4 characters

Figure 2.4: Illustration of the eye-hand span. If the preview window is reduced to 4 characters (in this situation the typist would not be able to see past the word "brown") which is smaller than the desired 5 character eye-hand span, the typist will slow down and will not be able to type at optimal speeds. With a larger preview window of 13 characters (the typist could now see the word "jumped") the typist can see further ahead and can therefore type at the optimal speed.

Salthouse Parsing Module

The parsing module is responsible for the eye-hand span and replacement span. The eye-hand span represents the minimum 'preview window' of the input material that allows the typist to type at their normal rate. When the preview of the input text being copied becomes smaller than (on average) five characters, typists slow down their typing speed. An example of the eye-hand span can be seen in Figure 2.4. This span suggests that the components of the transcription model are indeed working in parallel as the input module is able, and indeed required, to run ahead of the execution module in order for typing to occur at optimum levels.

One particular aspect of the eye-hand span makes it particularly interesting when considering how these typing spans might be affected for numbers. The eye-hand span is larger for words than it is for non-words (Salthouse, 1984; Hershman & Hillix, 1965). This difference is again thought to be due to words having representation in memory whilst non-words do not. When considering how this span may vary for numbers, it is possible that the eye-hand span for familiar numbers could be greater than that of non-familiar numbers, making this an important tool to explore the affect of familiarity on number transcription.

The final of the four spans considered is the replacement span. This span investigates the point at which a typist commits to the text that they have read. This is measured as the point at which a character in the input text can be replaced for another without the typist noticing. This represents the moment when the parsing module passes the word to the execution module. The replacement span is typically around three characters; a character two ahead of the one that has just been typed could be replaced and the typist would not make the change. See Figure 2.5 for an example of the replacement span.

With the knowledge of these spans, it is possible to assess what processing might be occurring when transcription typing takes place. We know that people are aware of input material up

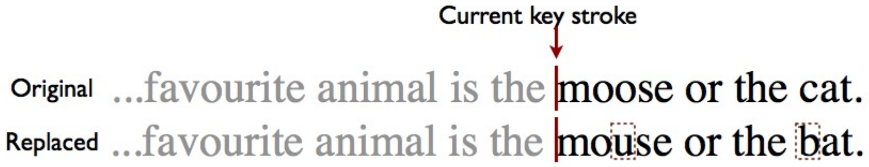

Current key stroke
Original ...favourite animal is the moose or the cat.
Replaced ...favourite animal is the mouse or the bat.

Figure 2.5: Illustration of the replacement span. The typist first sees the original text, but then one character is changed. If that character is changed within the replacement span (the 'u' in this case) then the typist will still type the old character (typing 'o'). If the change is made outside of the replacement span (replacing the 'c') then the typist will type the new letter (typing 'b').

to eight characters in advance in some circumstances, we are shown this from the detection span. When the input text is suddenly removed and the input module can provide no more additional information, we know from the copy span that around six to seven characters are stored in the input module. Additionally, the eye-hand span tells us that character information is required five characters before execution, any less and typing slows down. This provides some idea of the size of the parsing module. The replacement span of 3 characters shows the point at which the input buffer relinquishes information to the execution buffer. If the input module was still taking note of characters up to the next character to be typed, then the replaced letter would be noted and altered.

Text transcription and timing

One other notable phenomenon in Salthouse's review is how typing speed is affected by varying levels of sense in the input material. Salthouse states that in many prior experiments, word order has very little effect on our ability to copy and type text. However, as the words themselves approach random mixes of letters, the speed of transcription typing sharply decreases. This is thought to be due to inefficiencies when copying text that has no representation in long term memory. Words can be read and remembered quickly because they can be memorised as whole words. Non-words however, must be memorised as smaller chunks, because they have no prior representation in memory.

This effect can also be seen when typists transcribe different languages. In one study, participants whose first language was Finnish, but who were also proficient in English, were asked to transcribe text from both languages (Isokoski & Linden, 2004). Transcription in English, the less familiar language, was 16% slower than transcription in Finnish. Applying this effect to numbers, in terms of familiarity, most numbers will be treated as though they are randomly jumbled letters, as most numbers do not have meaning for us and are not available in long term memory. Therefore, we might expect typists' ability to transcribe numbers, to look similar to the performance of typing words comprised of randomly arranged letters; numbers will be transcribed slower that text.

Application of the Salthouse typing model to number transcription

The spans reported by Salthouse relate specifically to the typing of alphabetic text. The possible transcription spans for number typing will now be explored.

Some of the typing spans are affected by the predictability of the input as a whole, that is, how predictable the set of words are and not just the internal structure of the words. The copy span is extended if the input material is ordinary prose, rather than randomly arranged words (Salthouse, 1985). This is because words taken from prose have some link between them and the text has some meaning behind it. For instance, if you were to read the words "the piano was very heavy" compared to the words "cat macro gap near serve", you are likely to remember more from the first list, as there is predictability between the words. With numbers however, there is often little one can do to predict what number may come next and thus the copy span is likely to be smaller for numbers than for alphabetic text. It may be comparable to the copy span seen when transcribing random words, which is smaller.

However, there is also internal predictability within words; often it is possible to predict the end of a word having only seen the first few characters. If one is transcribing four letter words and has seen the first few characters of "bal-", it may be easy to predict the ending as "ball". However, with numbers, if one is transcribing four digit numbers and reads "586-", it is likely, without context, that one could not predict the end digit accurately. In some circumstances, prediction may be possible. For instance, if a person was asked to transcribe a set of prices taken from a shop, and saw "£12.9-", it might be possible to predict that the last digit is a 9, or perhaps a 5. However, this type of predictability is rare for numbers. In this way, the copy span may be even further reduced for numbers, and may be more like the copy span for non-words that have been created from randomly chosen letters. We have already seen in the previous section that numbers and non-words are potentially read using different processes to those used when reading words. However we also saw how the familiarity of a number can mean it is treated differently to a non-familiar number, because it is associated with some semantic knowledge. For this reason, we might predict that the copy span for familiar numbers is greater than that for non-familiar numbers.

The replacement span may also be affected by the predictability of the input. If the typist is confident of the word being read, they are less likely to double-check the word before typing it. If they have read it, and it agrees with the orthographic rules they know about how words are formed, then there is no need to ensure it has been read correctly. Most numbers however, do not have such predictability and therefore may need to be checked more often, meaning the replacement span will be smaller. However, as before, familiar numbers may be read more like words and will have more internal predictability (for example, when seeing the six digits that represent your date of birth). These numbers may therefore have a replacement span closer to that of text.

The eye-hand span is indicative of how quickly the input information is being processed by the parsing module. We know that non-words have a smaller eye-hand span than real words; it is thought that the cause of this is the internal predictability of words compared to non-words (John, 1996). For this reason we might expect that the eye-hand span of numbers will also be smaller. Equally, if a person were to type familiar numbers, we might expect there to be an increase in the eye-hand span.

The detection span is more difficult to predict. It is likely that the span will also be affected by the familiarity of the numbers in much the same way as the replacement span. However, the detection span is often incredibly variable (Salthouse & Saults, 1987) and so it is difficult to know how this will change if the input material were numbers rather than words. Additionally, it is difficult to know how the detection span could be tested for numbers. In word-based studies, typists are asked to detect a capitalisation of a letter. This not only allows for the word to be read and copied as normal, but also acts as a visual flag, as capitalised letters stand out amongst lower case letters. There is no such equivalent for numbers and detection span experiments would have to substitute a different character into the number for detection. This would compromise the integrity of the number and change how it was read. For these reasons, the detection span will not be investigated further in this thesis.

2.4.4 Summary

These spans offer a validated series of metrics that, when tested with words, have provided an understanding into the processes involved in alphabetic transcription typing. In order to know more about the process of numerical transcription typing, initial hypotheses have been formed that explore the possible implications that numbers could have upon these spans. These hypotheses will be expanded and span results from typing experiments will be used to provide comparative data points for number transcription compared with word transcription.

These spans can also provide additional information about the differences in numbers themselves, that is, between familiar and non-familiar numbers. Typing speed has shown how differences in the memory representation of words and non-words can be measured. The Salthouse phenomena present themselves as a reliable method for testing cognitive processing during text transcription; for this reason they will be used to understand more about number transcription.

2.5 Modelling Transcription

In addition to the five spans that Salthouse identifies, he also lists a number of other typing “phenomena”. These phenomena include predictions about error rates, speed of inter-key intervals, alternate hand typing speed and so forth (see Salthouse (1986) for a full review). These phenomena not only provide a way to describe transcription typing, but also provide a set of constraints for any computational model that attempts to mimic the process of transcription typing.

In his 1990 book “Unified Theories of Cognition”, Newell describes transcription typing as an ideal candidate for cognitive modelling: it is a well-defined task and thanks to work by Salthouse and others, there are clear phenomena that must be simulated by any potential model. There have been a number of cognitive computational models that aim to model transcription typing. It is possible that these models may offer some insight into the process of number transcription.

2.5.1 TYPIST Model of transcription typing

One notable model is the TYPIST model (John, 1996). The **T**heor**Y** of **P**erformance **I**n **S**killed **T**yping was modelled using the Model Human Processor framework (Card et al., 1983). The aim of this work is to create a usable model that would recreate the Salthouse phenomena in an easily readable manner that meant it could be used in conjunction with other tasks, not solely typing. For this reason, the model was made necessarily simplistic, aiming to recreate realistic typing practices, without aiming to explain in much detail where these behaviours arose and what caused them.

The model is based upon assumed timings for each step in the transcription process, that is, reading a word and encoding it takes 340 msec, using the cognitive processor takes 50 msec and typing a character takes around 230 msec. These hard-coded timings produce fair results but, due to their use, mean that certain Salthouse phenomena cannot be recreated. For example, these timings do not take into account commonality of characters and digrams (a pair of two letters), which Salthouse identifies as being key to certain inter-key timings: the more practice we get with common digrams, the better we become at typing them. For instance, we are far more used to typing ‘qu’ than we are ‘qg’, which does not occur in the English language; the hard-coded timings used in this model are unable to simulate such details. The model also requires small timing ‘buffers’ to be added in certain places to pad out the timings. These timing buffers are not justified with any cognitive reasoning, they are merely inserted in order to make the results more accurately match the Salthouse phenomena.

However, as stated previously, these hard-coded timings do come close to predicting many of the Salthouse phenomena. This begs the question of what the hard-coded timings would look like if this model were to be applied to the task of number transcription. It appears as though numbers may be read differently to words and so the timings may need to be altered. This model may actually match number transcription more accurately than it does word transcription as, depending on the specific domain the numbers are taken from, the numbers may not have ‘common digrams’, and thus this simple timing may be accurate. This model would predict that the only factor affecting the time taken to transcribe a number is its length.

The TYPIST model processes text information using chunks. Chunks can vary in both size and meaningfulness: from words, to syllables, to characters. The chunks are used when reading the input text and converting it into to-be-typed units. Chunks are highly related to representation of text in long term memory; if the words being typed are real words, then the material is read at a word-by-word level; if they are non-words and not available in memory, then the reading might break down to a syllable level; if they are unreadable then the chunks become character sized.

These chunks dictate how fast the typing is completed. Chunking explains the Salthouse phenomenon of words being quicker to type than non-words. Only a certain number of chunks can pass through the model at any one time, this is regardless of the size of chunk. Larger chunks, therefore, are more efficient. If the chunks are at a word level then many words can be read and typed quickly; at smaller levels, less information is passed through the model and so the process slows. This simulates the slowing that occurs when typing non-words compared to typing words.

The chunks are also affected by preview size; if only a few characters at a time are available, the typist is required to chunk words into smaller amounts as whole words may not fit in the preview. These smaller chunks again account for the slowing of typing when preview windows are reduced in size. This is also John’s explanation of the eye-hand span.

It is not clear currently how typists chunk numbers. It may be that whole numbers are chunked in one go, but it also may be the case that numbers are broken up into digits, which become chunks. For example, the number 3565 may be read by some as four separate chunks, “three”, “five”, “six” and “five”. But for some people, this number may be recognisable, for instance it may be a PIN. This means the number is stored in long term memory, and is thus read as a single chunk.

Another aspect of the TYPIST model that may show that numbers are transcribed in a different way to words, is the observation regarding the eye-hand span. In the model, it is necessary to allow the typist to see the entire word they are about to type, before they encode it. That is, if only the start of a word is visible, the typist won’t encode that word until the entire word is in the preview window, hence the necessity for the eyes to look ahead of the character currently being typed. The typist waits for the end of the word before encoding it as a chunk because a memory check will be performed on the word once it is encoded; checking whether the word is a known word. However, as stated before, numbers do not adhere to the same strict orthographical rules as words and so there are infinite possible numbers. Only a few key familiar numbers will have meaning enough to be stored in memory. If it is so unlikely that a number will have been stored in memory then this memory look up may not be necessary and so the typist won’t need to wait until the entire number is in view before encoding. If the typist is chunking numbers at a digit level, rather than a number level, then they may not need to preview the entire number before beginning typing and so the eye-hand span may be much smaller than when transcribing words.

Ultimately, the TYPIST model makes a good attempt at modelling Salthouse’s typist phenomena but in the end falls short of the level of detail required to accurately simulate all of them. It can say nothing of the errors that Salthouse observed, nor can it make any predictions based upon the frequency of words and digrams. It also does not explain in a great level of detail what processes are occurring to cause these phenomena. However, the details it does explore regarding the memory look-ups per word and chunking levels have interesting implications for number transcription that may provide some insight into the possible differences between number and text transcription.

2.5.2 Queuing Network Model of transcription typing

An alternative, and more recent model of transcription typing is the Queuing Network Model (QNM) of transcription typing Wu & Liu (2008). Unlike the previous model, this one uses the mathematical modelling method of queuing networks and provides a large amount of detail about the processes within the model. This model consists of a series of servers, each responsible for different tasks within three different categories: the perceptual subnetwork, the cognitive subnetwork and the motor subnetwork. As with the TYPIST model, this model was verified using the Salthouse phenomena. However, unlike the TYPIST model, the QNM of transcription typing is able to recreate all but two of the Salthouse phenomena, including errors and digram frequency related phenomena. It also takes into account motor constraints.

As with the previous model, the QNM also demonstrates that the eye-hand span is related, not to information flow as Salthouse suggested, but to how words are broken up into chunks. This model suggests that a smaller eye-hand span results when common digrams are broken up, making typing slower. We know that common digrams are typed faster than infrequent ones, causing a speeding up effect when typing. Like TYPIST, this theory suggests that as numbers do not have the same digram frequency effects, the eye-hand span for number transcription may be smaller than that for word transcription.

Digrams are an important aspect in this model, and determine how quickly the model is able to read and type the letters. The digram frequency information for the QNM is learnt as it reads the source material and not from any prior knowledge about digram frequency in the English language. This means it may be a suitable fit for number transcription. The QNM is a learning model; all the information it has, it receives not from hard-coding but from training on the data. For this reason, the digram frequency information is naturally acquired from the input material. If the model were to be trained on a number entry task, then any natural patterns appearing in the numbers would also be learnt. Patterns in the number data would result in similar increases in speed, as seen with letter digrams.

This model is the closest to recreating Salthouse's transcription typing phenomena and goes some way to explaining the reasoning behind certain behaviours in transcription typing. There are interesting possible applications of this model to number transcription and the model makes clear that number digram frequency would be an important factor in training the model. When combined with empirical results from experiments with participants, this model could be verified and adapted as necessary to the number transcription task.

2.5.3 A model of numerical transcription typing

The QNM model of transcription typing has also been applied to the task of number transcription (Lin & Wu, 2011). However, this model is based upon the auditory transcription task reported previously. This model of number transcription therefore simulates verbally presented random digit string transcription, which is a subset of the larger number entry task. Using the results from a participant-based experiment, Lin and Wu were able to adapt their QNM model to simulate the number transcription task. The model was successful at simulating the users' typing speeds but was slightly less accurate when predicting error rates. Throughout the paper, the authors speak of the motor causes of error (over shooting and no visual checking) but do not consider any other cause of error, for instance, those that occur when memorising (Wiseman et al., 2011). They acknowledge this issue, suggesting that memory may have an effect upon the types of error made in this experiment, but do not explore it.

The transcription model aims to replicate the phenomena in Salthouse (1986) without consideration as to whether results found from the text transcription domain can be directly applied to a number entry task. In addition to this, the task itself is very different to the task that the Salthouse phenomena are based upon. The distinction between a task where text is read, compared with a task where text is heard, may alter the expected values of various spans. Previous models have shown how important chunking methods are to the typing metrics; in an auditory transcription task, the chunking is not controlled by the typist, but by the aural presentation of the numbers. Indeed, in the experiment that informed this model, the numbers are presented in blocks of three, forcing the typists to chunk the numbers accordingly. This may artificially have an impact on any typing spans measured in this experiment.

Although the accuracy levels when predicting error rates are not perfect, the model does have useful insights into the way that tasks and interfaces should be designed. The model is also able to generate an "optimal" voice speed for reading out digits to reduce errors, yet not slow the task down by significant amounts. This is an example of how a model can have real world applications, whilst still contributing to our knowledge of the potential cognitive processes associated with tasks such as number transcription.

This work overall is useful in the field of number transcription. The QN MHP model is detailed and has basis in neurology. It is testable with lab experiments and, when applied to real world problems, can make recommendations about safer design of interfaces. However, the work is not as generally applicable to number transcription as it is presented to be. The model only represents the process of hearing numbers, and not transcribing numbers through sight. The model shows that investigating and modelling number entry can be successful, but does not apply to all aspects of number transcription.

2.5.4 Summary

The current models of transcription typing vary between simple and usable, to highly complex and accurate, yet less accessible. The models have recreated the Salthouse phenomena with varying success; the TYPIST model is unable to make predictions about errors and other key phenomena, but the QNM model has recreated all but two of the phenomena.

The design and reasoning behind these models has highlighted how important it is to know about how numbers are chunked. It appears that the chunking strategy will affect the way that input text is processed, and it can change the eye-hand span depending upon how much input information is available. The currently available model of number transcription does not account for this variation in chunking size, nor can it simulate a reading-transcription task. If a model of number transcription were to be produced, more needs to be known about number chunking strategies.

2.6 Current Text and Number Entry Research

The work explored in the previous sections has been largely theoretical. The research has generally aimed to understand the cognitive processes involved during a transcription task. In this section, the current research in the fields of text and number entry are explored to understand what aspects of the task are important in an applied domain.

2.6.1 Text Entry

The research into text entry in an applied context is diverse, as there are a wide-variety of different situations in which a user may need to enter text. Much of the research has focused on how text entry on mobile devices can be improved (Dunlop & Levine, 2012; Zhai & Kristensson, 2003; Zhai et al., 2009), but investigations can vary from understanding how text can be entered using gesture (Jones et al., 2010), to how it can be entered using minimal interaction, for people with limited speech and motor control (Baljko & Tam, 2006). This section does not aim to explore the entirety of the innovations within the text entry domain, as this is outside the scope of this thesis (recent review can be found in Dunlop & Masters (2009); Kristensson (2009)), but instead aims to understand a particular aspect of text entry methodology.

One of the key resources within the text entry domain is the MacKenzie and Soukoreff phrase set (MacKenzie & Soukoreff, 2003). This phrase set contains short alphabetic text phrases, with no punctuation and minimal upper case letters. The aim of this phrase set is to accurately simulated realistic typing behaviour, whilst still providing the researcher with control over the text that the participants are entering and allows accuracy to be easily assessed. This allows for direct comparisons between text entry interfaces.

The use of this phrase set in text entry evaluation is important. It has been seen previously in this chapter that the meaningfulness of text can affect how it is typed (Salthouse, 1986). The work of the MacKenzie and Soukoreff phrase set aims to ensure that participants during experiments are typing meaningful text that accurately reflects text that a typist might enter. The ability for this phrase set to create text entry evaluations that are valid and comparable means that it has been used by many researchers for multiple experiments.

Further work into the methodologies used when evaluating text entry tasks has suggested that this approach may not be externally valid when evaluating other text entry interfaces. For this reason, further phrase sets have been created that apply to more domain specific situations, such as when entering text on a mobile device (Vertanen & Kristensson, 2011), when testing a text entry interface with children (Kano et al., 2006), or when simulating SMS messages (How & Kan, 2005), for example. The range of available phrase sets shows how important it is when evaluating text entry interfaces to accurately simulate the text entry tasks that a user will be performing when using the interface.

Summary

This brief exploration of text entry research methodology has highlighted the importance of understanding realistic patterns of words when evaluating transcription tasks. Similar phrase sets are not available for number entry evaluation. This may be because the patterns in different number entry domains are not different enough to warrant distinction. But there is no evidence currently to show that this is the case.

2.6.2 Number Entry

The literature in the previous sections has focussed heavily on the theory behind number transcription. Much of the current number entry research, however, is more focussed on applied tasks such as number entry in the medical domain. In this section, examples of current number entry research are presented. The work is largely split into two different methodologies: simulations and lab experiments.

Briefly, a common number entry task in the medical domain is outlined. In modern hospitals, drugs are often administered to patients via a device called an infusion pump. This pump controls the levels that a drug enters a patient. The pump needs to be programmed so that it administers the correct amount of drug, over the correct amount of time. Errors during number entry in this process can result in patients receiving overdoses of the drug that they were prescribed, which can often result in patient harm or death. From this example, it can be seen that improving number entry and reducing error rates has the potential to reduce unnecessary patient harm.

Lab experiments

In order to design to prevent errors, it is important to first understand the errors that are possible. This work was conducted by Wiseman, Cairns, & Cox (2011). In this study, participants were asked to transcribe numbers under non-ideal circumstances, either having to store more than one number at a time in memory, or by being placed under time pressure in the form of alarms. Participants entered numbers on a touchscreen tablet. The study highlighted 21 possible different types of number entry error, more than had been reported previously (Thimbleby & Cairns, 2010; Oladimeji et al., 2011). What was not clear from this experiment however, was exactly what caused each of the error types. The paper explores the possible causes: mis-reading the number, mis-remembering the number, mis-typing the number; but it can do no more than hypothesise which failing resulted in which errors. Additionally, the errors collected in this work pertain solely to the 10-key number pad. Some of the errors collected, such as a finger slipping when typing a key, will not be applicable for other number entry interfaces, such as a dial interface. Finally, the ecological validity of this experiment is questionable. The aim was to inform number entry on medical devices, and yet the numbers used were randomly generated and the task artificial. This paper should be seen as an initial step towards understanding number entry error in the medical domain, but the results can be improved upon.

Other studies have tried to situate themselves more specifically in the medical domain. Oladimeji et al. (2011), for example, specifically compared two interfaces commonly found on infusion devices in the hospital. The aim was to understand how each interface facilitated the user's error checking behaviour. One interface, the 10-key number pad, was highly familiar to the users, whilst the other, the chevron interface, was not. The chevron interface controlled number entry by adding incremental amounts to the display. Pressing the "up" chevron would add a set amount to the number in the display, whilst pressing the "down" chevron would subtract that amount from the number. During the experiment, the researcher artificially inserted slip errors into the number being displayed. It was found that users were significantly more likely to notice these errors when using the chevron interface, compared with the 10-key interface. This result suggests that the chevron interface encourages more error checking behaviour. However, the applicability of these results to the real world raises questions. The experiment was conducted using a mouse to press simulated keys on a computer monitor. This may have artificially slowed down interaction, and altered the natural gaze of participants, compared to where they would ordinarily look when interacting with physical buttons, which participants could interact with through touch. The result also does not consider the amount of errors caused by the participants themselves. It may be that, although the chevron interface in this instance resulted in more errors being noticed, it actually resulted in more errors occurring. This difference is not explored.

A follow-up study by the same authors was conducted on a physical prototype of an infusion pump, with users entering numbers with physical buttons and dials (Oladimeji, Thimbleby, & Cox, 2013). This thorough review of number entry interfaces tested five different possible designs with a number of metrics, including speed of use, errors made, error corrected and severity of error. The results of participant testing highlight the complexity of choosing the “best” number entry interface. The number pad allowed for fast entry and low error frequency, but the errors that did occur were serious. The incremental interface such as the chevron once more focussed user attention to the number display, encouraging users to reflect on the number that they had entered and check for errors. This exploration of number entry interfaces offers the most in depth understanding of the trade-offs involved in number entry interface design. Although not providing a universal answer, it can help designers assess which interfaces are best suited to the tasks for which they will be used.

All of the studies presented here aim to inform the design of medical devices. The results of the studies highlight potential issues with medical device design and have increased our knowledge of number entry. It is clear, however, that more work can be done. These studies were conducted to varying degrees of ecological validity, often using numbers not representative of a medical device programming task. These studies also involved relatively small numbers of participants, performing a small number of tasks. One potential solution to this problem is to run computer simulations of number entry tasks.

Simulations

Simulations allow faster evaluation of designs, allowing researchers to test interfaces on a larger scale than is possible with participants in a lab setting. This approach has been used to explore potential number entry interface designs, particularly with the aim of preventing errors and reducing their impact.

Some of the first work in this area aimed to prevent error by blocking ambiguous number entry strings (Thimbleby & Cairns, 2010). Strings such as ‘1.2.3’ can be processed a number of different ways by devices, often leading to unexpected results for the operator. After designing an interface that would not allow ambiguous number strings to be entered, Thimbleby & Cairns simulated thousands of interactions with the machine by randomly inserting errors into the simulated keying process. A comparison of how many of these errors were blocked by the newly designed interface showed that “out by 10” errors (errors where the amount entered is ten times larger or smaller than intended) could be halved using error blocking.

This result is clearly beneficial in a medical domain, where preventing out by 10 errors could reduce the number of accidental overdoses programmed. However, it does rely upon simulated error and only takes into account key slip errors. It does not consider errors caused by mis-reading or mis-remembering (Wiseman et al., 2011). However, the use of simulation allows for designs to be tested quickly in large volumes.

Similar simulation analysis has been performed to analyse potential configurations for a 5-key infusion interface (Cauchi et al., 2012). This interface allows the user to enter numbers by moving a cursor on screen using “left/right” directional keys. The use of the “up” key then increments the digit that the cursor has currently selected, whilst the “down” key decrements it. The fifth key is the confirmation key. This interface can be configured in a number of different ways: pressing the up key when the cursor rests over the digit 9 can either cause the digit to change to 0 or can continue to display 9. Moving right at the right-most digit can cause the cursor to wrap-around to the left-most digit or remain where it is. Using a simulation approach, as in Thimbleby & Cairns (2010), thousands of simulated interactions with this device were run, with simulated key slips during typing. This analysis was able to suggest the configuration which reduced the amount of harm key slips could have.

Once more however, this method, although allowing thousands of interactions to be tested, did not model all errors, only keying slips. It is currently unknown how many errors are caused by key slips and how many are accounted for by mis-reading or mis-remembering. The method of simulation is strongest when used to highlight and explore potential design flaws and solutions. From here, it should be combined with lab-based experiments in order to understand how user variation affects results.

In both simulations, the numbers used were randomly generated. This is an issue with the analysis, as it may be that numbers in the medical domain are not accurately represented with random numbers. The particular interfaces tested in these simulations may be specifically designed to enter medical numbers rather than random numbers, meaning the predicted error reductions are over estimations.

2.6.3 Summary

The current work in the number entry domain has shown that there are many issues to be addressed when it comes to the design of number entry interfaces, particularly in the medical domain. Issues with the trade-off between speed and accuracy have been raised. Additionally, there are compromises to be made with regard to error rates and error severity.

The current work has so far aimed to both understand the causes and types of error (Wiseman et al., 2011; Oladimeji et al., 2011, 2013), and some insights have been gained in this area. The work has also attempted to make design decisions that can reduce the chances of error (Thimbleby & Cairns, 2010; Cauchi et al., 2012). So far, there appears to be no solution that successfully does this job. Much of the work in this area is lab-based and does not fully simulate the medical domain, often using randomised number lists, which may not be representative of the numbers used in hospitals. This means that potential domain-specific design solutions may be being overlooked. It is clear that there is more work to be done in the area of applied number entry research.

2.7 Motivation for future work

From the literature review, it has been shown that we read numbers and words using different processes in the brain, suggesting that directly applying results from the alphabetic text entry domain to number entry should not be done without justification. It has also been shown that familiarity of a number affects how it is read, in much the same way that words and non-words are read differently. This result may have implications for numeric transcription and suggests that the use of random numbers in number entry research up to this point may not be ecologically valid. However, it is not clear currently what a "familiar" number might mean.

Previous text transcription research provides a solid methodology for investigating possible differences in transcription typing behaviour for words and numbers. The Salthouse metrics provide benchmarks for comparisons between text types. These metrics can also be used to inform the design of models of number transcription. The current work in the number transcription area does not apply to reading-transcription tasks and does not consider the effect of familiarity.

Research in the text entry domain relies upon an understanding of the variations between text entered in different contexts. There are multiple phrase sets available depending on which domain the text evaluation will apply. Similar phrase sets are not available for the number entry domain, and it is not clear currently if similar context-dependent variations will occur.

The current work within the number entry domain does not rely upon any theoretical knowledge of number entry. This means that random numbers are often used and effects of familiarity are not considered. Additionally, the research often does not aim to replicate numbers used in applied domains. This means that the generalisability of results is questionable and that the design space is not being fully explored.

This research leads us to the following questions:

1. Can knowledge of alphabetic text transcription apply directly to number transcription or do the two tasks differ?
2. Does the familiarity of a number affect how it is stored in memory and, therefore, typed?
3. Is it possible to define a "familiar" number? What numbers do we consider to be familiar?
4. What numbers are used in hospital settings? Is it possible that medical workers use familiar numbers during number transcription?
5. How can number entry research be improved by understanding more about the numbers used in a particular domain?

With this established, we can begin to understand more about what matters when transcribing numbers and how we can use this information to design better interfaces and tasks.

Chapter 3

What Is A Familiar Number?

3.1 Introduction

Throughout the literature review, it was made clear that during transcription tasks, words and non-words are treated differently. Non-words are typed slower than words, they cannot be memorised as effectively, and they do not require the typist to look ahead as far as when transcribing words (Salthouse, 1986). Firstly this provides insight into the cognitive processes involved in transcription typing of words and non-words (John, 1996). Words are processed using larger chunks due to their stronger representation in memory. Secondly, this finding has provided clear guidelines for text entry design and evaluation: text entry interfaces should be tested with real words, because non-words are not an ecologically valid substitution.

Knowing how important this factor of familiarity is within the alphabetic text domain, it is also important to understand whether there are similar effects within the number entry domain. Studies with aphasic patients have shown similarities between words and numbers. Much as some aphasic patients lose the ability to read non-familiar words, whilst being able to read words, some lose the ability to read non-familiar numbers, yet retain the ability to read familiar numbers (Cohen et al., 1994).

One aim of the work in this thesis is to address the question of whether all numbers can and should be treated equally when considered in a number transcription task. In order to test this theory, it will be necessary to have a corpus of both familiar and non-familiar numbers. There are particular issues with these resources however, as a dictionary of numbers does not exist, making it difficult to objectively decide which numbers constitute familiar and which are non-familiar. The aim of this chapter therefore, is to produce two sets of numbers that can be deemed familiar and non-familiar. These sets will then be used in experiments to test the effect of familiarity of a number on typing metrics.

An additional aim for this chapter is to produce a set of “causes of familiarity”, which might be used to generate familiar number sets in the future. Previous definitions of familiar numbers are sparse and do not provide the detail needed to create a list of familiar numbers. The causes suggested in this chapter might provide a guide for familiar number generation in future experiments.

3.2 Defining familiar numbers

In order to create a set of familiar numbers, it is important to have a solid definition of the term. Familiar numbers are those that, when viewed by the reader, will not be construed as being ‘random’. That is, they are numbers that evoke some meaning for the reader and have some representation in memory.

Previous work has been conducted into the effects of “meaningful” numbers compared to ordinary numbers when read out by a patient with aphasia (Cohen et al., 1994). The researchers highlight the issues with determining these types of numbers; they are likely to be highly personal. However, the study still uses a set of meaningful and non-meaningful numbers. Meaningful numbers are described as follows: “highly frequent numerals such as famous dates or brands of cars”. However, no more detail than that is given. In the quotes that are highlighted in the paper, it is clear that the meaningful numbers they have chosen are highly specific to the location of the study as the numbers revolve around key dates in France’s history, and popular French brands.

Anderson & Schooler (1991) argue that our memory is shaped by the environment; that frequently encountered items have stronger representations in memory and can thus be accessed more easily. The strength in memory represents how quickly that particular item can be accessed. This stronger representation in memory is another way of defining familiarity: the more frequently we are exposed to a number, the stronger its representation in memory, and the more familiar it becomes.

It is clear then that a new set of numbers will need to be defined for forthcoming experiments, as previous experiments using meaningful numbers have not provided a list of resources. Using Anderson & Schooler’s theory, a list of possible causes of familiarity are explored.

Personal Importance

One way to define familiar numbers is those which are important to the reader in some way, that is, numbers which may represent significant aspects of their life; for instance, their birth date or year, or perhaps digits in theirs or a loved one's phone number. It is possible that these types of personally familiar numbers could be further classified into those regularly typed and those not typed.

Initially, it may appear that for there to be an effect in the way they are typed, that these numbers be regularly entered into devices, and that the typist should have plenty of experience in typing those digits in that order. However, typing is not simply the muscle memory involved in pressing keys, it also involves the cognitive stage of reading and memorising the input text meaning that experience with typing the number is not important. Personally familiar numbers will therefore include both numbers typed regularly (personal identification numbers for instance) and those that are typed less frequently (for example the birthday of a loved one).

Cultural Significance

In addition to the numbers that are personally familiar to us, there is a large set of numbers that are more globally familiar to a larger group of people. For instance, in the following list of four digit numbers, it is likely that one in particular will stand out to the reader if they are a European citizen who has been taught history in a school in the latter half of the 20th Century:

5234, 4478, 1945, 5111

Of the four numbers, one represents a year in recent memory. Furthermore, that year holds extra significance as it is the year that World War II ended. This therefore could constitute a familiar number. In terms of Dehaene's model of number reading (Cohen et al., 1994), this type of number is associated with an encyclopaedic memory and reading it will cause a look-up to occur, which is different to the other numbers in the list, which will not only have no information associated with them, but may not even be stored in memory at all if the number has never, or rarely been encountered.

This raises the question of whether all years in recent memory would have the same effect. Presenting that list again with another year inserted, for example 2004, would likely cause the reader to list 2004 as a "familiar" number as it is one they hold in memory. Some readers may class this as familiar due to some personal significance (perhaps the birth year of a child) or due to some cultural meaning (the year their football team won a particularly important trophy for instance), however, many others may have no such associations. Are all four digit numbers which could plausibly be considered a year familiar? How far into the future can this definition be stretched? Readers in the year 2013 may still view 2015 as a year, but is 2115 still as familiar? At what point do these four digit numbers begin to look random? It is not a trivial task for the experimenter to create a list of familiar numbers by just taking years as it is not clear how readers will view the years.

Culturally significant numbers do not necessarily have to entail an historical event associated with them. There are other numbers that are important in our everyday lives that are also familiar to many. For instance, the emergency service numbers in the UK and US are likely to be familiar: 999 and 911. Numbers from popular culture may be familiar to those who are aware of the references (for instance, 90210 from the show 90210 or 1337 from popular internet culture). Defining these familiar numbers poses a problem, as not every reader has the same life experience and thus not every reader will have the same cultural points of reference. A deep knowledge of a particular academic field may also affect the numbers some people find familiar. For instance, being shown 31415 may mean nothing to some people but to some people, who have studied mathematics, they may instantly see the first few digits of pi and therefore find the number familiar.

Patterns

Some numbers may appear familiar simply due to a pattern they contain. For instance, some readers may consider the numbers 1234 or 9876 to be familiar, despite them having no significant meaning. Equally numbers that have a visual pattern to them, for instance 4664 may also be considered familiar simply because of the ease by which they can be memorised. They may also be familiar due to the fact that they couldn't be classed as random and so default to familiar.

3.3 Study 1: Gathering a set of familiar numbers

It is clear from these descriptions of possible familiar numbers that it is not possible for one person to collate a set of familiar numbers. Personal experiences, the culture a person is brought up in, and specific knowledge of a subject area will all likely have an influence on the numbers a person considers to be familiar. Additionally, there may be other possible causes of familiar numbers that the experimenter and researchers up until this point have not considered. For these reasons, a set of familiar numbers will be created by collecting the opinions of many people, not just the single opinion of the experimenter.

Here a study is presented that aims to elicit a list of numbers considered familiar by a large group of people, with the intention of creating a more universal set of numbers than any one person could create. By asking the opinion of many people, small variances in familiarity should be ignored and only those numbers with high familiarity ratings for many people will be used, making it more likely that participants in future experiments will also consider the numbers to be familiar.

Although this study is initially aimed at generating a set of familiar numbers, additional metrics may be taken throughout the study that can provide more evidence for the thesis argument that familiar and non-familiar numbers are processed in different ways, much as words and non-words are.

The structure of this task is similar to a lexical decision task: a task where participants are asked to make decisions about the word(s) they are seeing and their lexical status. For example, they may be asked whether the text they are seeing is a word or a non-word. In studies, it has been shown that participants provide the 'word' response significantly more quickly than a 'non-word' response (Meyer & Schvaneveldt, 1971). The causes of this difference are believed to be due to the different ways that words can be read. An analysis of these different models can be found in Chapter 2. Briefly, non-words may be more difficult to recall as they have not been read before and thus have very low activation levels in memory, making their representation in memory very weak. A dictionary model would predict that non-words would take longer to classify because firstly every known word that is similar to the non-word has to be exhaustively checked before a no-result can be returned. Both approaches predict a slower response time for non-words, and may potentially do so for non-familiar numbers additionally.

In a similar fashion to previous lexical decision tasks with words, the response times will be collected during this number decision study. The hypothesis is that familiar number decisions will be made faster than random number decisions. As with words, it is hypothesised that this difference is due to the different pathways used when reading familiar numbers that are stored in memory.

3.3.1 Method

Participants

The study took place online. Participants were recruited from an opportunity sample of responses to online adverts placed in a number of locations, including on email mailing lists, twitter and other social media. A participant is defined as someone who gave more than one response during the study.

In total, 100 participants took part in the study with a mean age of 34.96 (*SD*=8.61). The country of residences reported can be seen in Table 3.1.

Country of Residence	**Number of Participants**
United Kingdom	94
United States of America	2
Ireland	1
Canada	1
China	1
Malaysia	1

Table 3.1: Countries of residence for participants in the Familiar Number study

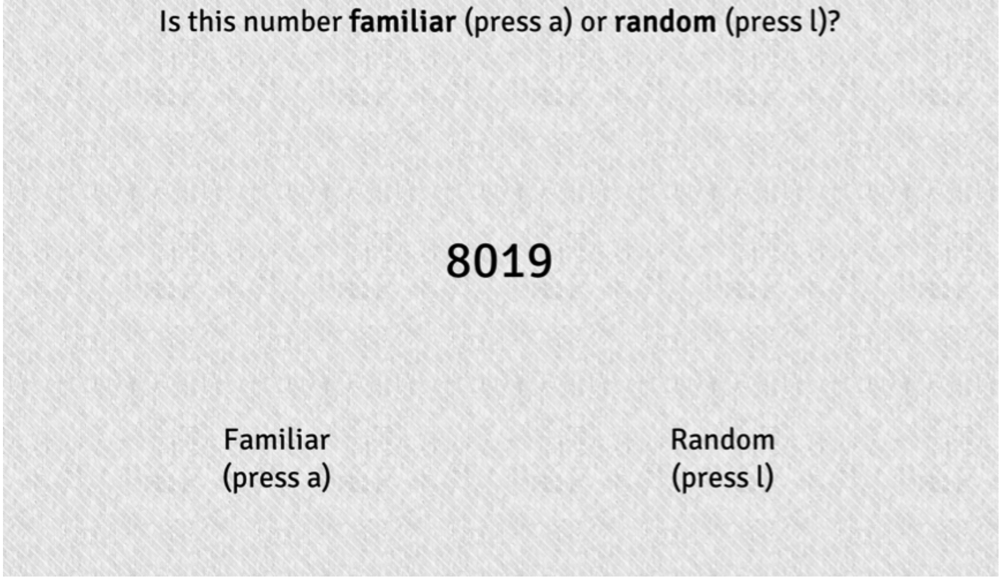
Materials

The numbers presented in this study were initially taken from results of a pilot study. A set of possibly familiar numbers was generated by the experimenter and a randomised list of numbers produced. These were given to pilot participants to rate as random or familiar. Pilot participants were also asked to generate their own familiar numbers.

The aim was not to present a set of familiar and a set of random numbers to the eventual participants in this study, but rather to present a wide range of numbers that could possibly be considered familiar or random. Participants in this study would not be tested on whether they agreed with this initial seeding of numbers. Any assessments of the numbers' familiarity would be based entirely on ratings providing during the study.

The initial pool was comprised of 97 potential familiar numbers and 666 potential random numbers, which were randomly generated and spanned multiple different magnitudes.

From this pool of initial numbers, 50 at a time were chosen to present to the participants. The 50 numbers used were randomly selected each time a new participant loaded the page,


Is this number **familiar** (press a) or **random** (press l)?

8019

Familiar
(press a)

Random
(press l)

Figure 3.1: Screenshot of the Familiar Numbers study, showing one number to be classified.

these were taken from the potential familiar, potential random and user generated pools. These were displayed one at a time on a web page. The numbers were displayed in black 30pt Signika Negative font.

On the page, instructions explained that, if the number was "familiar" the 'a' key should be pressed and if the number was "random" the letter 'l' key should be pressed. These keys were chosen rather that the 'r' and 'f' keys as they were physically separate on the keyboard and mapped to the position of the words "familiar" and "random" on screen. Providing a response to the number caused the next number to appear. A screen shot of the study screen can be seen in Figure 3.1.

The decision was made to use the term "random" in this study rather than "non-familiar", as it was thought to be more easily understood by participants.

In order to explore more numbers than those generated by initial pilot studies, the online study also collected suggested numbers from participants at the end of the study. Once 50 numbers had been rated, a page prompted the user to enter in any numbers that they deemed "familiar". Participants were not given any instructions on the length of number they should suggest. These were then processed to ensure they contained only numbers and were fed into the pool of numbers for future participants to rate.

Procedure

The study was advertised widely on social media and email lists. Participants were made aware that there would be no reimbursement for taking part. The advert provided a link to the study website.

When arriving on the site, the study was introduced to the participant by explaining the purpose of the study (to understand which numbers they considered to be familiar/non-familiar), the participant was also informed that they could leave the study at any point. Before continuing with the study, the participant was asked to provide two pieces of information: their year of birth and country of residence (it should be noted that the default country of residence in the drop down menu was set to United Kingdom). The study then began and the participant rated 50 numbers individually, using the ratings of "familiar" or "random". After the study was completed, participants were given the opportunity to add any numbers to the pool that they considered to be familiar.

3.3.2 Results

A complete study is composed of 50 responses from the participant. A total of 88 of the 100 participants (88%) completed all 50 responses for the trial. On average, participants completed 47.41 responses (*SD*=8.62).

The final pool of numbers in the study, including both those used at the start of the study, and those suggested by participants, was 995 numbers. Including duplicated numbers, participants of the study suggests 571 numbers. The number set ranged from 0 to 1818118181. The median number was 3878.

Familiarity of Numbers

Firstly, the familiarity rating of the numbers will be considered. The results of this study produced a list of numbers, showing the number of times each was rated as familiar, and the number of times each was rated as random. The aim of this study was to produce a set of numbers that were familiar and a set that were unfamiliar.

It is important to consider how many times each number was rated. It would not be appropriate to consider a number as familiar if only one participant had deemed it so; the aim of the study was to avoid subjective biases.

The mean number of times each number was rated was 4.76 times (SD = 7.52), the median number of ratings was 2. The amount of times each number was rated varied between 1 and 62. This variation was due to the random selection process that chose the numbers each participant would see. A threshold of 4 or more ratings was decided upon, as this was near to the mean number of ratings for all numbers.

Next, each number was assigned a familiarity ratio. This ratio represented how often a number was marked as familiar compared to how often the number was marked as random. This was calculated using the following equation:

Familiarity ratio = Number of familiar ratings / Total number of ratings

A familiarity rating of 1.0 therefore represented a number that was rated as familiar by every participant who saw it. A familiarity rating of 0.5 meant a number was rated as familiar half of the times it was seen and random for the other half.

The mean familiarity ratio was 0.2 (SD = 0.32), though the median was 0.0. The ratio ranged from 0 to 1.0.

The thresholds for familiarity and randomness were set at above 0.75 for familiarity and below 0.25 for randomness. A cutoff was required, as the familiarity ratio is a continuous construct in this study, rather than a binary decision. This threshold was chosen as it is close to the value of the interquartile range of 0.833. Numbers above this are deemed to be different enough from the rest of the data as to be noticeably familiar.

Using this sorting algorithm, 73 numbers could be considered familiar and 143 as random or non-familiar. The familiar numbers ranged from 0 to 118118, with a median of 1877. The familiar number magnitudes can be seen in Figure 3.2. A full list can be seen in Appendix A.

Timing

Here the response times for rating are analysed. The aim of this analysis is to see if, as with words, there are differences between the time taken to rate a number as familiar compared to the time take to rate it as random. This analysis is performed on all participants' responses.

The mean response time for all numbers was 1724 ms (SD = 1920). The mean response time for a participant choosing the familiar option was quicker ($M = 1601$, $SD = 1668$) than for choosing the random response (M = 1829, = 2107). A paired t-test performed on this difference found that familiarity significantly effects response time $t(92) = 3.16$, p = .002.

3.3.3 Discussion

Familiar Numbers

The study was successful in producing a set of familiar, and equally importantly, non-familiar numbers. Participants were capable of making decisions about a number's familiarity. Ad-

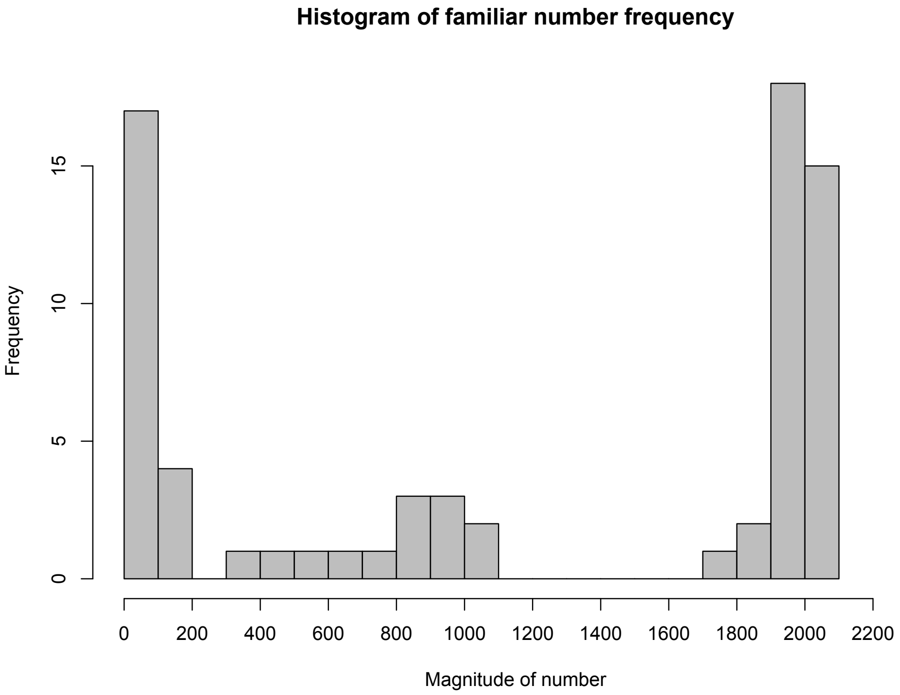


Figure 3.2: Histogram showing the range of numbers rated as familiar in this study. Excluded from the histogram are the numbers 10000, 12345, and 118118.

ditionally, participants were able to produce their own familiar numbers in addition to those they had seen throughout the study.

However, what is not clear from this portion of the study are the causes for familiarity. It is possible to see from Figure 3.2, that many numbers had a magnitude of between 1900 and 2100, which could support the hypothesis that number years would be considered familiar. Although this analysis again relies upon experimenter bias and is highly subjective.

Response Time

The response times gathered during this study support the hypothesis that familiar numbers will be rated more quickly than non-familiar numbers. This in turn provides evidence for the hypothesis that familiar and non-familiar numbers are processed using different cognitive pathways and are represented differently in memory.

Previous lexical decision research performed on words and non-words may be able to explain this difference in response time. There are a number of models that can explain different response times for decision tasks (see Ratcliff et al. (2004) for a review). The quicker response time for familiar numbers could mean that participants performed something like a “dictionary look-up” when reading the numbers. This theory suggests that when determining whether a familiar number is being read, all known numbers will be checked. If the number is present then a ‘familiar’ response can be given. It is only after exhausting the entire list, and not finding a corresponding entry, that a ‘random’ response can be given. This search technique would mean more familiar targets are labelled quicker than less familiar targets. This theory is based upon the Serial Search Model approach (see Rubenstein et al. (1971)).

Alternatively, using an activation level approach, this difference in response time could be accounted for by differing activation level thresholds. Various models of lexical decision tasks suggest that once a familiar target is read, it increases activation of similar targets, once this activation reaches a threshold, a ‘familiar’ decision can be made. After a cut-off time period has passed, and the threshold has not been reached, then a ‘non-familiar’ decision is given (Paap, Newsome, McDonald, & Schvaneveldt, 1982).

Without further testing, it is unclear which of these theories account for the findings of this study. It is likely that in order to fully contribute to this debate, more experimentation needs to be performed in order to provide more results for testing with a theoretical model. It is unclear currently what the error rate is for this numerical choice task. This in itself is a difficult problem, as the concept of an ‘error’ when classifying numbers as familiar and non-familiar is a difficult one: familiarity is something highly personal and subjective.

Regardless of this, the results from this study suggest that a model of number reading must account for the significant difference in response time when deciding if a number is familiar or non-familiar.

3.3.4 Limitations

One limitation of this study is the lack of diversity in the respondents. That nearly all participants were from the United Kingdom suggests that this familiar number set may contain numbers of cultural significance to that particular population. For instance, the number 999 was deemed familiar which is likely to be due to its key status as the emergency services number. It is possible that participants from other countries may not share the opinion that this is a familiar number.

It may however, be possible to consider this lack of diversity as a benefit. It ensured that any culturally significant numbers were indeed rated as such. A participant population consisting of multiple different nationalities may have not allowed culturally significant numbers to arise due to conflicting opinions of the participants.

3.3.5 Conclusion

The study was successful in its aim to generate corpuses of familiar and non-familiar numbers for use in future typing experiments. Additionally, the study provided some evidence for differences in reading between familiar and non-familiar numbers through analysis of response times.

However, it is clear that more research is required in order to determine the causes of familiarity. Suggestions have been made by researchers previously (Cohen et al., 1994) however, these have neither been exhaustively explored, nor externally validated. The next study will therefore aim to explore further the possible causes of familiarity using the numbers generated during this study.

3.4 Study 2: What makes a number familiar?

In previous literature that has required the use of familiar or meaningful numbers, the researchers have relied upon their own assumptions to generate this data (Cohen et al., 1994). As discussed in the introduction to this chapter, it is likely that familiarity of number may be a highly subjective thing, meaning self-generated "familiar" numbers may not be generalisable to others. In addition to this problem, the previous research has provided no information about the set of numbers that were used and assumed familiar. This means that future research could not benefit from the work.

In Study 1, a set of familiar and non-familiar numbers were generated which could be used in future studies where familiarity is important. However, these numbers are likely to contain familiar numbers that are specific to a specific audience and may therefore, not be entirely generalisable to different populations. A similar issue arises in previous work into familiar numbers: Cohen et al. tested a French aphasic patient who deemed the number '504' familiar, as it is a popular model of car in France. It is unlikely that this number would be as familiar to a British resident.

There is currently no method of producing a set of numbers that are familiar for a particular population. The analysis in this section aims to firstly understand the reasons behind the familiar numbers generated in the previous study. This is done by means of a card-sorting exercise with participants. Secondly, with this knowledge, a set of heuristics will be presented that can aid in the generation of familiar number sets for future studies that require them.

3.4.1 Method

Participants

Nine people completed a card-sorting task. The participants for this study were taken from an opportunity sample of people available to the experimenter. The participants were a combination of people who had taken part in the first part of the study and people who had not seen the numbers previously.

Design

This study used the set of familiar numbers generated in the previous study, printed on cards. The cards were given to each participant in a shuffled order, which was not controlled by the experimenter, so as to avoid bias affecting the presentation of the cards. The variables recorded from this study included the names of groups of categorised cards along with each participant's reasoning.

Materials

The 73 familiar numbers collected from the previous study were printed onto individual pieces of paper around 60mm x 60mm in size. Each number used the same font (Calibri) and font size (36pt).

A large, blank table surface was used in each card sort to allow the participants to fully spread out the cards they were presented with. This allowed participants to create clearly separated groups.

Procedure

In the first iteration of this study, participants were told they were going to be given a set of numbers that they would have to group. The participants were then presented with all 73 numbers cards and asked to create groups that were meaningful to them. This task appeared overwhelming to the participants, who struggled to organise the numbers presented to them.

To rectify this and help participants create possible groupings, an additional first step was added to the card-sorting task. In the first step, participants were presented with three cards at a time (a randomly created set to avoid experimenter bias) and asked to create a group that two of the cards belonged to that the other did not; the participant was asked to say aloud their reasoning for that decision. This group was then left on the table, whilst the third card was placed in a discard pile. All the cards were presented in this manner (save for one card, as 73 does not neatly divide into 3). During this group forming stage, participants were allowed to use the same group multiple times and were encouraged to be as creative, or simplistic as they liked.

Once this step had been completed for all cards, the participants were asked to tell the experimenter what the rationale behind each group was. This step helped remind the participants of the groups they had created and also aided the identification of weaker groups. The participants were then allowed to manipulate the groups they had created: they could break them up, join them together, or form new ones as they saw fit. Participants were told they could create groups that were 1 card size or larger. They were then given the discarded cards to add to the groups.

A key aspect of this method was that participants were able to have a "junk" pile of cards if they needed it. This meant that participants were not forced into creating artificial groups of cards, or adding cards that did not suit a particular group. This avoided over-fitting of numbers to categories.

3.4.2 Results

The participants produced between 6 and 14 categories each, with the mean number of categories being 10.22 (*SD*=2.68). Three of the participants made use of the "Junk" category. In total, participants produced 62 different categories.

As participants were able to name their categories with any label they chose, many of the differently named groups classified the same aspect of the numbers within. Further to this, it was clear that the groups had similarities with regard to rules they had used for inclusion in the group. An analysis was performed to group the categories with similar meaning. Table 3.2 shows these groupings and the number of times that grouping was applied to a number during the study. These groupings exclude the "Junk" categories, which accounted for 2.44% of all ratings. Here, the rules for categorisation are elaborated:

Culturally significant These groupings allude to the number's importance within a particular society. These might relate to the history of the society or key numbers relating to tradition (for instance important ages such as 16, 18 and 21).

Pattern These numbers are those that adhere to a particular pattern and are not necessarily important for what they stand for. It might be possible to replace the digits with other characters, for instance and still see the pattern.

Format These numbers are familiar because they adhere to a particular format for a number, for instance, any number that is 4 digits long and between 1900 and 2000 can constitute a "20th Century" number.

Mathematical Categories which were mathematical looked at the mathematical properties of a number.

Personal Personal numbers had highly subjective groupings, groupings that were only known to the participants creating the set at the time. The rules for inclusion in these groups could not be articulated easily.

3.4.3 Discussion

The groupings generated from the card-sorting technique were largely aligned with those hypothesised at the beginning of this chapter. Participants found that the familiar numbers could easily represent years. For example, one participant went so far as to group the numbers entirely according to which century they came from, including categories such as “Bronze Age” and “Dark Age”. However, others were able to see the potential of numbers to represent years without applying this metric to every number they saw.

Cultural significance and pattern of the number were also important to participants when grouping the numbers they saw. These again were hypothesised to be key causes of familiarity.

One group that was not initially considered was the mathematical properties of the number. Upon reflection, it seems clear that this would be important to participants as numbers, ultimately, are a mathematical construct. In some instances, some participants looked only for mathematical meaning.

The necessity of a “junk” pile showed that, despite the fact that all the numbers presented in this study had been rated as familiar in the previous study, familiarity is still a personal and subjective measure. Whereas many people may see a number as significant and familiar in some way, others may not see this.

The results of this card-sorting study support the reasons for familiarity suggested by Cohen et al. (1994) who suggested familiar numbers might include “famous historical dates [...] French brands of car [...] familiar zip codes”. However, the work conducted here extends these categories and generalises them.

3.4.4 Limitations

This study did not take into account the familiarity ratio for each of the numbers presented to the participants. A future study might investigate a correlation between the confidence in the familiarity ratio and the ease of a participant assigning that number to a particular group. Much as familiar numbers were labelled more quickly than non-familiar numbers in the previous study, it may be that more familiar numbers are also grouped more quickly.

A future study might also test the strength of the participants’ groups by asking participants to generate new numbers that fit within each category created. This would have ensured that some of the more specific and ‘difficult to use’ groupings might have been rethought by the participants. For instance, the participant who suggested “Useful everyday” might have investigated that category further to understand what about the numbers were “useful”.

3.4.5 Conclusion

The rules generated by this study suggest that if a set of familiar numbers is needed for an experiment, and it is not possible to generate these from a large participant pool, then a researcher might like to consider seeding the numbers from the following categories: *Culturally Significant, Mathematical Importance, Patterns,* or *Adherence to a particular format.* The "Personal Significance" category is not included in this set as this is subjective to each participant and thus unusable when generating a general set of familiar numbers.

3.5 Summary

This chapter has succeeded with the primary aim of generating a corpus of familiar and non-familiar numbers. These will be used in the following chapter. This contribution extends beyond just this thesis, as the corpuses may be used in other future studies where familiarity of a number is important.

Additionally, the data taken from response times in Study 1 have provided evidence that reading numbers is affected by familiarity and representation in memory. Links have been made to existing models of lexical decision models taken from the alphabetic text reading domain. These links show that numbers, like words, should not all be considered equal and that there are key differences in cognitive processing that need to be understood further and tested. This finding provides more support for the experiments that follow in this thesis.

Finally, the causes of familiarity of a number have been investigated and initial work has been done to create heuristics that can be used in the future if familiar numbers are required in a different context, for a different population.

Categorisation	Category	Example
Culturally Significant	Phone numbers	*118118*
114 (17.35%)	Ages	*21*
	Semantically meaningful	*1945*
	Recognisable	*101*
	Pop culture	*1969*
	Nerdy stuff	*404*
	Meaningful numbers	*42*
	Information	*999*
	Important	*365*
	Familiar but don't know why	*1066*
	Cultural	*666*
	Country code	*44*
	Common amount	*64*
	Area code	*808*
	Historical years	*1918*
	Important years	*2000*
Pattern	Repeated	*888*
63 (9.59%)	Begins with 10	*1024*
	Three digit	*200*
	Single digit	*3*
	Symmetrical	*404*
	Repeated pattern	*2020*
	Palindromic	*1991*
	One numeral	*555*
	Oh pattern	*101*
	Numbers when upside down	*1066*
	Number twice	*118118*
	Binary	*1000*
Format	Years	*2014*
336 (51.14%)	Time	*24*
	Twenty-first Century	*2000*
	Twentieth Century	*1992*
	Nineteenth Century	*1850*
	Years lived through	*2007*
	Victorian	*1877*
	Recent past	*2006*
	Other years	*1962*
	Future	*10000*
	Dark Ages	*999*
	Bronze Age	*128*
	Consecutive years	*2011*
	Boundary case	*404*
Mathematical	Sequential	*123*
110 (16.74%)	Small numbers	*7*
	Prime	*5*
	Multiple of 10	*900*
	Mathematically useful	*100*
	Ends in zero	*2020*
	Useful everyday	*50*
	Twenties	*25*
	Odd	*21*
	Nine related	*999*
	Hundreds	*200*
	Even	*24*
	Divisible by hundred	*800*
	Divisible by five	*25*
	Divisible by four	*16*
	2ˆn	*64*
Personal Importance	Pleasing numbers	*123*
18 (2.74%)	Personal years	*1996*
	Nice numbers	*24*

Table 3.2: Table showing the categories generated during the card-sorting activity and their subsequent grouping.

Chapter 4

Applying Transcription Metrics to Numerical Typing

4.1 Introduction

In Chapter 2, the different cognitive processes involved in reading numbers were highlighted. Studies with aphasic patients showed that it was possible for patients that have suffered brain trauma to read numbers that are familiar to them, whilst being unable to read all other numbers (Cohen et al., 1994). The exploration of different models of number reading shows that the semantic meaning of a number can cause it to be read using different pathways.

Reading is a key component in the transcription typing process, which means that these differences in memory representations may well have an effect upon how familiar and non-familiar numbers are transcribed. Research has shown that this is indeed the case when transcribing words; meaningless material comprised of non-words is transcribed slower. A typist copying meaningless words will not look as far ahead in the copy text as when transcribing meaningful words (Salthouse, 1986).

The different semantic processing routes are well-known for words and non-words and many studies have shown that these differences can be highlighted by measuring various transcription typing metrics (see Salthouse (1986) for a review). In this chapter, a similar approach will be used to investigate typing metrics for numbers. The results from these studies will investigate the different cognitive processes used for typing familiar numbers compared to non-familiar numbers.

In order to investigate this possibility, key typing metrics used in alphabetic typing studies will be replicated using the sets of numbers collected in Chapter 3. This chapter will present the results of four different typing experiments aimed at eliciting four different typing metrics: the Replacement span, the Eye-Hand span, the Copy span and Interkey Interval timing. The experiments will test not only familiar and non-familiar numbers, but words and non-words in order to establish if any differences seen when typing alphabetic text are replicated when transcribing numerical text.

It is not expected that the typing metrics should be the same for familiar numbers, as they are for familiar words, but rather the relationship between familiar and non-familiar numbers may be similar to that between words and non-words. An additional contribution of this chapter will be to understand how numerical typing as a whole compares to alphabetic transcription typing, something which until now has not been explicitly investigated.

The key hypothesis throughout this chapter is that the typing metrics tested will highlight a difference between the familiar and non-familiar numbers, much as they will for words and non-words. This will provide evidence that the two types of number are indeed processed differently when transcribed, much in the same way as words. Such a finding will endorse the previous research involving aphasic patients, but will do so in a healthy population. This result will provide further support representation in memory affects typing behaviour.

4.2 Typing metrics

The experiments presented in this chapter will recreate some of the typing phenomena presented by Salthouse in the 1986 paper "Perceptual, Cognitive, and Motoric Aspects of Transcription Typing" with an aim to understand more about how the familiarity of a number affects how it is processed when it is transcribed. This paper has been discussed extensively in the Literature Review chapter of this thesis. Here, the specific hypotheses of how these phenomena may differ according to numerical familiarity are formalised.

In total, 29 typing phenomena are highlighted by Salthouse. These phenomena range from understanding speed and types of errors, to the awareness of the text being copied. Not all are appropriate to investigate if the aim is to understand the cognitive differences between two types of target text. Before introducing the phenomena that will be used in this study, firstly the reasons behind discounting certain other phenomena will be discussed briefly.

Many of the phenomena are related to the motoric aspects of typing. For instance, it is known that the first letter after a space is typed more slowly than the rest of the letters in the word (Phenomenon 10). For many of these motoric phenomena, there is no reason to suspect that they would be any difference for familiar and non-familiar numbers, as there is no reported difference for words and non-words. In these instances, recreating these experiments would not provide any information about the cognitive processes occurring during numeric transcription.

Other phenomena highlight the effect of typist skill on the transcription process. Although this would be interesting to investigate (for instance, do medical workers who have dealt with the same numbers for many years use different cognitive processes when compared to new medical workers?), this work will not look at these phenomena, as skill-related metrics involve a more advanced hypothesis which should be tested after having established the basic numerical typing phenomena.

The remaining set of phenomena relate to the causes of error whilst typing. Again, this is an interesting avenue of research, however, some of it is less relevant to numerical typing (for instance many transposition errors happen when both hands are typing adjacent letters; in numerical typing, often only one hand is used (Cairns et al., 2014; Lin & Wu, 2011)). Investigating errors is another aspect of numerical typing that would have beneficial results for number entry research (see Thimbleby & Cairns (2010); Wiseman et al. (2011)), but again it requires more base-level knowledge about the cognitive processes involved in numerical typing before error information can be useful. This means at this point, the error phenomena are not relevant to the current thesis.

The phenomena that this thesis concentrates on are the transcription spans and typing speeds. These spans provide the most information about the motoric, cognitive and perceptual processes occurring during typing, and can therefore be used as a means of investigating the different processes occurring during the transcription of familiar and non-familiar numbers.

Chunking is thought to be a key factor in the differences in typing metrics when copying familiar and non-familiar text. Researchers aiming to recreate the typing phenomenon with cognitive models have speculated that the differences in the metrics may be due to the breaking up of incoming text into meaningful input units, or chunks (John, 1996; Wu & Liu, 2008).

A brief summary of the TYPIST model will now be presented, as using assumptions from this model can help to make predictions about how numbers might be typed during a transcription task. The TYPIST model considers the task of transcription typing in terms of the Model Human Processor (MHP)(Card et al., 1983). The MHP consists of three processors: the Perceptual, Motor and Cognitive processors. The task of transcription typing involves moving information from the perceptual processor, through the cognitive processor, resulting in movements controlled by the motor processor. The way this information moves through these processors is by means of chunking.

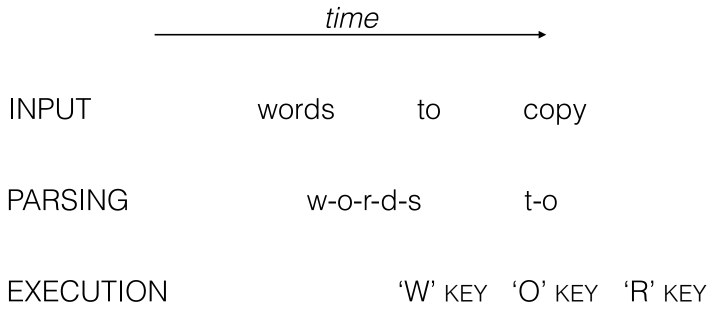


Figure 4.1: Diagram showing the parallel processes involved in transcription typing. The typist continues to read the input text whilst still typing the first word. A serial process would only allow the typist to read and type a single letter at a time.

The size of a chunk may vary depending on the task and the user. Chunks are often related strongly to the meaning of the text, for instance in transcription typing, a word is often considered a chunk. A chunk that a typist has seen before, and has a previous represention in long term memory (LTM), will be more efficiently memorised when read. This will affect transcription, as the process of typing involves the perceptual processor taking in a chunk, the cognitive processor looking up the chunk in long term memory (LTM) and then passing the individual letters of that chunk on to the motor processor. A chunk that is not stored in memory will need to be memorised at that point.

To operate at full capacity, the motor processor requires a series of motor actions to be stored in the form of letters from the cognitive processor. For the cognitive processor to provide instructions to the motor processor, it requires chunks from the perceptual processor. Although the processors are separated in this way, and each relies on the processor before it, they are still able to work in parallel, that is, a typist can strike the keys for the necessary letters of a word whilst still reading ahead to see what the next words will be (see Figure 4.1). This process is replicated in the TYPIST model.

The chunking strategy changes when the input text becomes less meaningful. If non-words are being copied, the words are no longer recognised as full chunks. If there is no long term memory store for the word then it is broken down into either pronounceable chunks (if the words are orthographically legal words) or, if the words are truly random strings, then the chunks simply become characters. Smaller chunks mean that due to the limited number of chunks that it is possible to hold in working memory (WM), the TYPIST model now processes a smaller amount of information.

The TYPIST model is able to accurately simulate human transcription typing based upon these assumptions. Here, the details of each of the spans that will be studied in this chapter are discussed in terms of what this chunking theory might mean for familiar and non-familiar numbers.

4.2.1 Eye-Hand Span

The eye-hand span measures the distance between the character being typed by the typist and the character being viewed by the typist. This span represents how far ahead in text a typist will look when transcribing in order for them to type at a normal rate.

The eye-hand span is examined by testing a range of preview sizes whilst asking a participant to type text. The smallest preview size at which the typist is still able to work at their normal speed represents the eye-hand span. The eye-hand span for words is greater than the eye-hand span for non-words (Phenomenon 16 Salthouse (1986)).

If the number of characters that a typist can see is reduced, that is, if the preview window of the material they are copying from is made smaller, there is a detrimental effect upon the typing speed and the typist slows down (Salthouse & Saults, 1987). A preview window that is smaller than a typist's natural eye-hand span will result in a slowing of the typist's speed. This is due to the parallelisation taking place during a typing task: we are able to view and take in material whilst still being able to type earlier material.

Indeed it is this parallelisation, combined with our ability to chunk incoming text, that is thought to explain the eye-hand span. This represents the perceptual processor looking ahead to store chunks for the cognitive processor to pass on to the motor processor. When the preview window is made smaller, the chunking process cannot function in the same way. If the typist cannot see the entirety of a word, because the latter part of it is not in the preview window, then the word can no longer be chunked as a whole and must therefore be chunked into smaller units, as a syllable or letters.

Applying this to the TYPIST model, it is possible to see how this would result in a slowing down. The model only has a limited capacity for processing chunks of text and, as such, can only hold a limited amount. If the chunks become inefficient, that is, they begin to only hold parts of words, then the typing slows down as the amount of information passing through the processors diminishes.

When the material a typist has to copy becomes sets of meaningless and random characters, the eye-hand span decreases (Salthouse, 1984). This again is due to inefficient chunking strategies. Meaningless material requires the typist to store the non-words as syllables as there is no long term memory version of the word to refer to. Non-words that can't be pronounced require even smaller chunks. As the chunks are smaller, it is less likely that a smaller eye-hand span will split them, meaning that typists can proceed at their normal rate even with a small preview window. This results in a smaller eye-hand span for non-words compared with words.

When considering how this will apply to the process of transcribing numbers, we can consider how familiar numbers might be chunked compared to non-familiar numbers. A number stored in memory may allow a typist to chunk the full number, rather than having to rely on breaking it up into smaller digit chunks. A non-familiar number however, could not be stored as a whole chunk. Much as the smaller chunks lead to a smaller eye-hand span with non-words, this would mean that the eye-hand span for non-familiar numbers would be smaller than that for familiar numbers.

4.2.2 Replacement Span

The replacement span measures the point at which a character is committed to memory and moved from the perceptual buffer. It is tested by changing single characters in the input text at various distances away from the current character being typed. For example, a typist might be asked to copy the sentence "I own a cat". When the typist has finished typing the word 'own', the word 'cat' might be changed to 'hat'. The typist will then perform one of two actions: type 'cat' if they had already committed the word to the cognitive processor, or 'hat' if they had not. The closer the altered character is to the letter being typed, the more likely it is that the typist will type the original character, because the typist has already converted that character into a keystroke command. The 1986 Salthouse paper suggests that the replacement span "can be assumed to reflect the point at which typists commit themselves to particular characters". Studies consistently place the copy span at around three characters.

The replacement span tests a typist's confidence in the text they are copying. If a typist is familiar with the text, they are more likely to create larger chunks and move on to read the next portion of text. Unfamiliar text is more likely to result in typists creating smaller chunks, and thus looking at the text they are copying more often, meaning that changes are likely to be noticed.

There is at present no study detailing the difference in replacement spans for words and non-words, however studies have been conducted with skilled and non-skilled typists. These studies have shown that skilled typists have a larger replacement span than novice typists (Phenomenon 27 Salthouse (1986)). Skilled typists are able to commit words to the motor buffer more quickly as they are familiar with, and have greater experience of typing them. Novice typists do not have this experience and so the replacement span is smaller.

It is possible to extrapolate from this finding that the replacement span will be smaller for non-words than it is for words. Skilled typists have a lot of experience with typing words, whereas non-skilled typists lack this experience. All typists lack experience of typing non-words and so it may be the case that typing non-words is similar to typing “without skill”. We may, therefore, assume than non-words will have a smaller replacement span, much as non-skilled typists have a smaller replacement span.

Furthermore, given that typists are more likely to be familiar with words, than numbers, it is likely that words will have a larger replacement span than numbers. If this familiarity effect holds for numbers, it it hypothesised that familiar numbers will have a larger replacement span than non-familiar numbers.

4.2.3 Copy Span

The copy span tests how much material the typist can recall when the text they are copying from disappears. It is suggested that the copy span is representative of a temporary input buffer (Salthouse & Saults, 1987) and is “expected to vary in size with the familiarity [...] of the material”. This is due to the copy span’s relation to chunking and working memory: the more chunks a typist is able to hold in memory, the larger their copy span will be. The less familiar the material becomes, the smaller the chunks will get, and the smaller the copy span will be.

Previous research reports a wide range of values that the copy span can take. In the 1986 Salthouse paper, copy spans of 7-40 characters are reported, yet in the updated paper in 1987, they report spans averaging 6.6 characters, which is smaller than was previously found. This is thought to be due to the predictability of the text. In the first study, the text used semantically meaningful sentences. The narrative of semantically meaningful text allows whole sentences to be stored in working memory. In the second experiment, the text was a randomised list of words. This lack of predictability and meaning is likely to have reduced the typist’s copy spans. This shows that familiarity of material will affect the copy span of a typist; the more familiar the material, the larger the copy span. A similar finding was reported in an experiment conducted by Rothkopf (1980). This copy span experiment found that the difficulty of the text affected the copy span: texts that were difficult to read due to unfamiliar content decreased the copy span of the typists. These findings suggest that familiarity of targets might affect copy span.

Considering how this would affect how numbers are typed, we might expect them to have a lower copy span than words, as there is rarely any predictability between numbers. It can also be hypothesised that internal predictability will play a part; numbers that are familiar may be encoded as larger chunks, meaning more information can be stored, which will lead to larger copy spans for familiar numbers compared to non-familiar numbers.

4.2.4 Interkey Interval

The speed at which users type can be measured with the Interkey Interval (IKI) which measures the speed between key presses within a word. Typing speed during a transcription task is ordinarily faster than reaction time (Phenomenon 1), as it is not a case of reacting to a letter and typing it, but is a process facilitated by previous knowledge of how words are composed. The cognitive processor can retrieve the memory of how a word is spelt without having to rely on the perceptual processor providing it with one letter at a time. However, when the perceptual processor takes in only one character at a time, as with random source material, the cognitive processor cannot facilitate the process by doing any more than passing the character on to the motor processor and therefore the speed of typing reduces to something closer to reaction time. The speed of typing is, therefore, affected by the familiarity of the text.

The speed at which text is copied is key to understanding how it is being processed. Faster typing speeds suggests larger chunks are being formed from the text. Salthouse suggests the slowing effect is due to the fact that when text becomes unfamiliar, typists are no longer able to rely on well-known text strings to help encode the text they see. This results in the text being broken down into smaller chunks.

The IKI becomes slower as the text that is to be copied becomes less structured, and features fewer legal words (Phenomenon 5); a finding which has been replicated many times. Importantly, this effect has been replicated when the digrams within the words are controlled for, that is, that the pairs of letters in both the "word" text and "non-word" text are the same (Salthouse, 1984). This implies that this effect is not merely measuring the typists' motor abilities, and their familiarities with typing certain letter combinations, but is also affected by their familiarity with the material. It is known that real words are typed most quickly. Legal non-words are slower and non-words comprised of randomised character strings are typed slower still (Phenomenon 5).

When considering how numerical transcription typing fits into this spectrum of typing speeds, it is firstly hypothesised that numbers will be slower than words. The numbers are less familiar than words and are, therefore, likely to be chunked less efficiently. However, within the set of numbers typed, it is hypothesised that familiar numbers will be typed more quickly than non-familiar numbers, as they will benefit from the support of long term memory representations. These representations will mean that the numbers will be encoded more quickly, as they will not require constant information from the perceptual processor. Non-familiar numbers however, are more likely to be typed at a similar speed to non-words, as both will be processed on a character-by-character basis, without support from any representations in memory.

4.2.5 Error Rate

The error rates will be monitored throughout the experiments. There is little research to suggest that error rates for words and non-words would be different and there is consequently no directional hypothesis for the difference in error rate between familiar and non-familiar numbers.

It may be that familiar material allows typists to check for errors more easily, without having to refer back to the original text. The typists would have a representation of the familiar text in their long term memory and, as a result, would be able to notice errors more easily in the text they had typed. In order to recognise errors in the non-familiar text, a comparison would need to be made between the material being copied and the text being entered. Comparing sources in this way increases information access cost and a user is less likely to perform this check (Back et al., 2012). However, it is also possible that typists may take more care when copying the non-familiar text, a phenomenon that has also been documented (Logan, 1999). Other studies however, have shown that familiarity with the text does not necessarily have any effect upon error rates (Inhoff, 1991).

Despite this, error rates are an important metric, as they are strongly linked to typing speed. It would not be possible to say that in some condition, typing speed was fast and yet ignore the number of errors made. Often a quick performance is at the expense of accuracy, and high speeds lead to high error rates. Monitoring the errors during this experiment will ensure that any speed increases are not due to accuracy losses.

4.3 Investigating the phenomena

In this chapter, four experiments will be presented that test each of the three span-related phenomena listed above. Additionally, the IKI and error data will be recorded during each of the three experiments. Each experiment is evaluated separately before discussing the implications of all the results upon our understanding of the cognitive processes that occur during numerical text transcription and whether these are affected in any way by the familiarity of the numbers being transcribed.

Each experiment will test typists' ability to not only copy numerical text, but alphabetic text too. This will act as both a check that the experiment has been designed well and replicates previous results, but also to compare results between numbers and words. In addition to investigating the effects of familiarity on number typing, it is also important to know how numerical typing compares to alphabetic typing.

The statistical tests performed on the results in this chapter will firstly aim to establish whether there are significant differences between the spans for words and numbers. It is hypothesised that the numerical spans will be smaller than those for words. To test this, ANOVA tests will be performed.

The primary aim of this thesis, however, is to determine if there are differences in the typing metrics for familiar and non-familiar numbers. This will be examined using planned comparisons.

All reported spans are measured in characters. All reported IKIs are measures in milliseconds. Error rates are reported as a percentage of possible numbers or words that did not match their targets. Error bars shown on the graphs are generated using standard error.

4.4 Study 3: Eye-Hand Span

4.4.1 Overview

The first experiment will investigate eye-hand span, as this measurement has been previously applied to both words and non-words. This will ensure that the experiment is able to recreate previous findings.

The eye-hand span is representative of the parallelisation of the perceptual, cognitive and motor buffers and highlights the different chunking strategies used when typing real and non-real words. Eye-hand span represents how large a preview window is required for a typist to transcribe at a normal rate. The preview window is the amount of material that a typist can see of the text that they are copying from. Smaller window sizes result in slower typing. The familiarity of a word also affects the eye-hand span: non-words have smaller eye-hand spans than words.

Hypothesis

The eye-hand span for numbers will be lower than that for words. The participants are likely to be more acquainted with the task of reading and typing words and so are likely to look further ahead. They are also likely to be able to recognise the alphabetic word strings more easily than numbers, meaning they will be processing larger chunks from the text, causing a larger eye-hand span. This will also cause the eye-hand span to be greater for familiar numbers than for non-familiar numbers. The more familiar numbers will allow the participants to read more of the text and create larger chunks for processing.

4.4.2 Method

Participants

In total, 117 participants completed the study (43 female, 1 other, 6 undisclosed) with a mean age of 25.31 (SD = 9.33). The participants were recruited from two locations: the university subject pool and the aggregation website www.reddit.com. More detail into the recruiting methods can be found in Appendix B.

Design

A 2 x 2 (target-type x familiarity) within-subjects design was used. Targets were words or numbers. For each of these a familiar and non-familiar target was used (see materials for more details). The dependent variables were: Eye-hand span, Interkey Interval (typing speed) and uncorrected errors.

Eye-hand span was measured in terms of number of characters. Eight eye-hand spans were tested between one and eight. Calculation of the eye-hand span is discussed in the results section.

Interkey interval was measured in milliseconds. Only trials resulting in a correct submission were counted in this measure.

Uncorrected errors are reported as an error rate, representing the percentage of words or numbers submitted that did not match the target.

Materials

During the study, participants were required to transcribe both numbers and words that were familiar and non-familiar.

The 80 numbers used in the experiment were sourced from results of the study conducted in Chapter 3. From this data set, 40 familiar and 40 unfamiliar numbers were chosen.

The familiar words used were generated using the Paivio et al. Word List Generator[1]. These words were all strings found in the Oxford English Dictionary and had a high frequency rating. The forty unfamiliar words were generated using the ARC Nonword database[2]. These were non-legal but pronounceable words that adhered to orthographic rules.

Table 4.1 shows the distribution of word and number lengths. The differing lengths were not thought to be an issue, as previous research showed that IKI is not affected by word length. The length 8 words were included so as to fully explore the effect of preview windows. The lengths of numeric targets used in this experiment were determined by the available numbers collected in the study in the previous chapter.

Length	**Numbers**	**Words**
2	3	0
3	14	3
4	21	5
5	2	8
6	0	12
7	0	8
8	0	4

Table 4.1: Table showing word and number lengths used during this experiment

[1] http://www.datavis.ca/online/paivio/

[2] http://www.cogsci.mq.edu.au/ nwdb/

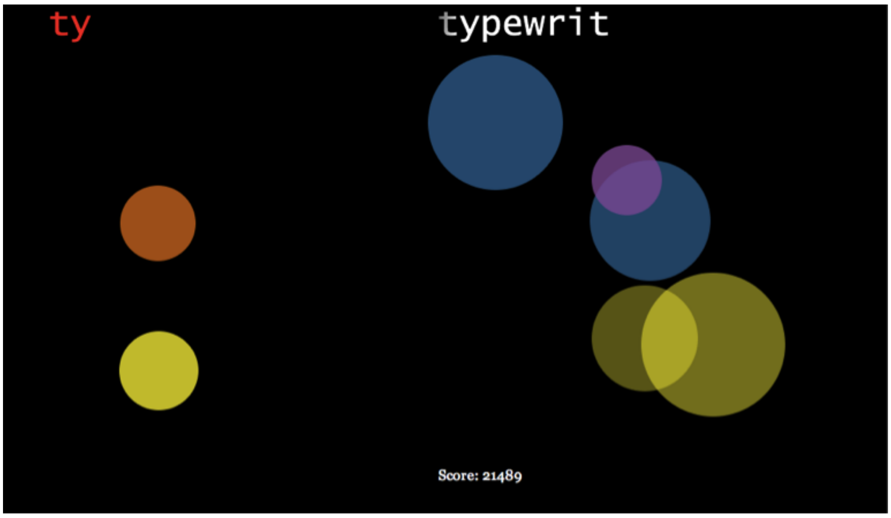


Figure 4.2: Screen shot of the eye-hand span experiment. The preview window is too small to show the end of the word "typewriter".

The text was presented to the participants on a black webpage. The text that participants were to transcribe was presented in white, 30pt Lucida Console. This text was displayed in the top, centre of the screen. The text that the participants entered was displayed at the top left of the screen in red 30pt Lucida Console.

As the participants transcribed the words and numbers, the white target text progressed character by character. Each character entered by the participant caused one character of the white text to fade and move one position to the left. This allowed the stream of text to update as the participants progressed through the text. The character that the participant had to type next was always the first bright white character in the text stream.

An underscore character was used instead of a space between words or numbers. This was due to participants in the pilot study mistakenly assuming they had reached the end of a word or number when the preview window obscured the ending of a word. An example of this can be seen in Figure 4.2, where the end of the word "typewriter" has been cut off by the preview window size. The underscore acted as an 'end of word' marker.

In order to test the participants' eye-hand span, a number of preview windows were used throughout the experiment. Participants were presented with a block of five familiar followed by five non-familiar targets per preview window size (this order was randomised). This block was presented with a set preview size between one and eight. At the end of each block, a new preview size was used. This was repeated eight times, so each preview size was tested with a block of ten targets. The order of the numbers and words presented to the participants was randomised for each trial.

The experiment allocated the participants a score, which was dependent upon their speed and error rate. The score was increased for fast typing speed: this was calculated as an inverse function of the time that the participant took to complete the word; the faster they were, the higher their score. Correct words and numbers also meant that a new coloured circle appeared on screen. The coloured circles were taken from a pastel-tone colour palette and slowly glowed on screen, moving with a floating-like animation. The aim was to create a more pleasing ambient environment for the experiment. These circles had a limited life span, and would disappear eventually. Incorrect responses resulted in half of the circles on screen disappearing. There was also a fixed points penalty applied to the score for incorrectly typed words and numbers. The score was used solely as a means of motivating the participants; it was not analysed as was not a dependent variable of the experiment.

Procedure

The experiment was advertised through four different methods. An advert was placed on a university subject pool offering participants £3 for their time to come into the lab to complete the experiment.

A later advert was placed on the same subject pool offering the same compensation for completing the experiment online.

Two more adverts were placed on the aggregation website Reddit[3]. These adverts were placed by different accounts in different locations. One advert offered compensation in the way of data and analytics about the participant's performance during the experiment. The other simply asked for help in a research project and offered no compensation. More information on these processes can be found in Appendix B.

All participants, regardless of location, were provided with a hyperlink that took them to the correct website for the condition they were in depending on whether they would see numbers or words first.

Upon loading the experiment in the required Google Chrome browser, participants were given information about the study and asked only to continue if they did not have RSI or other issues that would be exacerbated by typing.

After confirming they were fit to take part, participants were given instructions on the task. This included a text description and a video demonstration. The video contained subtitles and highlighted key elements of the task. Prior to each experiment, participants completed a trial period of 10 words or numbers.

[3] www.reddit.com

Participants then proceeded to the main body of the experiment where they were asked to transcribe 40 words or numbers. Once completed, participants were given the chance to rest before beginning the experiment in the alternate condition.

Once the experiment was completed in both conditions, the participants were taken to a debriefing page where their demographic details were collected. A text description of the research was provided.

Depending upon the method of compensation, at this point some participants were given a link to their data and shown interactive graphs of their performance. The experiment took no more than 30 minutes to complete. Participants were able to leave early if they wished.

4.4.3 Results

One participant's data was removed from analysis as they had skipped many of the trials. The analysis is therefore performed on 116 participants.

For statistical analysis a 2 x 2 (target-type x familiarity) repeated measures ANOVA was used to analyse each of the dependent variables, with a .05 significance level for judging the significance of effects.

Eye-hand Span

Eye-hand span shall be considered first. A participant's eye-hand span was calculated using the same algorithm as used in the original Salthouse work (Salthouse, 1984). This algorithm defines the eye-hand span as follows:

> "The smallest window at which the first quartile was greater than the second quartile of normal typing. This procedure effectively identified the span as the number of display characters at which 75% of the interkey interval exceeded the median interval from normal typing."

In order to apply this algorithm, normal typing is defined as the typing speed when the preview window was set to 8 characters.

The results of this analysis are shown in Figure 4.3. Participants had a marginally larger eye-hand span for familiar text (M = 3.15, SD = 0.81) than for non-familiar text (M = 2.97, SD = 0.78), this difference was significant F(1, 115) = 4.27, p = .041, η^2 = 0.04. Participants had larger eye-hand span for words (M = 3.28, SD = 1.04) than for numbers (M = 2.84, SD = 0.80), this is also significant, F(1, 115) = 12.87, p <.001, η^2 = 0.1.[4]

[4] The interaction effect between target type and familiarity was non-significant F(1, 115) = .66, p = .418, η^2 = 0.01.

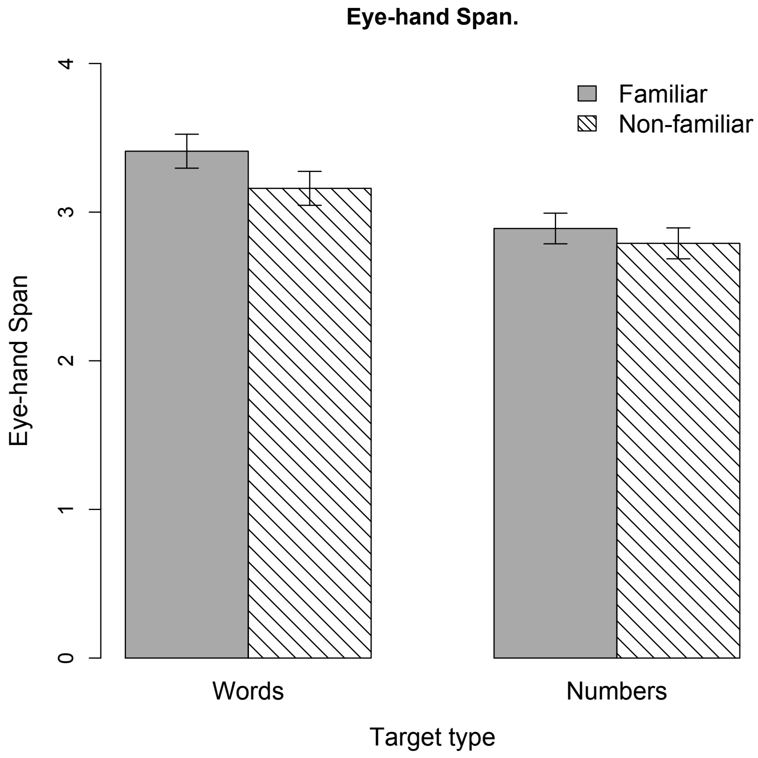


Figure 4.3: Mean eye-hand span for participants in all conditions

A planned comparisons test of familiarity for words and numbers was performed which showed that there was a significant simple effect of familiarity for words F(1, 115) = 4.59, p = 0.034, η^2 = 0.04 and not for numbers F(1, 115) = .441, p = .508, η^2 = 0.004.

Interkey Interval (typing speed)

IKI considers the time delay between two consecutive, correct keypresses (excluding the first character of a word or number). This is a reflection of typing speed. This data is shown in Figure 4.4.

Statistical analyses of these data show that participants had significantly faster IKI for familiar text (M = 291, SD = 123) than for non-familiar text (M = 382, SD = 143), F(1, 115) = 70.67, p <.001, η^2 = 0.38. Participants also had a faster IKI for words (M = 305, SD = 104) than for numbers (M = 367, SD = 140), F(1, 115) = 148.02, p <.001, η^2 = 0.56. [5]

The simple effects analyses found that familiarity had a significant effect for words F(1, 115) = 104.90, p <.001, η^2 = 0.48, and also for numbers F(1, 115) = 30.85, p <.001, η^2 = 0.21.

[5] The familiarity by number interaction was not significant, F(1, 115) = 2.71, p = 1.03, η^2 = 0.02.

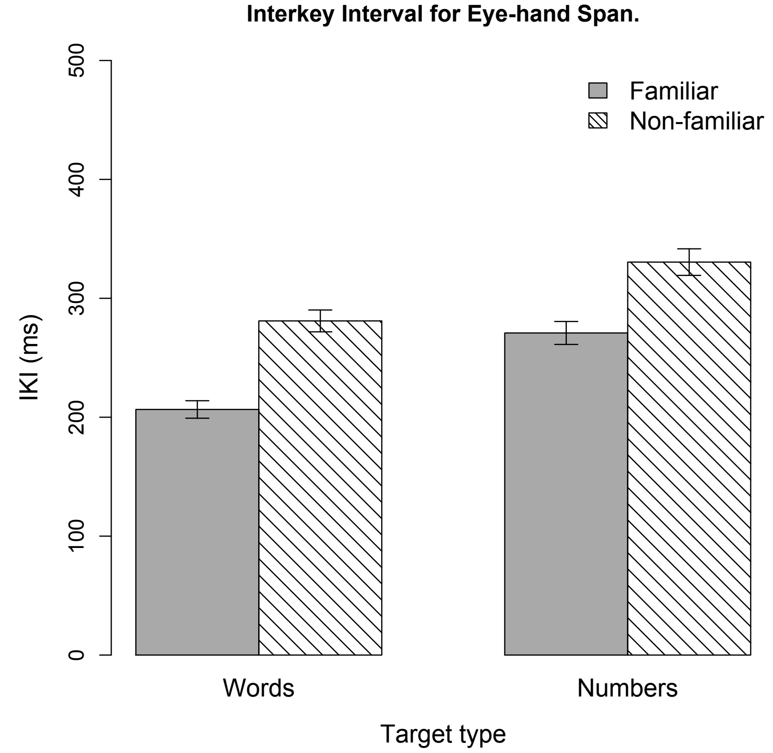


Figure 4.4: Mean interkey interval for participants in all conditions

Errors

The error rate reports the percentage of words and numbers that were entered by participants that did not match their target. Error rates are shown in Figure 4.5.

Error rates for familiar targets (M = 3.87, SD = 3.42) were lower than for non-familiar targets (M = 5.12, SD = 3.81). A statistical analysis shows this difference to be significant F(1, 115) = 16.80, p <.001, $\eta^2 = 0.13$. There was no significant effect of type F(1, 115) = .007, $p = 0.935$, $\eta^2 = 0$.[6]

4.4.4 Discussion

Eye Hand Span

In previous research, the eye-hand span has been found to be 3.45 for words and 1.75 for non-words (Salthouse, 1984). In the experiment in this chapter, the eye-hand span for familiar words (the comparable condition) was 3.41 and for non-familiar words was 3.16. This result is similar for familiar words, but larger than expected for non-familiar words. This is likely due to the types of non-words used in this experiment. In previous

[6] There was no interaction effect F(1, 115) = .141, $p = .708$,$\eta^2 = 0.001$.

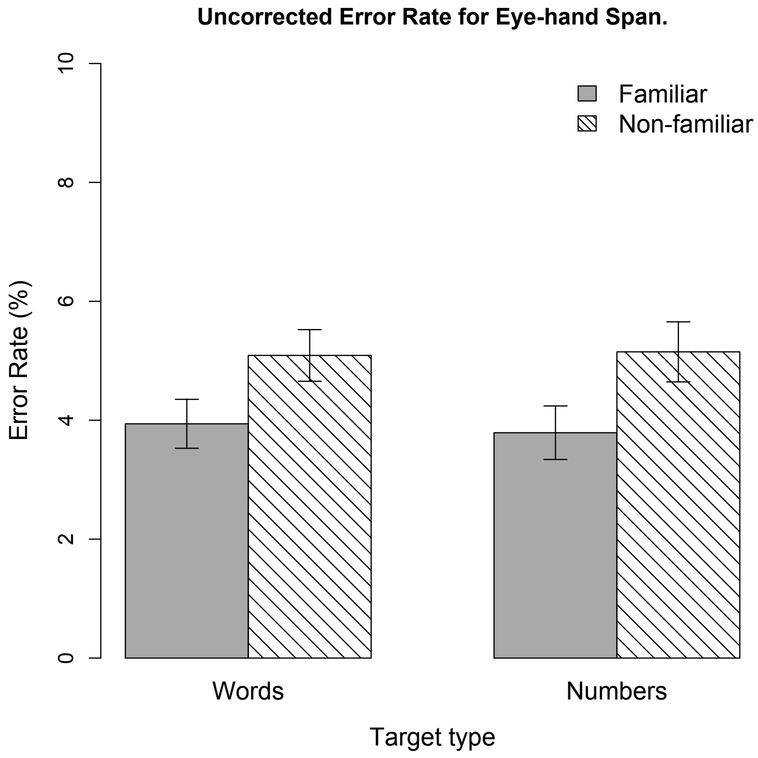


Figure 4.5: Mean uncorrected error rates for participants in all conditions

eye-hand span experiments, the non-words were non-legal non-words, that is, they did not adhere to orthographic rules and were, therefore, unpronounceable. Whereas in the present experiment, the non-words all adhered to orthographic rules and were easily pronounceable. This will have meant the non-words were more easily chunked (if not into full words, then into syllables), which would increase the eye-hand span. In previous experiments, where the non-words were unpronounceable, the chunks would have been at the letter-by-letter level, thus making the eye-hand span smaller. In order to investigate this, the experiment will be repeated using non-legal non-words.

Although the mean eye-hand span was higher for familiar numbers, there was no significant difference in the eye-hand span for familiar and non-familiar numbers. This suggests that the two may rely upon the same chunking method. This will be investigated a second time in the second iteration of the eye-hand span experiment (Study 4).

Interkey Interval

The analysis of IKI supports the hypotheses that words would be transcribed faster than numbers and that familiarity would also result in a speed increase during a transcription task.

Non-familiar words were typed notably faster than non-familiar numbers which was not predicted. However, this may be explained by the type of non-word used in this experiment. As stated before, the non-words used in this experiment were pronounceable, which has not been the case in previous research (Salthouse, 1984). This may have made them more easily memorised and thus the cognitive processor would have been able to facilitate the recollection process more easily for these non-words than previously. Indeed, Salthouse finds that as material approaches randomly arranged characters, the typing speed slows. The non-words used in this experiment appear more random than the words, in that they will not have been seen by participants before, but will appear less random than the non-words used in the previous iterations of this experiment (Salthouse, 1984), which used unpronounceable strings. This will be tested once more in the second eye-hand span experiment, where the non-words will again be unpronounceable.

Error Rate

The error rate analysis in this experiment produced significant results. It was found that non-familiar targets incurred more errors than familiar targets. This finding will be explored further when all experiments are considered together. Based upon initial suggestions for change in error rates however, this suggests that participants were less likely to check for errors in the non-familiar targets.

One reason for measuring error rates was to ensure that any increases in speed were not due to a loss of accuracy. Here, it can be seen that non-familiar targets were more error-prone, however, this has not translated into a gain in speed. The IKI statistical report shows that non-familiar target types were slower, showing that the higher error rate did not cause an artificial boost in speed.

4.4.5 Conclusion

The results from the eye-hand span for familiar words replicated those found in previous studies of the eye-hand span, however, the non-familiar results did not. This suggests the experiment is not a fair replication of the previous research and should be repeated with a different set of non-words. The non-words used in the current experiment were pronounceable, which may have caused them to be more easily chunked, which increased the eye-hand span and IKI. This follow-up experiment is presented next.

4.5 Study 4: Eye-Hand Span Replication

4.5.1 Method

Participants

Twenty participants took part in this study (10 female) with a mean age of 23.45 (SD = 3.2). The participants were recruited online from a university subject pool.

Design

The design of this study was the same as that used in the previous eye-hand span experiment.

Materials

The 80 numbers in this experiment were the same as those used in the previous experiment. A new set of words was generated using the Paivio et al. Word List Generator[7]. In this experiment, the word lengths matched the number lengths. The 40 unfamiliar words were generated using the www.random.org string generator. These non-words were different to those used in the previous study as they were not orthographically legal non-words, and were thus unpronounceable.

All four texts were 142 characters long. Each set was comprised of the same amount and length of words and numbers. Table 4.2 shows the distribution of word and number lengths.

Word/Number Length	Amount
2	3
3	14
4	21
5	2

Table 4.2: Word and number lengths used in the eye-hand span study

The materials for this experiment differed from the words used in the previous eye-hand span experiment as the non-words were illegal strings, that were unpronounceable. The length of the words was also altered to match the lengths of the number strings used.

[7] http://www.datavis.ca/online/paivio/

Procedure

The study was advertised online through a university subject pool. The reimbursement advertised was a £3 Amazon voucher. The procedure of the experiment was the same as in the previous eye-hand span experiment.

4.5.2 Results

Eye Hand Span

Eye-hand span is first investigated. The eye-hand span was calculated using the same algorithm stated in the previous experiment.

The results of this analysis are shown in Figure 4.6. Statistical analysis shows that there was no main effect of familiarity on the eye-hand span, F(1, 19) = 2.28, p = .148, η^2 = 0.11. There was a significant effect of target type. Participants had larger eye-hand span for words (M = 3.03, SD = 0.82) than for numbers (M = 2.55, SD = 0.76), F(1, 19) = 11.85, p = .003, η^2 = 0.38.[8]

The simple effect test of familiarity for words and numbers was performed and shows that there was a significant effect of familiarity for words F(1, 19) = 5.26, p = 0.03, η^2 = 0.22 and not for numbers F(1, 19) = .000, p = 1, η^2 = 0.

Interkey Interval (typing speed)

The IKI analysis is shown in Figure 4.7. Statistical analyses of these data show that participants had significantly faster IKI for familiar text (M = 227, SD = 38) than for non-familiar text (M = 327, SD = 58), F(1, 19) = 76.01, p <.001, η^2 = 0.80. Participants also had a faster IKI for words (M = 257, SD = 51) than for numbers (M = 297, SD = 50), F(1, 19) = 9.67, p = .006, η^2 = 0.34.[9].

A simple effects analysis shows that familiarity has a significant effect on both words F(1, 19) = 68.88, p <.001, η^2 = 0.78 and numbers F(1, 19) = 9.49, p <.001, η^2 = 0.33. A further simple effects analysis of target type shows that target type significantly affects familiar targets F(1, 19) = 29.42, η^2 = 0.61, p <.001, but not non-familiar targets F(1, 19) = .021, p = .885, η^2 = 0.001.

[8] The interaction effect between target type and familiarity was non-significant F(1, 19) = 2.57, p = .126, η^2 = 0.12.

[9] The interaction between target type and familiarity was also significant, F(1, 19) = 9.73, p = .006, η^2 = 0.34

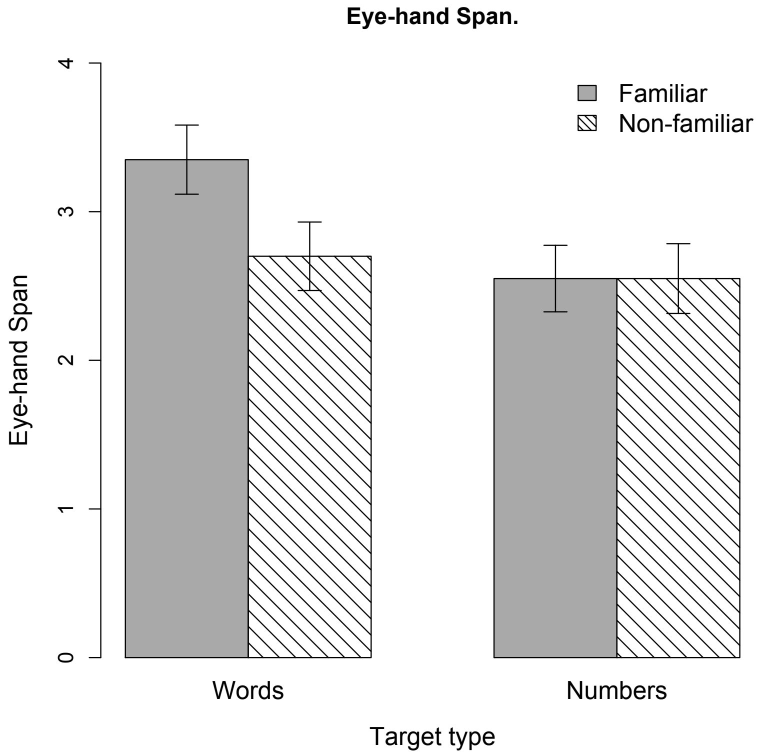


Figure 4.6: Mean eye-hand span for participants in all conditions

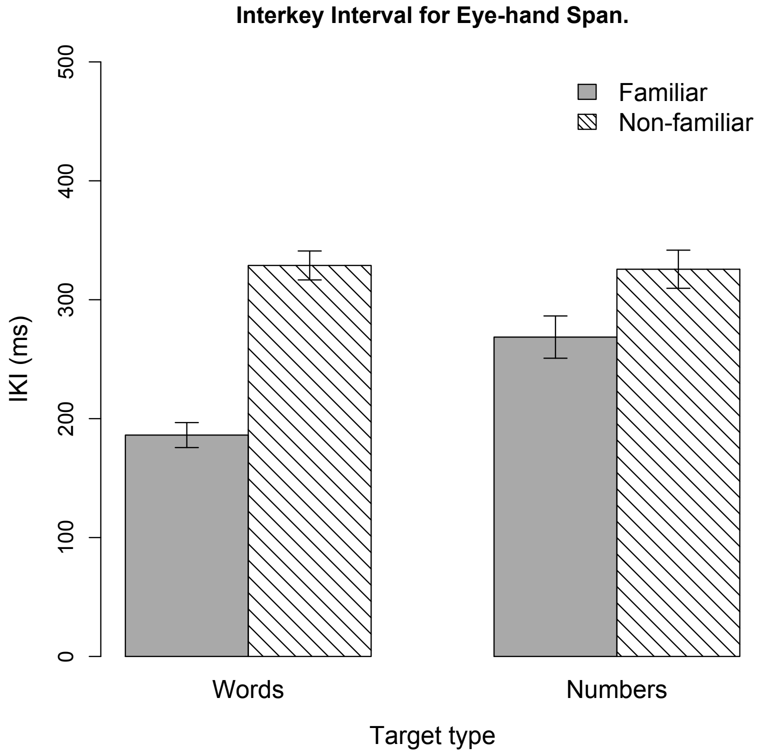


Figure 4.7: Mean interkey interval time for participants in all conditions

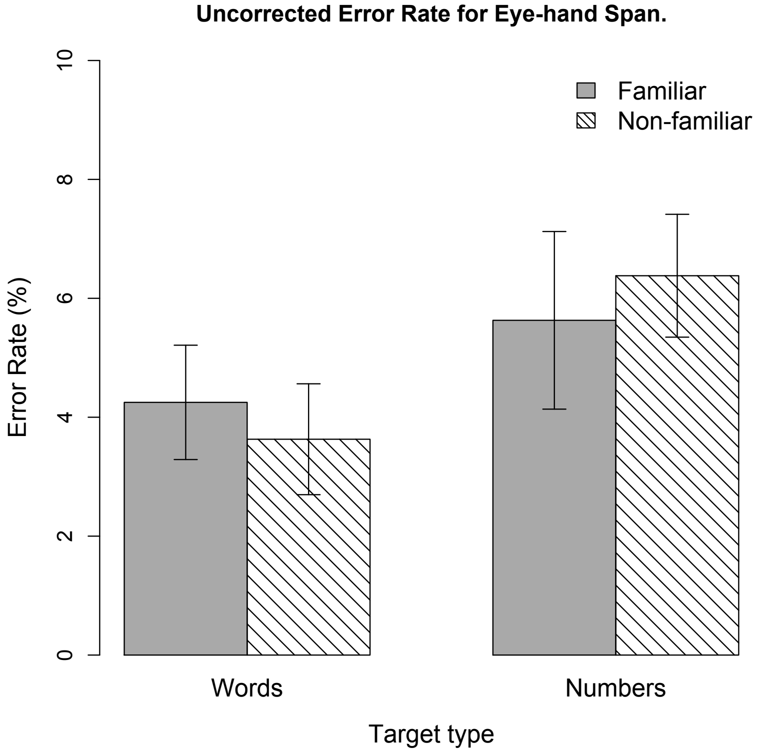


Figure 4.8: Mean uncorrected error rate for participants in all conditions

Errors

Error rates are shown in Figure 4.8. Statistical analysis shows that error rates for familiar and non-familiar targets were not significantly different F(1, 19) = .003, p= .954, η^2 = 0. There was also no significant effect of type F(1, 19) = 4.16, p = 0.055, η^2 = 0.18.[10]

4.5.3 Discussion

Eye Hand Span

In the first iteration of this experiment, there was agreement for the familiar word eye-hand span with previous research, but not for non-words. Salthouse reports an average eye-hand span of 3.7 for words and 1.75 for non-words. The updated, second eye-hand span experiment in this chapter found an eye-hand span of 3.35 for words and 2.7 for non-words. Once again, the result for words has been replicated. The eye-hand span for non-words, although smaller than previously, is still not comparable to previous results.

There is no clear reason why this difference still arises. It may be that the non-words used in this experiment were in some way different to those used in Salthouse's studies. For instance,

[10] There was no significant interaction effect F(1, 19) = 714, p = .409, η^2 = 0.04.

the word length in Salthouse (1984) varied between three to eight characters, whereas length in the current experiment varied between two to five (as a limitation of the number lengths collected in Chapter 3). Smaller words would lead to fewer words being broken by the smaller preview windows. In the present experiment, the preview windows of 5 and above would mean that all words could be presented whole, which was not the case for Salthouse's study.

Another possible explanation of the increased non-word eye-hand span in this study is the possibility that the typists in the current study are more accomplished at copying random character strings than those used in previous research. Often, early typing studies were conducted using skilled typists as participants. These typists would be very accustomed to typing formal language in the form of letters or dictations. A present day "typist", in contrast, has a wide variety of experiences with transcription typing, no longer is the skill used just in the workplace. Today, we may be asked to transcribe random character strings in the form of WiFi passwords or account details. It may be that such practice has caused modern typists to become more adept at typing non-word strings.

The eye-hand span for words was larger than numbers, as was found previously. Additionally, the familiarity in this experiment was a significant factor for the eye-hand span of words, but was not true for numbers. As before, this is likely due to the chunking strategies used for words being larger than those used for numbers. The result suggests no difference in chunking strategy for familiar and non-familiar numbers.

Interkey Interval

Previous research showed that as the material becomes less meaningful, the speed of typing slows. Here, words were typed faster than non-words, a replication of previous findings. What is interesting to note here, is how the two types of number fit into this relation. Familiar numbers were slower than words, but were faster than non-words. Non-familiar numbers were typed at a similar speed to non-words.

This similarity between non-familiar numbers and non-words was hypothesised at the beginning of these experiments. Speed of copying text is facilitated by the cognitive processor and its ability to provide information about the ordering of characters from long term memory, rather than from the perceptual processor. Neither non-familiar numbers nor non-words are represented in long term memory and, therefore, the typing task for both is a case of reacting to the characters seen, rather than recalling the arrangement of characters from memory.

Familiar numbers were faster than both non-familiar numbers and non-words. This suggests that the familiar numbers are being facilitated by some form of cognitive look-up during transcription. The results from the IKI measurements, therefore, suggest that familiar and non-familiar numbers are processed differently during transcription.

Error Rate

The error rate for the experiment showed no significant effects of either target type nor familiarity. This suggests that any improvements in speed were not due to losses in accuracy.

Prior to the experiment, two possible effects of familiarity on error rate were suggested: one increasing and the other decreasing the number of errors. It was possible that the representation of familiar text in the long term memory would allow typists to check their errors more easily. However, it was also possible the less familiar text would result in fewer errors due to extra care taken by the typist. The results from this experiment provide support for neither of these theories exclusively. They suggest instead that familiarity is either having no effect upon error rate, or it is having an equal and opposite effect upon error checking for familiar and non-familiar targets.

The ability for a typist to check their errors during this experimental design is limited. Once a word or number has been typed and the enter key pressed, that word cannot be seen again. Typists taking a global strategy to error checking would, therefore, not be able to look back and check their work. The only error checking possible during this experiment was local error checking, performed after typing each word. An experimental design that allowed for both error checking strategies may have provided more information about true error rates.

4.5.4 Limitations

This study used the preview window method to measure eye-hand span. It is possible, however, to use an eye-tracking device. Indeed, this is a common technique when analysing how musicians look ahead whilst reading from sheet music (Truitt, Clifton, Pollatsek, & Rayner, 1997). This may have provided more accurate results for the eye-hand span whilst typing, as it would involve directly measuring the participants rather than testing various spans incrementally. However, for methodological purposes, eye-tracking was not a viable option in these experiments. The experiments in this chapter were conducted online, which allowed for large numbers of participants to be gathered and tested in a short period of time. Although beneficial for recruitment, this method of running the experiment meant that an eye-tracker could not be used.

4.5.5 Conclusion

This experiment partially replicated previous results with regards to the eye-hand spans for words and non-words in that the span for words was larger. However, the absolute value for the eye-hand span for non-words is larger than reported previously (Salthouse, 1984). This may be due to differences in modern typists.

Both experiments failed to highlight a difference in eye-hand span between familiar and non-familiar numbers. This suggests there may be no difference in chunking strategies between familiar and non-familiar numbers, however, both experiments within this section reported significant differences in the IKI for familiar and non-familiar numbers, suggesting that familiarity does affect the transcription of numbers.

This experiment hints at potential differences in cognitive processes for familiar and non-familiar numbers through the different IKI values for both. However, the test for differences in the eye-hand span has not been successful. This may be a function of the experimental design. The difference between the spans for familiar and non-familiar numbers may be smaller and so many have been lost at this level.

Future experiments within this chapter will investigate firstly, whether the familiarity of a number affects other typing spans and secondly, will aim to replicate the difference in IKI.

4.6 Study 5: Replacement Span

4.6.1 Overview

The replacement span measures the point at which typists notice dynamic changes within the text. When a character is replaced in the text that a typist is copying, they will either type that new character, or will have already committed the character to memory and the original character will be typed.

Hypothesis

The replacement span for words will be greater than the replacement span for numbers. The replacement span for familiar numbers will be greater than the replacement span for non-familiar numbers.

4.6.2 Method

Participants

Twenty participants were recruited from a university online subject pool. Participants were on average 25.15 years old (ranging between 19 and 57, *SD* = 9.07). Fourteen of the participants were female.

Design

A 2 x 2 (target-type x familiarity) within-subjects design was used. Targets were words or numbers. For each of these, a familiar and non-familiar target were used (see materials for more details).

The dependent variables were: Replacement Span, Interkey Interval (typing speed) and uncorrected error rate.

Replacement span was measured in terms of number of characters. Three replacement spans were tested: 1, 3 and 5. Participants were assigned the highest replacement span at which they were most likely to type the original character.

Interkey interval was measured in milliseconds. As before, only trials resulting in a correct submission were counted in this measure.

Uncorrected errors are reported as an error rate, representing the percentage of words or numbers submitted that did not match the target. Note that when a word or number contained a replaced letter or digit, both the original and new character were considered correct.

Materials

During the study, participants were required to transcribe both numbers and words that were familiar and non-familiar.

The 80 numbers and 80 words used in the experiment were the same as were used in the second eye-hand span experiment (Study 4).

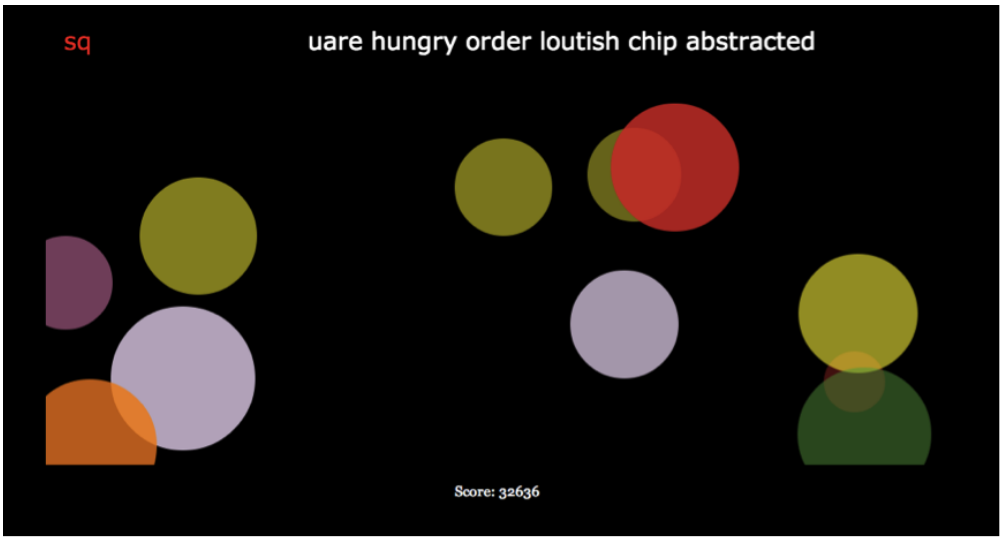
The text was presented to the participants on the same black webpage as used in previous studies, using the same text layout. The text was presented in 20pt Verdana. The text stream updated as the typist copied the words, this worked the same way as in the eye-hand span experiment. One variation was that the preview window of text was set to 35 characters, and so none of the words or numbers were ever split. This meant that the underscore method of signalling the end of a word was not necessary.

The participants completed four blocks of text transcription. One for each of the (word, number)×(familiar, non-familiar) conditions. The order of these blocks was randomised for each participant. The participants had to copy a stream of 40 words or numbers.

A screenshot of the trial period of the task can be seen in Figure 4.9. At this moment the participant is mid-way through typing the word "square". The text that the participant has entered, "sq" is in red in the top left. The word in the text stream has progressed accordingly and now displays the word "uare", letting the participant know which characters she has left to transcribe.

The study was designed to measure the participants' replacement spans. To do this, characters in the white stream of text were replaced at different intervals. Three possible replacement spans were tested: 1, 3 and 5 characters. These spans were the same as those used in previous replacement span studies. Each of these spans was tested 3 times per condition.

During a test of the replacement span, a character would be changed within the word. When testing a replacement span of 3, for instance, the character that was 3 away from the currently typed character would be changed. In the example given in Figure 4.9, this would mean the letter 'r' in the word "square" would be altered to a different letter, for example, it could be changed to the letter 'm'. The test then checked whether the participant typed the original 'r' or the updated 'm'. This letter switching occurred when the character was 1, 3 and 5 characters away from the current character. The position of the switched character


sq

uare hungry order loutish chip abstracted

Score: 32636

Figure 4.9: Screen shot of the replacement span experiment. The participant has typed the first two characters of the word "square" (shown in red). The white text shows what they have yet to transcribe.

within the target was also altered: the switched character would be either the first, second or third letter of the word or number.

The alteration occurred as soon as the participant had typed the character that was that particular span away from the character being changed. The position of the alterations was spaced out at semi-random intervals throughout the experiment.

In each condition, the familiarity of the word or number was maintained as far as possible. When a letter was switched within a word, it was ensured that the new letter still resulted in a legal word. The same care was taken to preserve the non-word status of the non-familiar words.

For numbers, maintaining the status was more difficult, as there was only a limited set of familiar and non-familiar numbers available. For this reason, subjective decisions were made when ensuring that the new digit still created a familiar number. Rules learnt from the card-sorting exercise in Study 2 provided guidance when making this switch. For instance, some numbers that appeared to represent a year were swapped with similarly plausible year numbers. Care was taken to ensure that the non-familiar numbers were replaced with numbers that did not appear to be familiar. The same scoring mechanic was used as in the eye-hand span experiments.

Procedure

The experiment was advertised as an online study through a university subject pool website. Participants were told they would receive a £3 Amazon voucher for taking part in the experiment.

Participants were provided with a hyperlink that took them to the correct website for the condition they were in: half performed trials with numbers first and half with words first.

As in the previous experiments, participants were given instructions and a briefing video. Participants then completed a trial where they transcribed 35 words or numbers depending on the condition they were currently in. During this trial period, 5 characters were replaced.

Participants then proceeded to the main body of the experiment where they were asked to transcribe 40 words or numbers. Once completed, participants were given the chance to rest before beginning the experiment in the alternate condition.

Once the experiment was completed in both conditions, the participants were taken to a debriefing page where their demographic details were collected. A text description of the research was provided.The experiment took no more than 20 minutes to complete. Participants were able to leave early if they wished.

4.6.3 Results

A 2 x 2 (target-type x familiarity) repeated measures ANOVA was used to analyse each of the dependent variables, and to investigate simple effects.

Replacement Span

Replacement span shall be considered first. A participant's replacement span was determined as the largest span at which they were most likely to type the original character. A ratio was calculated for each participant, for each span (1,3 and 5) that described the likelihood that the participant typed the original character compared to the replaced character. The last point at which that ratio was greater than 50% was deemed the replacement span.

It was possible, therefore, for participants to have a replacement span of 1, 3 or 5. A replacement span of 0 was also possible if the participant always typed the replaced character and never the original character.

The results of this analysis is shown in Figure 4.10. Participants had a larger replacement for familiar text (M = 1.95, SD = 1.00) than for non-familiar text (M = 1.10, SD = 0.70). Consistent with this, statistical analysis found a main effect of familiarity on replacement

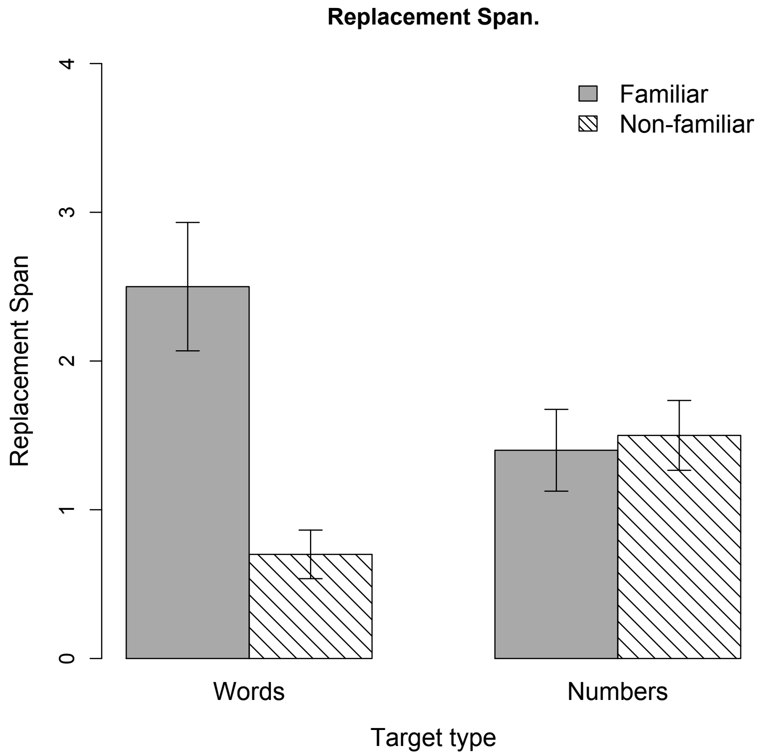


Figure 4.10: Graph showing the differences in average Replacement span in each condition

span, F(1, 19) = 13.69, p = .002, η^2 = 0.64. Participants had a larger replacement span for words (M = 1.60, SD = 1.19) than for numbers (M = 1.45, SD = 0.81). However, statistical analysis showed no main effect of target type on replacement span, F(1, 19) = .203, p = .657, η^2 = 0.01.[11]

The planned simple effect test of familiarity for words and numbers was performed. Confirming what can be seen in Figure 4.10, there was a significant simple effect of familiarity for words F(1, 19) = 22.30, p <.001, η^2 = 0.67. However, there was no significant effect of familiarity for numbers F(1, 19) = .076, p = .785, η^2 = 0.09.

Interkey Interval (typing speed)

IKI is analysed next. It considers the time delay between two consecutive, correct keypresses. This is a reflection of typing speed. This data is shown in Figure 4.11.

Statistical analyses of these data show that participants had significantly faster IKI for familiar text (M = 254, SD = 80) than for non-familiar text (M = 360, SD = 113), F(1,

[11] The interaction effect between target type and familiarity was significant F(1, 19) = 10.57, p = .004, η^2 = 0.31.

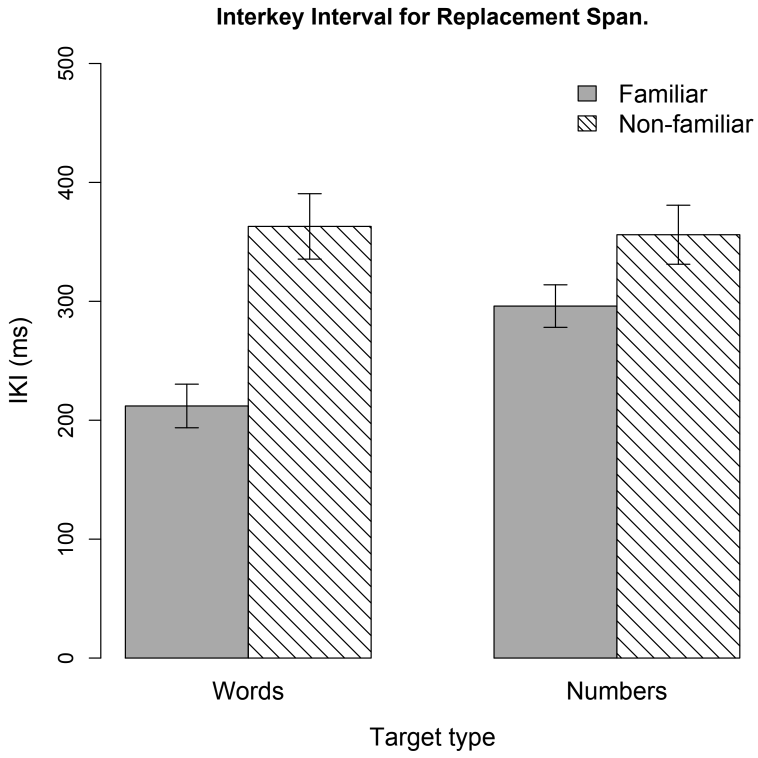


Figure 4.11: Graph showing the differences in average Interkey Interval in each condition

19) = 102.52, p <.001, η^2 = 0.84. Participants also had a faster IKI for words (M = 288, SD = 101) than for numbers (M = 326, SD = 92), F(1, 19) = 20.61, p <.001, η^2 = 0.52.[12]

The planned simple effects test showed that there is a significant simple effect of familiarity on IKI for both words, F(1, 19) = 197.49, p <.001, η^2 = 0.91, and numbers, F(1, 19) = 19.08, p <.001, η^2 = 0.50.

Errors

The error rate reports the percentage of words and numbers that were entered by participants that did not match their target. Error rates are shown in Figure 4.12.

There was no significant difference between the error rate for familiar and non-familiar targets F(1, 19) = 2.61, p = .122, η^2 = 0.12, nor was there a significant difference of error rate for numbers and words F(1, 19) = .279, p = .603, η^2 = 0.01. [13]

[12] The familiarity by target type interaction was significant, F(1, 19) = 44.77, p <.001, η^2 = 0.70.

[13] The interaction was non-significant F(1, 19) = .229, p = .638, η^2 = 0.01.

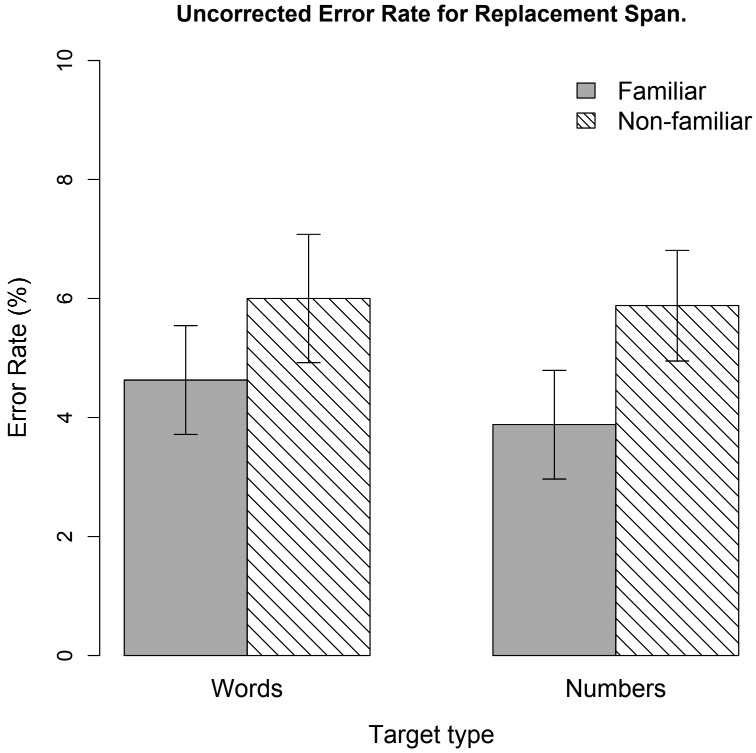


Figure 4.12: Graph showing the differences in average uncorrected errors rates in each condition

4.6.4 Discussion

Replacement Span

The replacement span reported in Salthouse (1986) is reported as being between 2.8 and 3 characters. This was based upon typists working with text comprised of legal words. The comparable condition within this experiment was the familiar words condition. Participants in this condition had a mean replacement span of 2.5 characters. This suggests that the experiment presented here is a fair replication of previous work and the results may, therefore, be compared.

The replacement span for words was found not to be significantly different to that for numbers. This is partially due to the large range in replacement span for words caused by the differing familiarity levels. Familiar words had a larger span than all number targets and non-familiar words had a smaller span than all number targets.

The fact that familiar words had a larger replacement span than numbers is not surprising. It was hypothesised that words would cause typists to use larger chunks and therefore overlook character replacements more than with the less familiar numerical targets. However, there was no reason to believe that the replacement span for non-words would be lower than for non-familiar numbers.

The reasoning behind the differing sizes of replacement spans relied upon the size of chunks used to read-in the text. Any incoming text that had some representation in memory would benefit from being read as a larger chunk. The result of this experiment, therefore, suggest that all numbers were read with larger chunks than non-words. This may be due to the fact that all numbers used in this experiment were in fact legal, that is, all the numbers, regardless of familiarity, represented a real number with a real meaning behind it. All the numbers could be processed and pronounced as a full number. This was unlike the non-words, which were random letter strings and could not be processed either lexically (pronounced) or semantically (given meaning). These differences may have caused the non-words to be read using smaller chunks and thus produced a smaller replacement span.

It was found that familiarity had a significant effect on increasing the replacement span for words. This supports the hypothesis that non-words would have a smaller replacement span than words. However, no such effect was found for numbers, where familiarity had no effect upon replacement span. This suggests that the familiar and non-familiar numbers were being read and processed in much the same way, using similar sized chunking strategies. This was unexpected, given previous research had suggested that the two were stored differently in the memory (Cohen et al., 1994) and that, in the study in Chapter 3, the response times to familiar versus non-familiar decisions was significantly different.

It is possible, however, that the granularity of measurement during this experiment was not fine enough to find any difference between the two familiarity levels of number. As previously discussed, the numbers within this experiment, including the non-familiar numbers, were all legal numbers, unlike the words used. This means that differences in reading strategy between familiar and non-familiar numbers will be smaller than those for words and non-words. In this experiment, replacement spans of 1, 3 and 5 were tested. This meant that typists with replacement spans of 2 and 4, would have appeared during this experiment to have replacement spans of 1 and 3 respectively. The lack of this experiment's ability to measure fine details like that, may have caused it to not record differences in replacement spans for familiar and non-familiar numbers.

Interkey Interval

The results of this experiment replicated those in the eye-hand span experiment, in that words were typed faster than numbers, and familiarity had a significant effect on typing speed for both words and numbers.

The IKI for non-words and non-familiar numbers is comparable in this experiment, with a statistical analysis showing that there is no significant difference between the IKI in the two conditions. The IKI in this experiment was also faster for familiar-numbers than for both non-familiar numbers and for non-words. This result is likely due to similar causes as in the second eye-hand span experiment.

Error Rate

The error rate, as in the replacement span experiment, is not significantly affected by either target type, or familiarity. This once again shows that the IKI results are not affected by increased error rates.

4.6.5 Limitations

In previous investigations of the replacement span only words have been tested. Choosing characters to replace in words is an easy task as there is a large set of words from which to choose. If one wanted to replace a letter in the word "band", one could replace any of the letters to get "hand", "bond", "bard" and "bank". The task for replacing characters in non-words is similarly straight forward; there are many ways to replace characters in the word "gsegk" and still retain the word's non-word status.

There is an issue, however, when it comes to performing the same task with familiar and non-familiar numbers. This experiment relies upon a small set of familiar and non-familiar numbers that have been assessed and created by a large set of participants. As there is a small set of familiar numbers, there is no way to precisely replace digits in a number and have it still originate the created familiar set. For instance, the number "888" is in the set of familiar numbers. There are no numbers in the set of familiar numbers that have only one digit different from this numbers. This is true for many of the numbers within the familiar number set. Of the numbers that are just one digit away from another number in the set, for instance "1993" and "1992", replacing a digit to make another number from the set would mean that the second number could no longer be used in the text to be copied. This would mean the participant would type that number twice in the same set, meaning they would have more experience and may be more practiced at typing that number than others. A similar issue arises with the numbers that made up the non-familiar number set.

For this reason, there has had to be some subjective input when choosing digits to replace for the number sets. The guidelines generated in Chapter 3 were used to create new familiar-like numbers. Although this is not a perfect solution, as it involved experimenter bias, it still allows for the replacement span for numbers to be tested in a similar fashion to words.

The predictability of replacement was also an issue during this experiment. One participant, after taking part, noted that in the word condition, it was sometimes possible to predict which words would have a letter switched. Some words have a smaller edit distance to other legal words, whereas some share very few similarities. In a stream of text containing the following words: “piano, dress, horn, courage”, one could predict that the word ‘horn’ could change to a number of words with only one letter difference, whereas the others in the string could not. This issue affects only the familiar word condition and so may have reduced the replacement span for words artificially compared to the other conditions.

4.6.6 Conclusion

This experiment successfully replicated previous findings for the familiar word replacement span. It also confirmed the hypothesis that familiar words produce larger replacement spans than non-familiar (this finding has not been reported previously). This result suggests that the experimental set-up is satisfactory for eliciting differences in cognitive processing: we know there is a difference for words and non-words and this experiment highlighted that difference.

The experiment did not show a difference in replacement spans for familiar and non-familiar numbers however, despite there being a significant difference in the IKI for the two different types of number. This replicated the results found previously in the eye-hand span experiments.

The final experiment within this chapter will aim to replicate the consistent results of IKI measurements, and will investigate whether familiarity of numbers has an effect upon the copy span.

4.7 Study 6: Copy Span

4.7.1 Overview

The copy span measures the amount of material that a typist stores in memory when typing. This is tested by asking a participant to copy text, which will at random intervals disappear from screen. The participant at this point is asked to recall as much text as they reliably can.

The copy span is larger for text that is predictable and easy to read. It becomes smaller the more difficult the copy text becomes. Salthouse & Saults (1987) found that randomly arranged words resulted in smaller copy spans than words arranged in order. This information led Salthouse to conclude in his 1986 paper (Salthouse, 1986), that "the size of the chunks appears to be dependent on the familiarity of meaningfulness of the material".

Hypothesis

It is predicted that the copy span for words will be larger than the copy span for numbers, as words are easier to remember and chunk than numbers. It is predicted that the copy span for non-words will be smaller than the copy span for words, as real words are easier to chunk as they can be recalled from memory. Additionally, familiar numbers will have a greater copy span than non-familiar numbers for the same reason.

4.7.2 Method

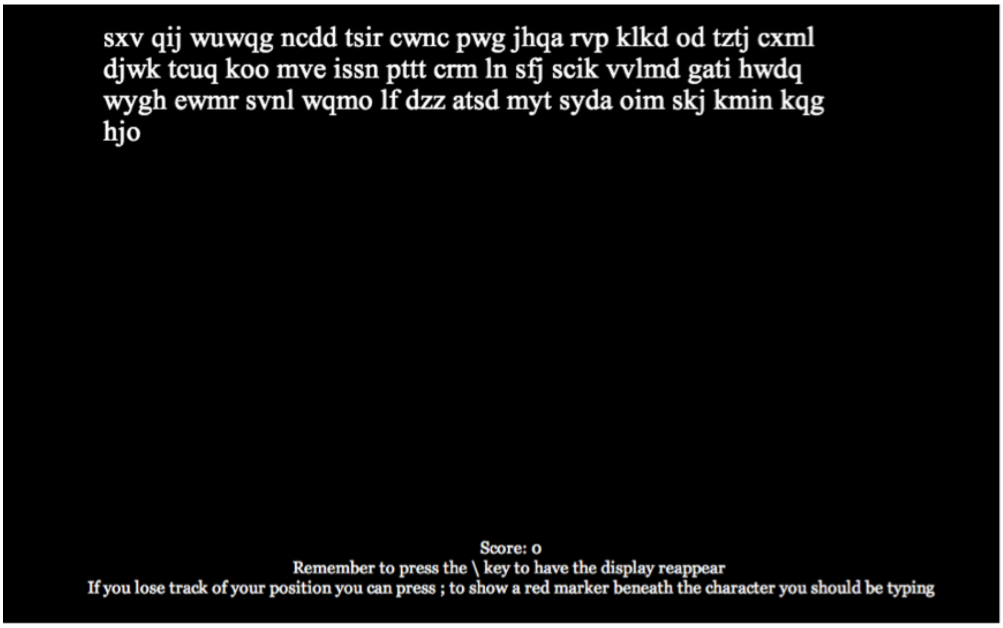
Participants

The experiment took place online. Participants were recruited from the UCL Psychology Subject Pool website. There were 21 participants in total, 13 female, 7 male and 1 no response. The mean age of participants was 25.38 years (*SD* = 8.83).

Design

A 2 x 2 (target-type x familiarity) within-subjects design was used. Targets were words or numbers and each were represented by both familiar and non-familiar text.

The dependent variables in this experiment were copy span, Interkey Interval (typing speed) and errors.



sxv qij wuwqg ncdd tsir cwnc pwg jhqa rvp klkd od tztj cxml
djwk tcuq koo mve issn pttt crm ln sfj scik vvlmd gati hwdq
wygh ewmr svnl wqmo lf dzz atsd myt syda oim skj kmin kqg
hjo

Score: 0
Remember to press the \ key to have the display reappear
If you lose track of your position you can press ; to show a red marker beneath the character you should be typing

Figure 4.13: Screen shot of the copy span experiment

Copy span was measured using the display-blanking technique; at random intervals during the transcription task, the screen went blank and participants were asked to type as much as they could confidently remember. The average number of characters recalled is taken as the participant's copy span. Interkey interval and error rate are measured in the same manner as the eye-hand and replacement span experiments.

Materials

The text transcribed during this experiment was the same as used in the second eye-hand span experiment (Study 4). Briefly, this involved participants typing numbers that were both familiar and non-familiar, and words that were familiar and non-familiar. In this context, non-familiar words are defined as meaning non-legal non-words.

The text in this study was displayed using white 20pt Times New Roman. Participants typed a block of either words or numbers, comprising of 40 familiar or non-familiar targets. Unlike previous studies in which a stream of text was displayed, this experiment presented a block of text to the participants. As this test would involve testing how far ahead the participant could copy text, a stream would not have been appropriate (see Figure 4.13 for a screenshot of the non-word condition). The text that the participants entered was not displayed, this was in order to recreate the conditions used in the original copy span study.

In order to test the participants' copy span, the text display went blank and required typists to recall as much text as they could confidently recall. This occurred five times during each condition. The copy span was recorded as the number of characters that the participant was able to recall when the display was blank. The timing of the display blanking was semi-randomised.

Participants were given instructions to press the '\' key once they had recalled all they could; again this feature was taken from the original copy span experiments. Upon pressing this key, the display returned and the participants could continue transcription.

During piloting it was found that due to lack of feedback of typed text, participants occasionally became lost as to which word or number they should be typing. In this situation participants were instructed to press the ';' key, at which point a red marker would appear underneath the character that they were to type next. This red marker also appeared when the participant consecutively typed three targets incorrectly, this was to provide feedback to participants who had not noticed that they were not typing the correct targets.

Unlike previous experiments, this study relied solely upon a numeric score, as the coloured circle feedback could not easily fit on the screen with the text passage. The scoring mechanic was used to encourage participants to both recall as much text as they could, whilst penalising guessing. The bonus for typing a word or number correctly was tripled if that word or number was recalled during a point when the text was not displayed on screen. The penalty for typing a word incorrectly at this point was also tripled.

Procedure

The experiment was advertised as an online study using the university psychology subject pool. Participants were reimbursed with a £3 Amazon voucher for their time.

As with previous experiments, participants were provided with a link to the experiment web page, where they were given health and safety advice, instructions and a video explaining the experiment.

Participants completed a trial period before the main experiment. This consisted of typing out a passage of 40 words and then 40 numbers, the order depended upon condition. Once the participants had completed the trial, the experiment began. This also involved copying passages of 40 words or numbers in length, which were either familiar or non-familiar. Participants therefore copied four different passages. Once participants had completed one condition, they were given the option to take a short break before continuing with the second condition. Participants were randomly assigned to complete the word or number condition first, the order of familiarity was randomly determined.

At the end of the experiment, participants were debriefed and the aim of the experiment explained. Participants were given access to their data if they wished to download it.

The experiment took on average 15 minutes to complete from beginning to end.

4.7.3 Results

For statistical analysis a 2 x 2 (target-type x familiarity) repeated measures ANOVA was used.

Copy span

Copy span shall be considered first. Copy span is defined as the average number of characters recalled and typed for each condition of the experiment.

This data is shown in Figure 4.14. Participants had a larger copy span for familiar text (M = 5.74, SD = 3.38) than for non-familiar text (M = 3.83, SD = 2.21). This difference was significant, F(1, 20) = 10.21, p = .005, η^2 = 0.34. Participants had a marginally larger copy span for words (M = 5.21, SD = 3.05) than for numbers (M = 4.36, SD = 2.67). However, statistical analysis showed no main effect of target type on copy span, F(1, 20) = 1.99, p = .174, η^2 = 0.09.[14]

The planned simple effect test of familiarity was performed for words and numbers separately. Reinforcing what can be seen in the figure, there was a significant simple effect of familiarity for words, F(1, 20) = 7.74, p = .011, η^2 = 0.28. However, there was no significant effect of familiarity for numbers F(1, 20) = 2.43, p = .135, η^2 = 0.11.

Interkey Interval (typing speed)

Statistical analyses of the data shown in Figure 4.15 show that participants had significantly faster IKI for familiar text (M = 256, SD = 75) than for non-familiar text (M = 350, SD = 96), F(1, 20) = 90.87, p <.001, η^2 = 0.82. Participants also had a faster IKI for words (M = 288, SD = 79) than for numbers (M = 318, SD = 94), F(1, 20) = 6.93, p = .016, η^2 = 0.26.[15]

A planned test was conducted to explore the simple effect of familiarity for both words and numbers. Here there is a significant simple effect of familiarity for both words, F(1, 20) = 99.39, p <.001, η^2 = 0.83, and for numbers, F(1, 20) = 33.34, p <.001, η^2 = 0.63.

Further follow-up tests were conducted to determine the source of the significant interaction. A pair of simple effects tests of target type were performed, finding that participants had a significantly faster IKI for familiar words (M = 219, SD = 73) than for familiar numbers (M

[14] The familiarity by number interaction was also non-significant, F(1, 20) = 3.33, p = .083, η^2 = 0.14.

[15] The familiarity by target type interaction was also significant, F(1, 20) = 53.147, p <.001, η^2 = 0.73.

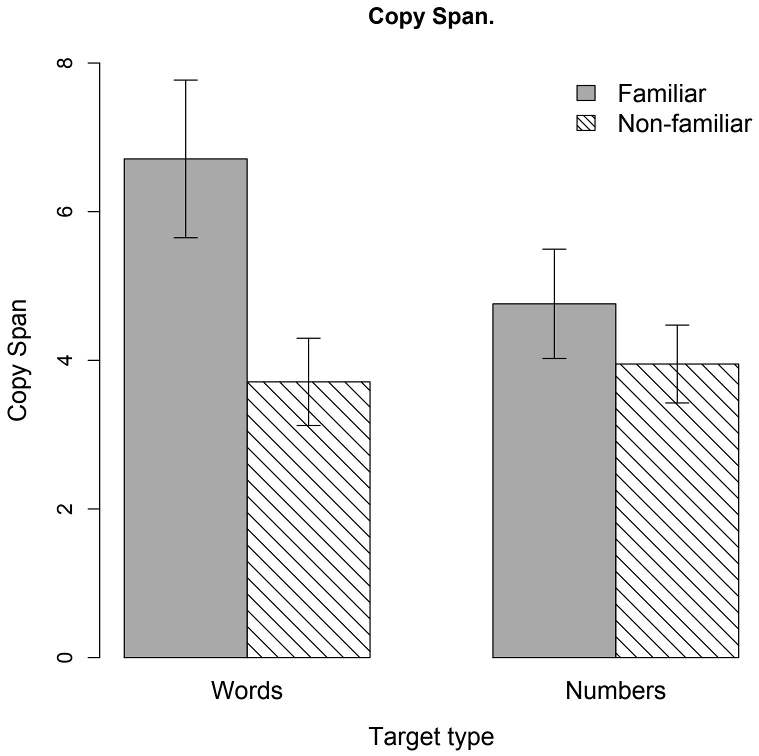


Figure 4.14: Mean copy span for participants in all conditions

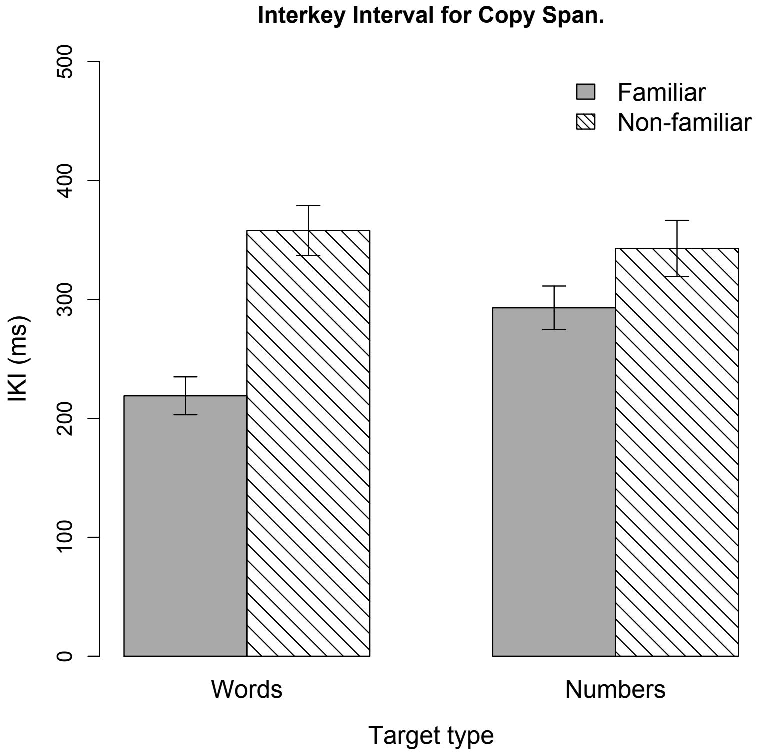


Figure 4.15: Mean IKI for participants in all conditions

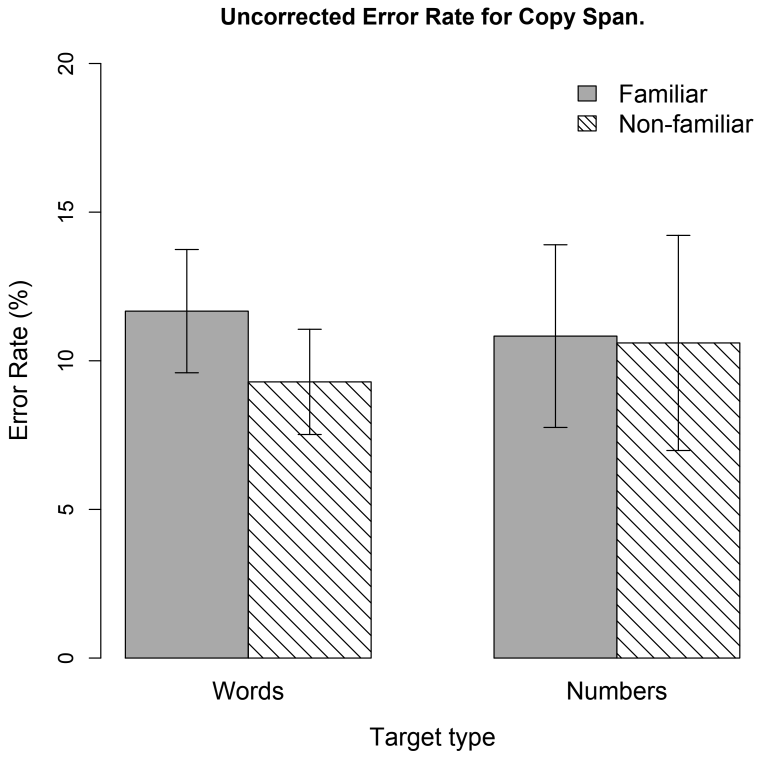


Figure 4.16: Mean uncorrected error rate for participants in all conditions

= 293, *SD* = 85), F(1, 20) = 55.26, p <.001, η^2 = 0.73. However, for non-familiar targets there was no effect of target type (i.e., words vs. numbers) F(1, 20) = .911, p = .351, η^2 = 0.04.

Errors

The uncorrected error rate can be seen in Figure 4.16. Although the rate was higher for familiar targets (M = 11.25, *SD* = 10.0) than non-familiar (M = 9.94, *SD* = 9.85), a statistical analysis of this difference shows it is not significant F(1, 20) = .43, p = .591, η^2 = 0.02.

Participants made marginally fewer errors when typing words (M = 10.48, *SD* = 7.59) compared to numbers (M = 10.71, *SD* = 13.43), but again this is not a significant difference F(1, 20) = .007, p = .933, η^2 = 0. [16]

[16] The interaction effect between familiarity and target type on uncorrected error rate is also non-significant F(1, 20) = .345, p <.564, η^2 = 0.02.

4.7.4 Discussion

Copy Span

Previous research into the copy span suggests that for words, the span ranges from 7 to 40 characters (Phenomenon 13 in Salthouse (1986)). This wide range is a result of differing levels of readability of text and predictability of words used in the experiments. In the current experiment, the copy span for familiar words was 6.71 characters, which is close to the lower end of the range of spans previously found. The words used in this experiment had no predictability, that is, the passages being copied were randomly arranged words, rather than a section of prose; this will have resulted in a reduced copy span. The result is close to previous findings and so it is possible to be confident that this experimental design is sufficient to replicate previous results.

As with previous spans, the copy span for words was larger than that for numbers. This again can be attributed to the efficiency of the chunking system: words were easier to chunk as wholes than numbers were, leading to more information being stored in the cognitive processor, which could then be recalled once the text disappeared.

The familiarity of text had a significant effect upon the copy spans for words: familiar words had a larger copy span. This was predicted at the beginning of the experiment, as the efficient chunking of words as whole units, compared to chunking as syllable or characters, meant that the copy span would be larger. A similar effect was not noted for numbers where, despite non-familiar numbers having a smaller copy span than familiar numbers, the difference was not significant. This finding suggests that the chunking span of the numbers was no different for familiar or non-familiar numbers.

Interkey Interval

As in all previous experiments, familiar targets were significantly faster to type than their non-familiar counterparts. No further discussion is offered here, as this effect has now been replicated multiple times within the experiments in this chapter.

Error Rate

The error rate was not affected by familiarity or target type, a finding seen in many of the experiments within this chapter. However, the standard deviation in error rate was very large, particularly for those in the number condition. This shows the different strategies that some participants took during the experiment: some were careful to have a high accuracy level whereas others chose not to make that as high a priority.

The error rate during times when the display text disappeared was no higher than during the rest of the experiment, suggesting that participants were being cautious when recalling the text. This may have contributed to the copy span appearing small compared to previous research. Error rate during previous experiments is not discussed, so comparisons cannot easily be made.

4.7.5 Limitations

As with the eye-hand span experiments, the copy span has been determined using eye-tracking in previous studies (Rothkopf, 1980). Such studies have shown that copy span is highly correlated with the number of fixations a participant makes when viewing the text. Using an eye-tracking approach would have allowed for the copy span to be constantly measured during the experiment, rather than tested at a finite number of intervals. However, the benefits of running these experiments online outweighed the benefits of testing copy span using an eye-tracking methodology.

Another limitation of this study was the length of passages copied by participants. The passages in this experiment were 40 words or numbers long. In previous research, passages of up to 150 words have been used (Rothkopf, 1980) which would allow for far more tests of the copy span throughout the passage. The limited passage length used in this study was determined by the limited number of familiar numbers generated by work in Chapter 3.

4.7.6 Conclusion

This experiment has provided the final piece of evidence in this chapter that familiar and non-familiar numbers do not have discernibly different typing spans associated with them. For a third time, the differences between the two have proven non-significant.

However, once again the speed of typing has been shown to be affected by the familiarity of text, for both words and numbers. This is a strong result in terms of replication of findings, and strengthens the argument that familiar and non-familiar numbers require some differing process when being read and transcribed.

These ideas will now be further discussed and all results from each of the experiments will be compared.

4.8 Discussion

The experiments reported in the chapter demonstrate the effects of familiarity on number typing. These have been briefly explored after each section. Here, the results are brought together and larger implications are drawn from viewing the findings as a whole.

4.8.1 Measures of span

The results of the span measurements did not support all span-related hypotheses made prior to this experiment. The experiments supported the hypotheses that spans would be larger for words than numbers and that words would have larger spans than non-words. Extrapolation from previous research into the replacement and copy span had suggested that this would be the case, but it was only the work in this thesis that explicitly tested the assumption. Using a chunking theory, these results suggest that words are stored as larger chunks than non-words, as discussed individually in each section.

A similar effect was not found for numbers in any experiment. However, data from the IKIs collected throughout the experiments suggests that familiarity does affect the way that numbers are transcribed, as significant differences were found in the speed of typing familiar and non-familiar numbers. This contradicts the results from the span measurements.

This discrepancy may suggest that testing typing spans for numbers cannot be easily done by using the same experiments that are used for words. This may be due to the size of differences expected between familiar and non-familiar numbers, compared to those expected for words. This is supported by the effect sizes found during the experiments: η^2 was greater for words than numbers.

Words and non-words are entirely different from one another, in that some sets of letter strings are clearly representative of words. In every language there exists a dictionary of legal words. Non-words are equally clear; if they cannot be found in the dictionary, or simply do not adhere to lexical rules, then a string of letters is a non-word. This means that there is a large difference in terms of the representation of words and non-words in the memory; non-words can be given no meaning. Consequently, differences in the spans involved in typing words and non-words are similarly large.

With familiar and non-familiar numbers however, the differences are not as large. All numbers used in this experiment were legal, and could be parsed to form a number word; and each number word had a meaning in that each represented a magnitude. The familiar words, as well as representing a value, also represented other information as discovered in Chapter 3. This meant that the distinction between familiar and non-familiar numbers, in terms of representation in memory was not as large as with words.

The spans, unlike the timing intervals, were measured using discrete values. The eye-hand spans and replacement spans, for instance, tested each participant with a series of possible spans. This meant participants were assigned to a particular span size, rather than the participant's span being directly measured. As the differences in representation for words was so great, this discrete testing of spans was able to highlight the differences. It is possible that, from observing the smaller effect sizes, the span tests are not detailed enough to highlight the subtle difference between familiar and non-familiar numbers in the same way that they can measure differences between words and non-words.

Another potential issue that would affect the effect size of number familiarity on the various typing spans is that it could not be guaranteed that all numbers within the set were familiar to each typist. It may be that only a subset of the numbers in the familiar number set were familiar to each participant, meaning that the set of familiar numbers was actually a set of familiar and non-familiar numbers. This effect was seen in Study 2, where during the card-sorting activity, some participants needed to place the generated familiar numbers in a junk pile, as they did not find them familiar despite being rated as such in Study 1. Similarly, it was possible that some of the non-familiar numbers were actually known to the participants, meaning that set could also represent a mix of both familiar and non-familiar numbers. This issue would result in any differences between the two levels of familiarity being diminished.

Ultimately, the difference between words and non-words is robust enough to withstand the lack of detailed tests using the Salthouse studies. Familiarity of a number is more subtle and, as such, may require more detailed testing. As a result, the Salthouse span measurement experiments may not be appropriate. To measure this subtle difference, continuous measures may be required, such as those used for timing data.

4.8.2 Interkey Interval

Every experiment provided evidence that words were faster to type than numbers (this supports the findings from Chapter 2 into typing job requirements), and that familiar targets were faster to type than non-familiar targets. This effect was found for both words and numbers; for both, the effect size was medium to large (Cohen, 1988). This result replicated many previous studies that found typing non-words to be slower than typing words (Phenomenon 5 Salthouse (1986)). This result also supports the hypothesis made prior to the experiments that non-familiar numbers would be slower to transcribe than familiar numbers.

Condition	Replacement	Eye-hand I	Eye-hand II	Copy	Mean
	Mean (SD)	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*
Familiar Numbers	296 (80)	271 (104)	186 (72)	293 (84)	261 (85)
Non-familiar Numbers	356 (111)	330 (120)	329 (80)	343 (108)	340 (105)
Words	212 (82)	207 (79)	269 (54)	219 (73)	227 (72)
Orthographic Non-words		281 (99)			281 (99)
Non-words	363 (113)		326 (47)	358 (96)	349 (89)

Table 4.3: Table summarising the mean IKI taken from all experiments for each condition (Timings in ms)

A further finding was that during each experiment, non-familiar numbers and non-words were typed at the same speed. Table 4.3 summarises the mean IKI results from all experiments (note here "lexical non-word" defines non-words that adhere to lexical rules and are pronounceable, this result is taken just from the first eye-hand span experiment). Prior to the experiment, it was stated that IKI becomes slower as the material being typed becomes more random (see Phenomenon 5 Salthouse (1986) for a review). The results from these experiments support that finding: familiar words were typed faster than lexical non-words, which were in turn typed faster than true non-words.

Prior to the experiment, a hypothesis was posed with regards to how numerical typing speed would fit into this spectrum of transcription IKIs. The hypothesis has been supported in that familiar numbers are typed slower than familiar words. In addition to that, familiar numbers are typed faster than non-words. The typing speed of non-familiar numbers and non-words is comparable.

Salthouse reports that the slower typing speed is representative of smaller chunks being used. Interpreting the results gathered during these experiments in a similar manner suggests that familiar numbers are read and stored using larger chunks than non-familiar numbers. This can be contextualised with general typing research in the following way: transcribing familiar words allows a typist to use large, efficient chunks. Familiar numbers, although slower, still allow the typist to create large chunks, due to the number's representation in memory. Orthographic non-words utilise smaller chunks: the full word is not represented in memory, yet it can be pronounced, and therefore memorised, at a syllable level. Non-words and non-familiar numbers have no representation in memory other than the smallest units, characters.

One issue with this understanding of the results is in the pronounceability and legality of non-familiar numbers. Although they have no full representation in memory, they are still legal strings and, as such, can be processed into a singular number-word, making them similarly pronounceable to orthographically legal non-words. The results from these experiments suggest that this may not be the case however, as orthographical non-words were still transcribed more quickly than non-familiar numbers (as seen in the first eye-hand span experiment, Study 3). This may be indicative of the method of number reading chosen by participants in this study. The numbers may have been read and parsed into number words, or equally, they may have simply been read as digit strings. This strategy may be preferable as it requires less processing of the perceived information. This strategy would also result in less efficient chunking strategies, thus making non-familiar numbers slower than lexical non-words. To investigate this further, it would be important to understand the reading strategies of the typists. One simple way to access this would be to ask typists to read aloud the numbers they are transcribing. However, this could confound the transcription task by giving the typists an extra action to complete, and potentially give them access to a phonological representation of the numbers in working memory. Further research into number reading strategies is necessary.

The discrepancy between the lack of effect found in span measurements, and significant effect found in IKI for familiar and non-familiar numbers, has so far been justified by the lack of detail available in the span experiments. However, it is important to consider alternative explanations. One possible cause for the increase in speed for familiar numbers is that typists are more well-practiced at typing the digrams within the numbers and can thus move more quickly on the keyboard, which would mean that differences in typing speed were not due to chunking strategies. Although studies in the alphabetic typing domain have shown that non-words are typed more slowly, even when the digram frequency is controlled for (Salthouse, 1984), which suggests that practice effects do not increase typing speed when transcribing non-familiar text. This does, however, provide the design for a future empirical study of familiarity in numerical typing which would investigate possible practice and digram effects on number transcription in order to replicate similar studies in alphabetic text entry (Gentner & Larochelle, 1988).

This result is important for the understanding of how numbers are transcribed. The theory so far has not investigated this difference and studies involving numbers treat them as though familiarity has no effect. However, this finding has shown that an interface tested with familiar numbers will appear to perform better than one tested with non-familiar numbers. This finding would not be accurate. Additionally, the results of the IKIs from each experiment have allowed a theory of number transcription to be incorporated in the current text transcription research.

Condition	Replacement	Eye-hand I	Eye-hand II	Copy	Mean
	Mean (SD)	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*
Familiar Numbers	3.88 (4.09)	3.79 (4.85)	5.63 (6.68)	10.83 (14.08)	6.03 (7.43)
Non-familiar Numbers	5.88 (4.16)	5.15 (5.44)	6.38 (4.62)	10.60 (16.58)	7.00 (7.70)
Words	4.63 (4.08)	3.94 (4.43)	4.25 (4.30)	11.67 (9.50)	6.12 (5.58)
Orthographic Non-words		5.09 (4.68)			5.09 (4.68)
Non-words	6.00 (4.83)		3.63 (4.17)	9.29 (8.11)	6.30 (5.70)

Table 4.4: Table summarising the mean error rate (%) taken from all experiments for each condition

4.8.3 Error Rate

There was some deviation in error rates between experiments: the copy span experiment elicited a larger number of uncorrected errors than the other experiment in this chapter. This result is likely due to the different experimental design during the copy span experiment. Other experiments involved participants typing a stream of text, whereas the copy span was tested by asking participants to copy a passage of text. This experiment also did not give participants any feedback about what they were typing, or where they were within the text. These reasons combined likely contributed to a higher error rate during this experiment. A summary of error rates across all experiments can be seen in Table 4.4.

In the first eye-hand span experiment (Study 3), the error rate was affected by familiarity of target, where errors in non-familiar text were more likely. This difference is difficult to account for, as not only was it not replicated throughout the other experiments, but it was also not replicated when the eye-hand span experiment was repeated. Many of the errors in this experiment were due to participants terminating the word or number before it had completed. This was as a result of small preview sizes; the preview size would often prevent the end of a word or number from being seen. This occasionally caused participants to think that the word or number had ended, and they thus submitted what they had typed prematurely. This effect would be greater for non-familiar text, as the participants would have no internal representation of the target they were typing and thus it would not look unfinished. When typing familiar text however, the participant would be able to judge that the word 'towe' had not finished and would thus wait to see the end letter. Participants however, would not know to wait for the end of the non-word 'pkha', for example, as there was no representation in memory and thus no information as to how the word looked as a whole.

However, this would be the case for both orthographical, and normal non-words, so does not explain the difference in uncorrected error rates. Orthographic non-words (pronounceable) had a significantly higher error rate than words. These non-words were typed more quickly than non-words in other experiments. Increases in speed often result in reduced accuracy.

For words, which were also typed quickly, this loss of accuracy could be counteracted by the participants ability to error check from memory, rather than the text being copied from. Orthographic non-words could not be checked as easily, as they would have no representation in long term memory, thus meaning any errors that occurred due to an increase in speed would not be as easily noticed. This may explain the significant result in Study 3 and non-significant result in Study 4. Additionally, these differences may be affected by the participants' adherence and attention to the instructions and brief.

Overall, the error rate results show that there is no conclusive difference in error rate for words and non-words, neither is there a clear difference in rate between familiar and non-familiar numbers. However, error research is still important, but it appears that approaching the research from a familiarity point of view does not provide applicable information about errors. Errors may be best studied as part of more ecologically realistic transcription tasks, where environmental and design factors can also be assessed as potential causes for error.

4.8.4 Implications for models of number transcription

Deep semantic model of arabic number reading

(Cohen et al., 1994)
The inspiration for these experiments came from research with aphasic patients, showing that it was possible for a person to suffer brain trauma, lose the ability to read and recognise meaningless words, but retain the ability to read meaningful words. This suggested that there were multiple routes for reading number and that a semantic route of reading was possible. The experiments in this chapter have provided support for this theory. Something within the transcription process meant that the familiar numbers, with semantic meaning, were typed faster than non-familiar numbers with only magnitude values. These results have previously been gathered solely from case studies with aphasic patients, the results in this chapter support these findings and show evidence of the multiple route model of number reading within a healthy population.

Salthouse and the TYPIST model of text transcription

(Salthouse, 1986; John, 1996)
Salthouse explains many of the typing phenomena in terms of chunking strategies. This theory is replicated in the TYPIST model, which operates on text in chunks. If the TYPIST model were to transcribe numbers, all numbers would be treated equally and thus would be typed similarly to non-words. The numbers would be broken down into small chunks

representing a few characters at a time and would not be considered as full chunks, regardless of their familiarity. The model relies upon a dictionary look-up of words and syllables, each input chunk is tested with a “get-spelling” step (John, 1993). If the chunk is found not to be a word or a pronounceable syllable at this point, then it is processed at a letter-by-letter level. Numbers will have no spelling and thus would be processed as non-words at a digit-by-digit level.

Results from these experiments have shown similarities between non-words and non-familiar numbers with regards to typing speed and, in this way, the model would accurately simulate the process of typing non-familiar numbers. However, it has been shown that familiar numbers are typed faster than non-words. In order for the TYPIST model to accurately simulate this result, the “get-spelling” step would need to have access to something like a dictionary of numbers, which would contain all the numbers that the typist was familiar with.

Although this appears to be an easy task, with the simple addition of a dictionary of familiar numbers, it is actually a more complex process. To suggest that there is a dictionary of familiar number is a simplification of the situation: it has been discussed previously that familiarity of a number is a subjective and continuous concept rather than a binary decision, as is the case with words and non-words. The model would need to artificially determine a list of familiar numbers and assume some cut-off for familiarity, as was decided in Chapter 3. This would result in many numbers, which had a high familiarity rating, being ignored and treated as non-familiar numbers.

QNM model of number transcription

(Lin & Wu, 2011)

This model was trained using randomly generated numbers, as it was assumed that frequency effects are not as important in number transcription as they are in text transcription. The work in this chapter shows this assumption to be incorrect. However, this model could be adapted to make use of the results of this study: the model could be made to see some numbers as familiar. As discussed in Anderson & Schooler (1991), repeated exposure of ideas in the environment can mean they are more strongly represented and more easily retrieved from long-term memory. This effect may well be the cause of number familiarity. Therefore, the model which learns from the numbers it is shown, could be trained to see certain numbers as familiar by presenting some numbers more frequently than others. In this way, the model is smarter than the TYPIST model and would be able to treat familiarity of number as a continuous variable, rather than a binary choice. The results of this chapter could, therefore, easily improve the accuracy of this model.

4.8.5 Implications for the evaluation of number entry interfaces

The results of the experiments conducted in this chapter raise questions about previous research in the number entry domain. As previously stated, much research has been conducted using randomly generated numbers (Wiseman et al., 2011; Oladimeji et al., 2011; Cauchi et al., 2012; Thimbleby & Cairns, 2010), and from this research, claims are made about how this might apply to the real world. However, the research within this chapter suggests it may not be possible to generalise from random numbers to numbers in an applied domain. The previously held belief that number entry is not affected by familiarity has not been supported in this chapter: familiar numbers were consistently typed more quickly than non-familiar numbers.

If the numbers entered on devices in the real world are truly random, then users would have no more experience of one number than another and thus would see all numbers entered as non-familiar. However, if there are patterns in the numbers entered, and some numbers are used more often than others, then it is possible that some users may become familiarised with certain numbers. This theory is discussed in Anderson & Schooler (1991), where repeated exposure to an item in the environment makes that item more readily retrievable from memory. This theory would mean that regularly used numbers would become familiar. Therefore, this would mean that to realistically evaluate a number entry interface, samples of numbers used in a particular task would need to be used.

In order to fully understand the implications of this chapter upon number entry research as a whole, more information is needed regarding the distribution of numbers used in specific contexts.

4.8.6 Limitations

One major limitation for all experiments conducted in this chapter was the relatively small set of familiar numbers that were available to work with. This limited the amount of text that participants could transcribe, and therefore, limited the number of tests that could be performed in each condition. It would not have been possible to test the copy span any more than three times during each condition without the participants becoming frustrated with how often the text on screen disappeared. The passage length used for the copy span experiment in this chapter was much shorter than in previous tests. Similarly, in the replacement span experiment, only three different spans could be tested. This limited testing meant that the results gathered during these experiments may not have been stable.

This small selection of familiar numbers also limited the lengths of numbers and words used throughout the experiment. Within the eye-hand span experiment, the words were sometimes split due to small preview windows. This is a key factor in Salthouse's theory of slowing. The words and numbers used in these experiments however, were shorter than those used in previous experiments, and thus were not as susceptible to being split, meaning that the eye-hand span may have been artificially lowered.

The length of numbers used in these experiments also means that the results may not generalise to larger numbers. Due the their short lengths, the numbers used in these experiments could generally be memorised in full, without the need for a chunking strategy. This approach would not be possible in all cases of number transcription, for instance when copying a phone number or credit card number. In such cases it is most likely that transcribing longer numbers would require chunking and may therefore mean that the metrics measured in these experiments are not applicable. Further testing with longer numbers is required to establish how number length affects the various number typing metrics.

The experiments, although providing evidence that familiarity affects IKI when typing, have ultimately only shown that this is the case for the particular numbers used within these experiments. This is another issue with the small set of familiar numbers collected here. Future experiments might aim to replicate the studies here, using different sets of familiar numbers, to ensure that the set of numbers used here truly was representative of many different familiar number sets.

4.9 Summary

The aim of these experiments was to investigate a difference in the cognitive processing of familiar and non-familiar numbers. Through typing studies, a clear difference between the two types of number arose when measuring typing speed. It was also hypothesised that there would be notable differences in the typing spans of the familiar and non-familiar numbers; this however, was not shown to be the case. This may be due to the style of experiments and the lack of detailed measurements available using the testing methodology.

The results of this experiment have clear implications for the evaluation of number entry interfaces. Familiarity has a significant effect upon how quickly familiar numbers are transcribed. Current methodologies that rely upon randomly generated numbers are only ecologically valid if the numbers typed in applied domains are random and entirely unfamiliar to the user. The next chapter, therefore, focuses on number entry in the real world and investigates the implications of these findings in specific contexts.

Chapter 5

Number Entry in Hospitals

5.1 Introduction

The previous chapters have shown that when considering how numbers are read and transcribed, it is incorrect to assume that random numbers can be representative of all transcribed numbers; typists treat some familiar numbers differently. This finding has contributed to the theoretical understanding of number transcription tasks and so, at this point, it is important to consider how these findings might apply to the real world. In this chapter, an analysis of numbers used in hospitals, a set of heuristics and two experiments into number entry interfaces in the medical domain are presented that aim to understand more about how number entry in the real world might verify the findings in the previous chapter.

Number entry is a key task in the financial domain. Transcribing numbers incorrectly can result in large monetary losses and companies collapsing (McCurry, 2005). However, number entry is also an essential task in many safety critical domains such as aviation, nuclear power and the military. In this chapter, the area of focus will be on number entry in the medical domain.

The work conducted in this thesis was done as part of the CHI+MED project (EPSRC Grant EP/G059063/1). The CHI+MED project aims to address a serious issue in the medical domain with regards to device design. Modern medicine makes use of technology in a number of ways, from patient record keeping to administering treatments. The use of technology in these situations can allow for more directed treatment, with finer detailed control. However, it is important that these devices still allow medical workers to fully understand the actions they are taking and reflect upon their interactions. A poorly designed device can make the process too abstract and thus prevent the user from understanding issues with the treatment. Such a case occurred in the Beatson Oncology Case (Cook, Nemeth, & Dekker, 2008), where the patient LN was given a series of radiation treatments that were 58% stronger than needed. This serious overdose eventually led to her death. One cause of this overdose was thought to be the machine's control over the process that took away calculation duties from the trained operators. Without control over the precise calculations, the operator was less able to reflect upon the dosages and information produced by the machine, resulting in a far larger dose of radiation than expected.

The issue with this case, and others, is that often the user is blamed; the solution, therefore, is often to provide more training or to sack the user. Unfortunately, it is rare that the device itself is evaluated to see if a better design could prevent such errors from occurring, or to see if fault lies with the interaction itself (Cook et al., 2008). The aim of the CHI+MED project is to reduce unnecessary harm to patients by making medical devices safer.

In the report, *To Err Is Human*, it is estimated that up to 98,000 patients a year die as a result of preventable error in the medical domain in the US alone (Kohn, Corrigan, & Donaldson, 2000). These errors relate to a range of issues, from leaving surgical implements in a patient after surgery to failures of communication. Some of the errors relate specifically to errors during treatment or the administration of drugs.

Number entry in the medical domain is, therefore, an important task. Here, number entry errors can cause serious harm to patients and can even cost lives. Unfortunately, news stories regularly report on number entry errors leading to a patient's death: from being given an overdose of food solution (BBC News, a), to morphine overdoses (BBC News, b), to incorrect dosages on prescriptions (Smith, 2012).

It is clear then, that number entry interfaces and tasks must be designed in order to minimise these risks. Indeed, research is now being conducted into numerical entry tasks in hospitals, looking at how both devices and humans check for errors in these situations (Oladimeji et al., 2011; Thimbleby & Cairns, 2010). The research shows there are possible solutions to a small set of errors that occur during number entry on medical devices, but the problem remains far from solved.

In this chapter, a number of approaches are taken to inform the evaluation and design of number entry interfaces in the medical domain. The four studies in this chapter firstly aim to investigate how applicable the results found in Chapter 4 are to the real world by understanding if familiar numbers occur in the medical domain. These findings are then converted to a simple set of heuristics, which allow for heuristic evaluation of existing medical device designs. Using these heuristics, number entry interfaces are adapted to better suit the tasks that they are used for. Finally, the knowledge of numbers used in the medical domain is applied to a new design approach to catching errors, this experiment looks at how checksum algorithms can be used to prevent errors in medical device programming.

The results in this chapter show how the landscape of numbers used in a particular task can affect both how they are transcribed by the user, and also how number entry interfaces are designed.

5.2 Study 7: Numbers used in infusion pump programming

This study and its results were published in the journal paper Wiseman, Cox, & Brumby (2013b) and were presented at the Human Factors and Ergonomics Society Annual Meeting in 2012.

5.2.1 Introduction

Much of the research in the number entry domain is often conducted under the assumption that the numbers being entered do not affect the way an interface is designed or evaluated. Much of the research in this domain tests scenarios with randomly generated numbers (Thimbleby & Cairns, 2010; Cauchi et al., 2012; Wiseman et al., 2011). This may be an appropriate approach to the research: if the numbers being entered in medical situations truly are random, then medical workers will become no more acquainted with one number than another, meaning that they are essentially transcribing non-familiar numbers on these devices.

However, if there is a possibility that the numbers are not random and do follow a pattern, then the medical workers using these devices may become familiar with certain numbers. Applying the findings of the research in this thesis, it is possible to show that evaluating and designing devices using randomly generated numbers will not be ecologically valid and will not simulate real world conditions where medical workers transcribe familiar numbers. This is important, as results from the previous chapter show that familiarity has a significant effect upon the speed that numbers are transcribed.

This is not a novel concept in the text entry domain. When testing text entry interfaces, participants are given a series of words to enter. The McKenzie word set has become a standard for usability testing on keyboards (MacKenzie & Soukoreff, 2003) as it simulates the words that an average user might use. This ensures that common digrams and words are tested more frequently than digrams that don't commonly occur in everyday usage. Levels of ecological validity continue to increase: recent work on new keyboard designs has suggested that a standard word set is not a realistic enough metric to test keyboards against and has championed the use of composition tasks, which ask participants to write words of their choosing during experiments (Vertanen & Kristensson, 2014).

In order to improve the research conducted in the medical number entry domain, and begin to bring it to a similar standard of ecological validity that is attained in the text entry domain, it is key to understand whether the differences between typing familiar and non-familiar number affect medical number entry.

Previous research in a number of other contexts suggests that numbers do not occur randomly and that digits do not have an equal likelihood of appearing in a particular place within a number. It might, therefore, be possible to look at existing research on number and digit distributions in order to better understand the numbers being transcribed in hospitals.

There has been a small amount of research into the distribution of numbers in certain domains and contexts. To determine whether experiences with numbers are similar throughout different cultures, researchers have investigated the frequency of number words and numerals occurring in set corpuses including government documents and newspapers from many different languages (Dehaene & Mehler, 1992). This research showed that the patterns for each country were very similar. For numbers 1 to 9 and 10 to 90, there was a noticeable decrease in frequency, with 1 and 10 being used most frequently and all numbers occurring less frequently after that. Interestingly, for all numbers from 1 to 1 billion, some numbers appear as spikes in the frequency graph, showing that they occur more often than expected; these numbers are 10, 12, 15, 20, and 100.

Beyond this research, little has been done to understand number distributions. More work has ben done in terms of digit distribution analysis. Although digit analysis will not provide information about the numbers that may be familiar to us, it can provide an insight into certain specific aspects of numbers. One of the most famous examples of digit distribution analysis is Benford's Law, which is noted in many naturally occurring data sets (Benford, 1938). For example, stock market prices (Ley, 1996), fraud detection (Durtschi et al., 2004) and country populations (Sandron, 2002). This law states, that in naturally occurring number sets, the probability that the leading significant digit is 1 is close to 30%, this is greater than the expectation of around 11% if one assumes all digits bar zero have an equal possibility of being the leading digit in a number. This trend continues with 2 being more likely than 3, which is in turn more likely than 4, and so on. We might, therefore, expect to see a Benford distribution in the numbers used when programming infusion pumps. This finding may not provide a complete understanding of the number distribution in the medical domain, but similarities between this work and the number distribution research previously presented can be seen; the highly frequent numbers found by Dehaene & Mehler adhere to Benford's Law, with leading digits of 1 and 2.

Benford's Law has an additional requirement that may mean it is not applicable in the case of medically important numbers. The numbers have to be cumulative and "naturally occurring", that is, they cannot have been created by a human. The numbers programmed into medical devices are likely to have been determined by humans when deciding what dosage to administer to patients and thus the numbers will not fulfil the naturally occurring requirement.

There has also been some effort to look at the distribution of terminal digit preference in a medical context (Nietert et al., 2006; Wen et al., 1993). In the study by Wen et al. (1993), it was shown that human preference has an effect upon which digit a blood pressure reading is likely to end in. Despite any number being possible when recording these readings, it was found that 78% of the time, the reading will end in 0; 5% of the time, it will end in 5; 15% of the time, in an even digit; and 2% of the time, it will end in an odd number other than 5. This finding shows the doctors' natural bias when choosing how to record blood pressure readings and suggests that this bias possibly has a large influence on these findings. This shows how numbers in the medical domain can be affected by human influence and why it might be that Benford's Law would not apply in these cases. However, this research considers only the last digit of numbers and says nothing about the distribution of all digits entered in such tasks. Therefore, it may not be appropriate when considering how best to represent the numbers entered in the medical domain.

This research combined, suggests that we may be able to predict medical numbers. The numbers found in Dehaene & Mehler (1992) adhere to a Benford's Law-like distribution, with smaller numbers occurring more frequently than larger numbers. This finding may mean we can expect the same sort of patterns with numbers used for programming medical devices. The findings in the Dehaene & Mehler paper also support the findings of the terminal digit preference work; the numbers that showed up as spikes in the distribution, and were unexpectedly frequent, terminated in 0, 5, or 2, which is similar to studies of terminal digit choice. Again, this finding may suggest that a similar pattern will be seen in the terminal digits of infusion pump programming data.

However, overall, it is very difficult to determine whether such research might predict the distributions found in numbers collected from hospital logs. Firstly, two of the reported distributions here predict only a single digit within the number. Secondly, it is difficult to say whether numbers used to program pumps represent "natural" numbers. The numbers used in hospitals may be based on the size of predetermined container sizes, such as drug bags or syringes. They may also be based on standard, predetermined rates and volumes, which will not allow for natural variation in the numbers used.

Due to this uncertainty, it is important to investigate the digit distribution and number frequency in a particular set of number input data. It is clear from the research that there

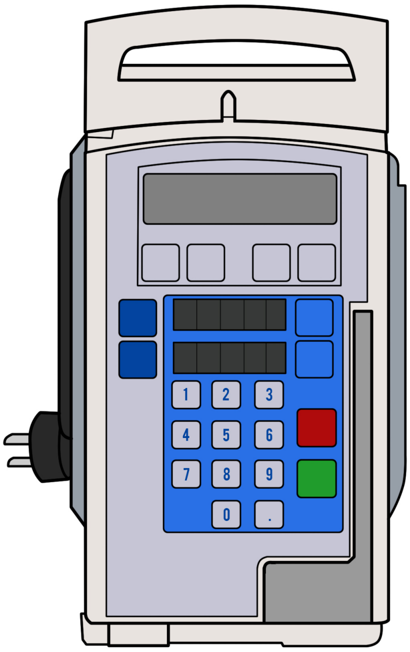


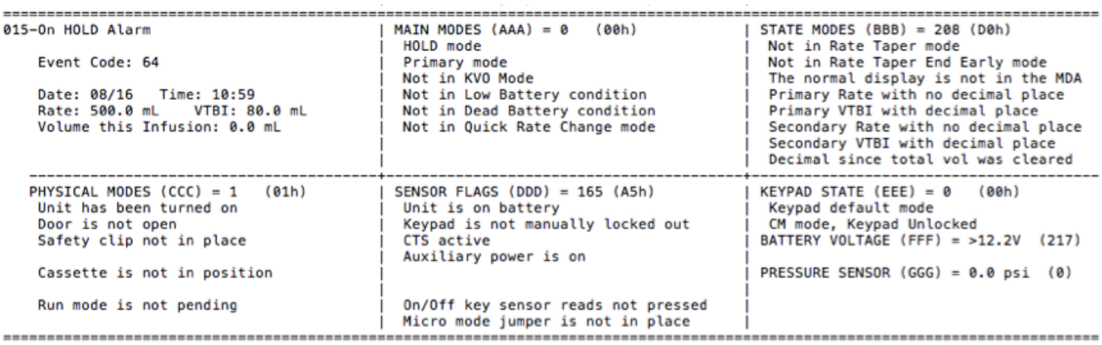
Figure 5.1: Illustrative example of an infusion pump. The interface allows for entry of dosage information, and control of the infusion. A display screen shows a small amount of information about the current infusion.

may be distributions we could use to mimic real life data, but at this point there is no evidence to suggest that this would be a viable route to understanding medical numbers. In this study, a set of numbers is taken from the medical domain, which will be analysed in order to investigate the possibility of patterns of number usage.

5.2.2 Method

The infusion pump is highlighted as a device that would require medical workers to transcribe and enter numbers and so is chosen as the subject of this investigation. An example illustration of an infusion pump can be seen in Figure 5.1. The infusion pump is used to administer fluids to a patient, this can be medication or nutrients. The pump controls the levels of fluid entering the patient's intravenous system. An infusion pump is often programmed by a medical worker, who will enter in the numbers that represent a subset of information about the infusion: the time the infusion will run for, how much fluid needs to be infused and the rate at which the fluid should flow.

Additionally, the device is interesting to explore as there are multiple different infusion pump interfaces available, ranging from the standard calculator or telephone layout keypad



```
============================================================================================================================
015-On HOLD Alarm                          | MAIN MODES (AAA) = 0   (00h)            | STATE MODES (BBB) = 208 (D0h)
                                           |  HOLD mode                              |  Not in Rate Taper mode
    Event Code: 64                         |  Primary mode                           |  Not in Rate Taper End Early mode
                                           |  Not in KVO Mode                        |  The normal display is not in the MDA
    Date: 08/16   Time: 10:59              |  Not in Low Battery condition           |  Primary Rate with no decimal place
    Rate: 500.0 mL    VTBI: 80.0 mL        |  Not in Dead Battery condition          |  Primary VTBI with decimal place
    Volume this Infusion: 0.0 mL           |  Not in Quick Rate Change mode          |  Secondary Rate with no decimal place
                                           |                                         |  Secondary VTBI with decimal place
                                           |                                         |  Decimal since total vol was cleared
   ----------------------------------------+-----------------------------------------+---------------------------------------
   PHYSICAL MODES (CCC) = 1   (01h)        | SENSOR FLAGS (DDD) = 165 (A5h)          | KEYPAD STATE (EEE) = 0   (00h)
    Unit has been turned on                |  Unit is on battery                     |  Keypad default mode
    Door is not open                       |  Keypad is not manually locked out      |  CM mode, Keypad Unlocked
    Safety clip not in place               |  CTS active                             | BATTERY VOLTAGE (FFF) = >12.2V  (217)
                                           |  Auxiliary power is on                  |
    Cassette is not in position            |                                         | PRESSURE SENSOR (GGG) = 0.0 psi  (0)
                                           |                                         |
    Run mode is not pending                |  On/Off key sensor reads not pressed    |
                                           |  Micro mode jumper is not in place      |
============================================================================================================================
```

Figure 5.2: Example entry from Infusion Pump Logs

to directional keys, a review of the range of number entry interfaces can be found in Oladimeji et al. (2013).

From a practical stand point, these devices are chosen as they record log data about the dosages programmed into them. This is important as it provides a clear set of data containing information about the numbers commonly programmed into the devices. Due to strict ethical rules, access to medical records where dosage information was stored is not possible. The infusion pumps retain this information and they do not associate it with any patient data, thus making it the least intrusive method for accessing sets of numbers used in hospitals.

Data Source

A total of 154 log files were collected from 104 pumps over 4 different years. All the pumps had a numeric keypad that was used to enter numbers. On average, each log covered a period of 15 days, with the longest log lasting for 173 days and the shortest covering a single day. The pumps used in this study were located in different wards throughout one hospital, including oncology, surgery, paediatrics, midwifery and general medical.

The log files contained a large amount of information, recorded at an event level. An event could be anything from starting and stopping the machine to an alarm sounding if an infusion had finished. At each stage, information about the rate of infusion, the volume of drug to be administered or "volume to be infused" (VTBI) and volume already infused, was logged. Figure 5.2 depicts an example of a log entry, as displayed in a text editing program. This log has been created due to an alarm sounding. The VTBI and Rate information can be seen at the left-hand side of the log.

Data Extraction

For this study, the information required for analysis included all numbers entered by a medical worker in order to program the pump for a particular infusion. The logs from the pump did not record data at this level and so this had to be extrapolated from the data. The pump being used required the Rate and VTBI to be entered to start an infusion, so it could be safely assumed that both the Rate and VTBI had been keyed in by the medical worker.

To determine the numbers entered on the pump, a python script was written to iterate over the log files. This script firstly looked for all instances of the pump being started. Then, from this set, the script collected all logs where the "Volume this Infusion" information was at 0.0mL (as in Figure 5.2). These two pieces of information combined (the information that the pump had just started and had until this point not infused anything) ensured that the log would represent a newly programmed infusion. Information about the volume already infused was important, as in some cases of the pump being started, it represented the pump being "unpaused" from a previous infusion. If the Rate and VTBI information was taken at this point it would firstly mean that the Rate information was being recorded for a second time and that secondly, the VTBI would no longer represent a number entered by a medical worker. As the infusions progress, the VTBI that was set by the user decreases; taking the VTBI information at any other point, apart from the very beginning of an infusion, would therefore not represent the number entered by the user.

The Rate is entered into the pump as the amount per hour to be infused. For example, a Rate of 550 mL/hr would be entered simply as the number 550. Equally the VTBI was entered as digits, with no mL unit entered by the user.

The log reported data to two decimal places as standard. Trailing zeros after the decimal place were therefore ignored, as they did not represent digits that a medical worker would have explicitly entered into the pump. The decimal point was also ignored if zeros followed it.

5.2.3 Results

Log extraction

A total of 6,040 numbers were gathered from the logs, 168 of which included decimal places, resulting in 17,122 keypresses being collected. In terms of digit distribution, 16,954 digits were collected in total. The spread throughout the years can be seen in Table 5.1.

Number Distribution used for programming infusions

The aim of this analysis is to investigate if some numbers in the hospital occur more frequently, thus potentially making them more familiar to the medical workers. The analysis

Measure	2007	2009	2010	2011	All
Number of digits	6877	3946	5268	863	16954
Number of keypresses	6964	3989	5301	868	17122
Number of full numbers	2442	1382	1924	292	6040
Number of logs	58	44	39	13	154

Table 5.1: Number of digits, keypresses, full numbers and logs collected for each year

here focuses on the full numbers that were gathered from the logs to determine whether there are patterns to these numbers and whether some are more common than others.

Measure	Rate	VTBI
Number of different numbers	170	153
Total amount of numbers	3020	3020
Maximum value	999	9999
Minimum value	1	0.1

Table 5.2: Variation in numbers used across all years and all wards

Table 5.2 provides data about the numbers being used for Rate and VTBI across all years and all wards. It can be seen that there is a larger range of numbers used for Rate than there is for VTBI (170 numbers used for Rate and 153 numbers for VTBI). The values for VTBI vary more greatly than those for rate: The variations in VTBI range between 0.1 and 9,999 mL, whereas Rate varies between 1 and 999 mL/hr. This variation may be as a result of the physical limits of the pump in terms of ability to vary rate. It may also be as a result of a pre-programmed minimum and maximum value having been set by a technician.

The 10 most common values for Rate and VTBI are listed in the Tables 5.3 and 5.4. These top 10 numbers represent 64.11% of all numbers used for Rate and 72.12% of all numbers used for VTBI.

For both Rate and VTBI, the most common values are predominantly round numbers, ending in 0s; in the case of VTBI, every one of the 10 most frequent values being used ends

Rate number	Frequency (%)
200	16.99
125	7.98
999	6.99
600	6.19
100	6.19
500	5.86
250	4.17
400	3.31
300	2.58
83	2.32

Table 5.3: Highest frequency numbers used to program Rate across all years and all wards

VTBI number	Frequency (%)
100	29.87
1000	9.97
50	9.07
500	6.36
20	3.81
250	3.54
10	2.98
30	2.62
200	2.12
80	1.79

Table 5.4: Highest frequency numbers used to program VTBI across all years and all wards

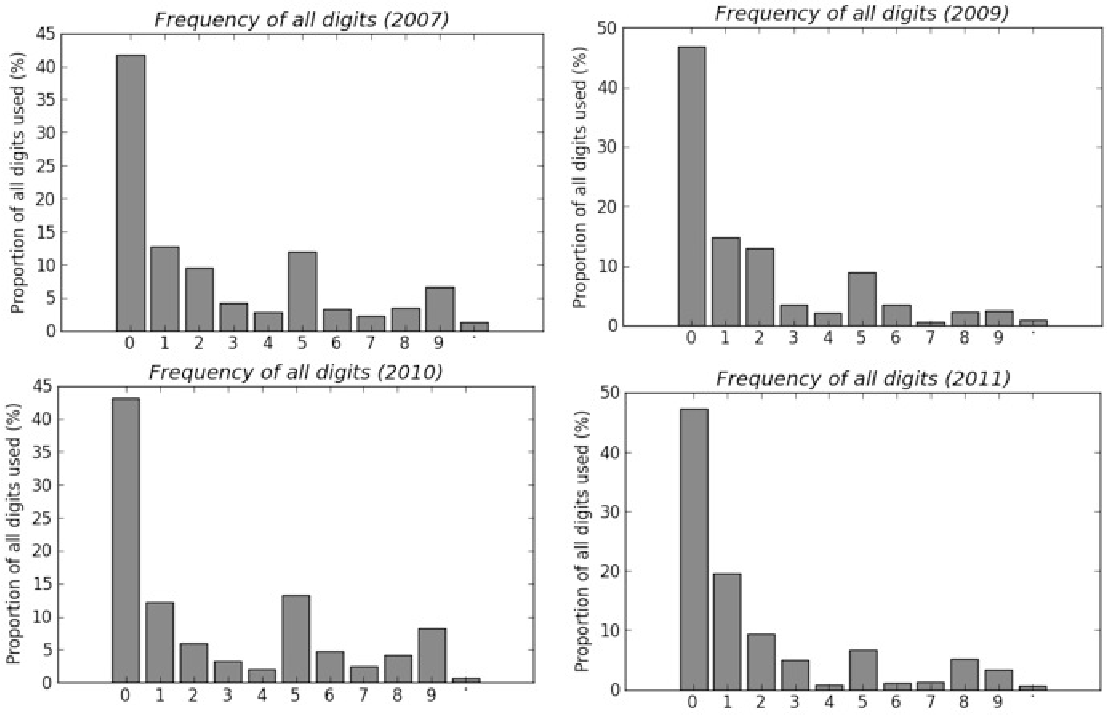


Figure 5.3: Year-by-year analysis of all digits used to program infusion pumps

in a 0. The most common number used for VTBI is 100, which makes up almost a third of all VTBIs programmed into the infusion pumps.

Digit Distribution in the hospital

The next analysis involves understanding more about the digit distributions used in the hospital. These digit distributions represent the keypresses on the devices. The year-by-year analysis of the digit distributions shows that there is a clear pattern of digits that remains stable across each of the 4 years (see figure 5.3). In all cases, 0 is by far the most common digit, accounting for slightly more than 40% of all keypresses on these pumps, with the digits 1, 2, and 5 being the next most common digits.

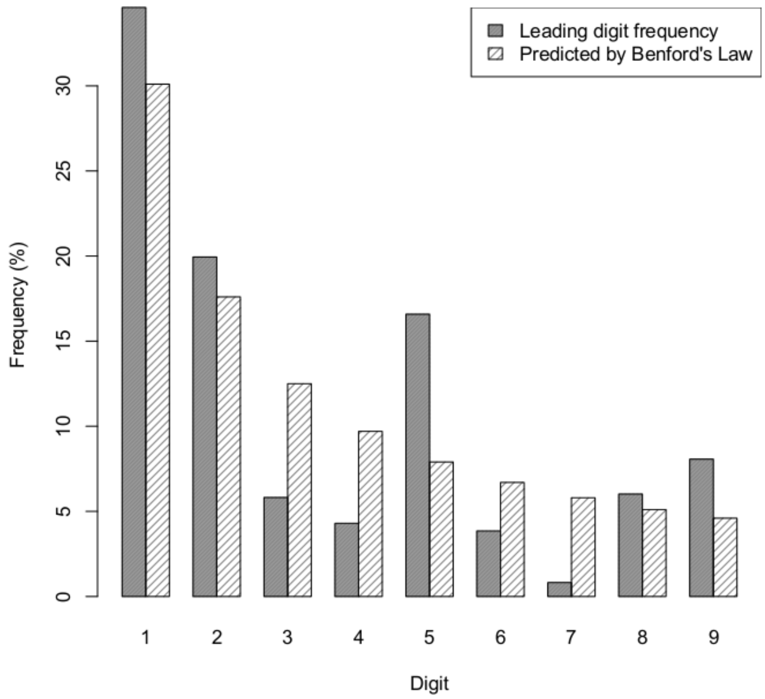


Figure 5.4: Analysis of leading digits used to program Infusion Pumps compared to Benford's Law Prediction

Although Figure 5.3 shows the distribution of all digits, it is important to consider the distribution of just the leading digit to determine whether Benford's Law is applicable to this data set. Figure 5.4 shows the distribution of first digits in all the data. The results show that digit 1 is the most frequent leading digit for the numbers gathered in this study, then the digit 2, but the distribution then diverges from a Benford's Law distribution with infrequent 3s and 4s and overly frequent 5s. This finding suggests that the numbers used when programming infusion pumps may not be naturally occurring numbers, as they do not adhere to Benford's Law.

It is also possible to compare the terminal digits of numbers in this data set with data found previously. Table 5.5 shows the proportion of numbers that terminate with each digit.

The terminal digit across all years is 0 in approximately 72% of cases, which is similar to the 78% found by Wen et al. (1993). In both instances, the digit 5 is the next most common digit for a number to end with. However, again, the distribution seen in this study diverges from previously seen patterns: 9 and 3 appear relatively frequently as terminal digits in this study, whereas in Wen et al.'s study, these two numbers are infrequent.

Terminating Digit	Frequency (%)
0	71.95
1	0.73
2	1.31
3	3.38
4	1.66
5	9.90
6	2.07
7	2.35
8	1.41
9	5.25

Table 5.5: Terminal digit analysis of numbers programmed into Infusion Pumps

5.2.4 Discussion

Familiar Numbers

The aim of this work was to determine whether it was possible for medical workers to perceive some numbers as familiar. The number distributions gathered here show that there are very clearly some numbers that will be more familiar to medical workers than others. The fact that the 50% of all VTBI entries used 100, 1000 or 50 is evidence for this.

There are noticeable similarities between this result and the work of Dehaene & Mehler (1992), who documented the distributions of number words occurring in large text corpuses. For instance, Dehaene & Mehler found there were unexpected “spikes” in frequency for the numbers 10, 20, and 100; these numbers appear in the top 10 numbers used for Rate and VTBI. However, Dehaene & Mehler’s number analysis does not account for other common numbers used in infusions, such as the most common Rate values, 200, 125, and 999. Again, it seems that there is no existing analysis of number or digit distribution that accurately predicts the digits and numbers used in infusion pump programming.

The work reported in Study 1 created a list of familiar numbers that were familiar to a general population. However, the chapter did discuss the possibility of different populations finding different numbers familiar. In the chapter this was considered from a nationality perspective, but the population could also be a group of people with specific training and experience. It is entirely possible that the numbers that medical workers see repeatedly in their daily working life could be considered familiar to them. This was seen in Study 2, where some participants saw numbers comprising of 1s and 0s as binary numbers due to their experience with maths and computers, whereas other participants without that knowledge did not. Knowledge of a domain, and experience with it, can cause a number to become specifically familiar to a person (Anderson & Schooler, 1991).

This study therefore suggests that medical workers are likely to be familiar with a small set of numbers that they use on a regular basis. This suggests that the findings from the previous chapter can have a real impact on the way that medical devices are designed and evaluated. Testing medical number entry interfaces with randomly generated numbers means asking people to transcribe non-familiar numbers. This may be similar to testing text entry systems with random character strings: the results from such an evaluation would not be ecologically valid. Not only this, but the ecological validity has implications for behaviour during a transcription task.

Digit Distribution

This work, in combination with the experiments performed in the previous chapter, has provided evidence that using realistic numbers when testing medical number entry devices is important. Obtaining such numbers to use in evaluation however, can be difficult due to privacy restrictions in place in the medical domain. Here, realistic number generation is considered by building numbers from digit components.

This research shows that there are digits that are used more frequently than others in the hospital. The digit 0 was used three times more often than any other digit, whereas the digit 4 for instance was used infrequently. Users are more likely to use and be familiar with numbers ending in multiple 0s.

Benford's Law is not an acceptable method of generating ecologically valid medical numbers. Although the digits 1 and 2 had a high probability of being the first digit in a number, the digits 5 and 9 were the third and fourth most likely. This pattern diverged from the Benford's Law predictions. Previous work investigating terminal digits appears to predict some of the numbers found in a hospital setting (Wen et al., 1993), but is not entirely accurate.

Other researchers have looked at the distribution of digits used in the medical domain during infusion pump programming (Cauchi et al., 2013). The results of the research shared some similarities to those found here, in that the digit 0 was most often used, and the 1 and 5 were common digits. However, the research diverged from the results presented here in that decimal points were frequently used. This may be due to the different locations of the hospitals and implies that the distribution found in the work in this chapter may not be fully generalisable to all infusion programming tasks. Interestingly, the interfaces on the infusion pumps in both of these studies were different: one used a number pad and the other a 5-key interface. It may be the case that the interface affects the numbers entered or that the interface is chosen depending on the types of numbers that will be entered on it.

5.2.5 Limitations

The number distributions found in this research specifically represent numbers entered in infusion programs and, what is more, were only represented for infusion programs used in this single hospital. In one replication of the study, differences in digits distributions show that the results here may not be instantly generalisable to all hospitals or to other contexts. If other domains are considered, we can intuitively see that this distribution would not likely be seen. For instance, if numbers were collected from the retail domain, we may likely see an abundance of the digit 9, as many prices end in double 9.

However, the aim of this work was not to create an overall distribution that would be applicable in many domains; the aim was to investigate the possibility that workers in a medical domain could have a specific set of familiar numbers and this work has shown that this is the case.

5.2.6 Conclusion

The information gathered in this research is not just important for making more ecologically valid testing conditions for number entry interface research. The results may also have implications for the design of number entry interfaces.

For instance, information about the abundance of the digit 9, and the frequency of the number 999 for the rate, suggests that entering the maximum possible value is very important. This finding should be considered when designing number entry interfaces for infusion devices. It is important that the user is able to input this maximum Rate quickly. Additionally, the top three most common values for VTBI were used in almost half of all infusions (100, 1,000, and 50). It might, therefore, be important that interfaces allow users to access these values as easily as possible. It might also be possible that shortcuts to these frequent values be considered on interfaces. Study 8 aims to summarise these findings into a succinct list of heuristics to be used when considering the design of a new interface. These heuristics will then be tested in Study 9, which aims to adapt existing infusion pump interfaces to better match the numbers that are programmed into the devices.

5.3 Study 8: Generation of Design Heuristics

5.3.1 Introduction

The results of Study 7 highlight some key functionality that number entry interfaces for infusion pumps need to offer users. Currently, understanding which of the results from that study are applicable to design is not clear. Here, the results are distilled into a set of design heuristics that aim to make the more important findings easier to apply to design evaluation. These heuristics are then tested against existing infusion pump interfaces to see how current number entry systems perform. The heuristics are then used to generate a possible future number entry interface design that is best suited to the task of programming infusion pumps.

5.3.2 Materials

Infusion pumps currently utilise a variety of interfaces for the entry of numbers (see Oladimeji et al. (2013)). In the previous study, the logs were taken from an infusion pump with a serial number entry keypad, but other infusion pumps use different interfaces, such as incremental chevrons or a five-key cursor-controlled input. These three designs are discussed in more detail here.

Number pad interface

This interface comprises 12 keys: 10 keys for the numbers 0 to 9, 1 key for a decimal place and 1 key for clearing entered numbers from the display (Figure 5.5). This interface is used to enter the numbers serially, one digit at a time. The interface can be laid out in one of two ways: as a calculator with the numbers 7, 8, and 9 in the top row or as a telephone with the numbers 1, 2, and 3 in the top row. Current guidelines suggest that number keypads on infusion pumps should make use of the telephone layout (NPSA, 2010).

Chevron interface

This interface allows the user to increment or decrement a number displayed on a screen by predetermined amounts (Figure 5.6). The double chevrons make larger increments and decrements, for instance, by adding or subtracting a whole number (e.g., from 1.0 to 2.0). The single chevron makes smaller increments and decrements, for instance, by adding or subtracting a decimal point (e.g., from 1.0 to 1.1).

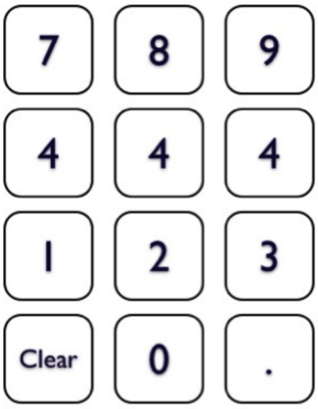


Figure 5.5: A Number pad interface

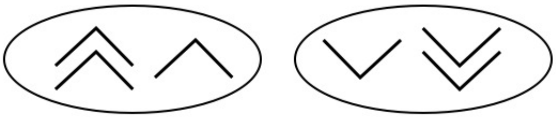


Figure 5.6: A Chevron interface

Five-key interface

This interface allows users to move an on-screen cursor in the left-right and up-down directions (Figure 5.7). The user edits one digit at a time, selecting it by moving the on-screen cursor left or right along the number. At the desired digit, the user presses up or down to change the value of that digit by one integer. This interface can be configured in different ways to allow a wrap-around from 9 to 0 if required.

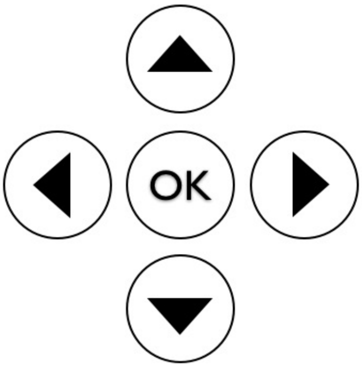


Figure 5.7: A 5-key interface

5.3.3 Creating the design heuristics

Using the results of Study 7, heuristics were developed to help test existing interfaces. The heuristics were based upon the patterns seen in the first study and were aimed at making the most common numbers easier to enter. It is important to note here that this largely addresses speed of entry, and does not take into account the likelihood of errors nor does it consider the scale of errors (as noted in Thimbleby & Cairns (2010)). These heuristics aim solely to translate the findings of the Study 7, heuristics aiming to cover a broader range of issues may differ from those generated here. Here the four guidelines are listed and the reasoning behind each explained.

Number entry heuristics

Heuristic 1: Entering a string of zeros should involve no more than one keypress per digit.

Reasoning: Across all years, the most commonly used digit was 0, making up over 40% of all keypresses. Much in the same way that the @ symbol is brought to the front of a mobile touchscreen keyboard to prevent extra keypresses when entering email addresses; this heuristic ensures that one of the most important digits is easily accessible.

Heuristic 2: Entering a decimal point should require the user to request that functionality to prevent accidental slips.

Reasoning: The decimal point was used very infrequently in the data collected in Study 7. The decimal point button is a powerful key in that it can change the magnitude of a number by a significant amount, often almost imperceptibly. The decimal point is small enough that it might be missed when reading a number on the screen of a device, meaning its accidental use could be dangerous. As it is so infrequently used, it can be made harder to insert. Much in the same way that emergency buttons are often covered to prevent accidental use, and require extra steps to push, the decimal point button should be harder to reach.

Heuristic 3: Entering the value 999 should be possible with, at most, three keypresses.

Reasoning: Unlike many of the numbers found in the study, 999 was a non-round yet common number. This represents the fastest Rate possible to program into the infusion pump. Entering multiple 9s would allow the maximum Rate to be set on the pump; which is a relatively common task (for instance, when treating local anesthetic toxicity (Holmes, Jefferson, & Ball, 2009)). From this information, it is thought that 999 could be used in emergency situations, where it is important that the patient receive the drug treatment as quickly as possible. For this reason the 999 setting should be easily accessible. To allow for serial entry, this heuristic allows for one keypress per digit, leading to a conclusion that the number should be able to be entered with three keypresses or fewer.

Heuristic 4: The numbers 1,000, 100, and 50 should be accessible in one keypress.

Reasoning: These numbers were seen to be used in over half of all infusions. For this reason, as with the digit 0, they should be easily accessible to the user, considering how that it is likely that the user will want to enter one of these three numbers.

5.3.4 Method

A heuristic evaluation was carried out by a single expert (Nielsen & Molich, 1990). Each of the three interfaces that would be evaluated were viewed separately and tested against each heuristic. The result was either a pass or fail for each heuristic.

Number pad

Heuristic 1: PASS The number pad allows for quick entry of multiple zeros. Each zero requires a single keypress.

Heuristic 2: FAIL There is a dedicated decimal point button on this interface, meaning that accidental entry of an unwanted decimal point is very possible (especially with poor number pad design; (Wiseman et al., 2011)). The decimal point button could be controlled by asking users to confirm that they wish to use the decimal key.

Heuristic 3: PASS The number 999 can easily be entered with only three keypresses.

Heuristic 4: FAIL There are no dedicated buttons for these common values on this interface. The user must type the whole number each time.

Chevron

Heuristic 1: FAIL It is not possible to enter strings of zeros on this interface with a single keypress per zero. The user must cycle through values before getting the number he or she wishes to enter.

Heuristic 2: FAIL There is no dedicated decimal point button on this interface, which means a decimal is less likely to be entered by accident. However, it is still possible if the user pressed the down chevron multiple times. This error can be guarded against by asking users to confirm that they want to enter a decimal place using an additional step.

Heuristic 3: FAIL Entering the maximum value 999 takes far more than three button presses. To fix this, the interface could use a maximum-rate button. This interface should require users to confirm that they wish to press this button to guard against accidental use.

Heuristic 4: FAIL There are no dedicated buttons for these values. Shortcut buttons should be provided, or the increments could be tailored so that each time the double chevron button is pressed, the display could show the next most common number, rather than simply adding 10 to what was displayed on screen. This function, however, may not be as predictable to the user, and so the increment values could be made larger to help users navigate to the large, common values more quickly.

Five-key

Heuristic 1: PASS All digits default to zero so the only button press necessary is to move the cursor left or right, away from that zero. This interface is well designed for multiple zeros.

Heuristic 2: FAIL There is no dedicated decimal point button, so entering a decimal by accident is less likely. However, it is still possible to enter numbers after a decimal point by pressing the right key too often; therefore, users should be prompted to confirm that they wish to enter a decimal.

Heuristic 3: FAIL Entering 999 can take 5 keypresses or 29, depending on how this pump is configured. Pressing the down arrow while the cursor is on a 0 can either cause that digit to change to a 9 (if wrap-around is enabled) or remain as a 0 (if wrap-around is not enabled). Allowing wrap-around means that the user can type 999 much faster than without. This feature could be implemented without having to make hardware changes. However, this interface may still require the addition of a maximum button, as the 5 keypresses it currently takes to enter 999 still does not meet the guideline.

Heuristic 4: FAIL There are no dedicated buttons for these values. Shortcut buttons should be provided.

5.3.5 Alternate Interface

Although none of the interfaces evaluated is ideally suited to entering the most common numbers used in infusion tasks, the number pad interface adheres to more of the guidelines than the other two interfaces and is the best suited of this set. However, it is also possible that alternative interfaces may better suit this task. As touchscreen technology becomes more affordable, it might allow more novel number entry interfaces to be used in the design of infusion devices. Entering numbers on a touchscreen scroll wheel (as is common on many smart-phones and tablets) has many benefits (Figure 5.8). This design would allow the user to control each digit independently by either scrolling on the wheel or tapping on the desired number. The decimal portion of the number would be greyed out and would not be usable until the user tapped and confirmed that they wished to select it. The default for the display would have every wheel set to 0. By assessing this design against the heuristics, we can see it performs very well.

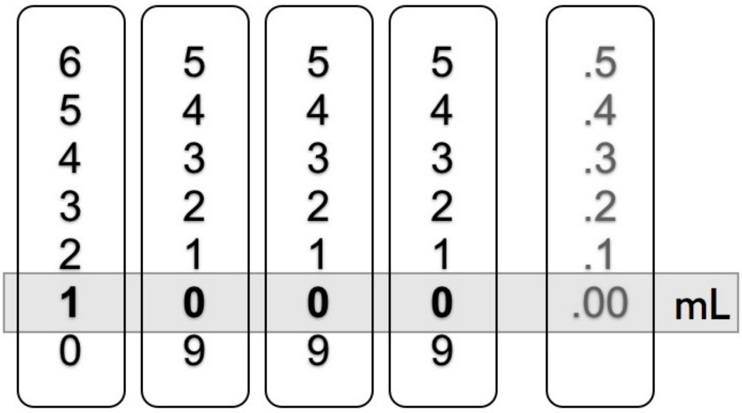


Figure 5.8: Potential scroll wheel interface for entering numbers on an infusion pump

Touchpad scroll wheel

Heuristic 1: PASS All scroll wheels default to zero, meaning no interaction is needed to enter multiple zeros.

Heuristic 2: PASS Entering a decimal point requires that the user interacts specifically with the decimal wheel. The user would have to confirm this intention before the wheel turned, making it harder to enter a decimal by accident.

Heuristic 3: PASS Entering 999 takes three presses; the 9s are on screen so simply need to be tapped by the user.

Heuristic 4: PASS The most common VTBI numbers can be entered by tapping one section on the desired wheel. For instance, 1,000 is entered by tapping the 1 on the first wheel, 500 by tapping the 5 on the second wheel and 50 by tapping the 5 on the third wheel. This essentially makes the most common numbers available with only one keypress, without needing to add dedicated buttons to the interface.

5.3.6 Discussion

By adhering to the heuristics based upon real numbers entered into the devices, one can design an interface that requires fewer presses than the previously analysed and existing interfaces, This, in theory, could reduce the potential for error as there would be fewer opportunities for keying errors, although this does not take into account other causes of error. However, this is currently a theoretical design only, and would require user testing to investigate whether it does have an effect upon error levels, and whether the touchscreen results in different error rates compared to a psychical interface.

There has been previous research conducted into two of the existing interfaces tested here, the chevron and 5-key interface. The chevron interface was tested in a lab environment and it was found that it performed better than the number pad in terms of numbers of errors made by participants (Oladimeji et al., 2011). This contradicts the suggestions made by evaluating these interface against the heuristics listed here, which found the number pad to be best for entering numbers relating to infusions.

The more accurate performance using a chevron interface was thought to be due to the participants' gaze during the experiment; the chevron interface encouraged participants to look at the display more than the number pad interface. One key aspect to note during this experiment is that the numbers being entered were randomly generated. A case has already been made that this may not generate the most ecologically valid results.

Despite this, the difference in results of the heuristic analysis performed here, and the experiment performed by Oladimeji et al., were ultimately testing separate aspects of the interface: one tests the number of errors made, the other the potential for error. This means that the two methods may lead to different interface recommendations and both should be taken into consideration when analysing interfaces.

The 5-key interface has been tested using computer modeling and simulations (Cauchi et al., 2012). This research simulated hundreds of instances of number entry on a 5-key interface. In a small proportion of these entries, incorrect key sequences were injected. This was tested over multiple different versions of the 5-key interface (for instance with and without digit wrap-around after the 9 key). These experiments found that preventing this wrap-around actually prevented more errors from occurring whilst programming the device. Again, this study used randomised numbers as the input. A later study used numbers taken from hospitals (Cauchi et al., 2013). This study again found that the best configuration for the 5-key interface was to have no digit wrap-around, which is contrary to the suggestion from the heuristics. This is explained by the differing aims of the simulation study and the heuristics. The simulation study aimed to lessen the impact of inevitable keying errors, whereas the heuristics presented in this study aim to make the device more usable for the tasks it is needed for. These two aims appear to work against each other; it is likely that a compromise is required between the two in order to make the device safe and usable.

5.3.7 Limitations

Using this heuristic approach is a quick method of assessing the interfaces. It is by no means a replacement for empirical testing. This method should be used in conjunction with user testing. Currently, the heuristic analysis does not take into account factors such as the environment, or user expectation. Additionally, it does not asses how well each of the interfaces perform in terms of error checking. For instance, as stated previously, the chevron interface encourages error checking behaviour from users (Oladimeji et al., 2011). This feature is not tested in this heuristic evaluation and ignoring it makes the chevron interface appear as a failure on all counts. This highlights the importance of the heuristic evaluation as just one tool amongst many that are required for interface evaluation.

These heuristics were generated from a log analysis performed at a single hospital, which used a single brand of infusion pump. That means these heuristics are likely to be highly specific to this one situation. It is not suggested that these heuristics be used in every circumstance unquestioned, but rather that heuristics might be developed for a number of different hospital settings. The contribution of this work is not necessarily the heuristics in their current form, but an exploration of the possible benefits of heuristic generation and evaluation for medical device design.

5.3.8 Conclusion

The design of number entry interfaces that are currently used on infusion pumps is not optimised for the tasks for which they are being used. By taking recommendations from the results of the log analysis, a series of heuristics can be created that ensure common number entry with fewer keypresses, which could reduce the chance for key slip errors.

The guidelines do not only have to be used as an analytic tool but can also be used as the basis of design. Here we have shown how a touchscreen scroll wheel interface may be designed to meet the needs of the infusion-programming task. Designs, such as the touchscreen scroll wheel presented here, need to be tested in laboratory environments before a recommendation can be made for real world usage, particularly as there is existing research that suggests alternative outcomes to those found from the heuristic analysis. However, this study has shown how results from the log analysis can be used to generate possible improvements in number entry interface design on infusion pumps.

When considering how to apply the findings from the heuristic analysis to the real world however, implementing an entirely new interface is not a trivial problem. This would involve the replacement of many existing devices and would require entirely new systems to be adopted by both manufacturers and hospitals alike. It would also require additional approval from regulators. Although such a program could be incredibly beneficial to staff and patients in terms of ease of use and guarding against errors, it is not the most practical solution. Instead of creating entirely new interfaces, the possibility of adapting and tailoring existing hospital number entry interfaces should be explored. The heuristics generated in this study can be used for this task.

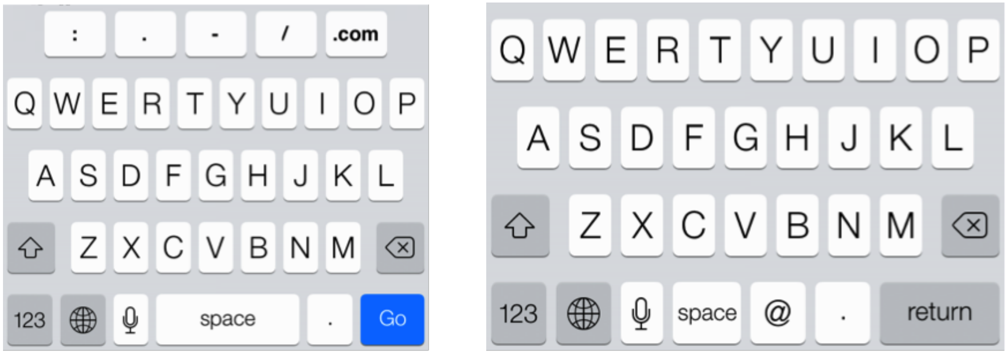
5.4 Study 9: Tailoring Number Entry Interfaces

The following experiment was conducted by Orla Hennessy, a Masters student at UCL. The experiment was run by Orla Hennessy, preliminary analyses are presented in her Masters thesis "Can task performance be improved by adapting number entry interfaces to fit the task?". My contribution during this Masters dissertation was the design and coding of the experiment. In this Ph.D Thesis, a more thorough analysis of the results is presented.

The results from this chapter were presented at the 2013 HFES Annual Meeting in San Diego, US. The work can be found in the paper Wiseman, Brumby, Cox, & Hennessy (2013a).

5.4.1 Introduction

Adapting number entry interfaces to fit the specific task they are used for could help reduce the opportunities for making errors by reducing the number of keypresses needed to enter data. This approach is not new: there are a number of examples of interfaces designed with character frequency in mind outside of the medical domain. One example is the .com button that appears on certain touchscreen phones when the user is typing a URL in the address bar (see Figure 5.9). The .com suffix is the most common of all top-level domains and designers of these interfaces have clearly incorporated its common use. Similarly, the @ symbol appears as a button on touchscreen keyboards when an email address is being typed.



(a) URL Keyboard

(b) Email Keyboard

Figure 5.9: Keyboard layouts for URL and email address entry on iOS 7.0.2.

On many financial calculators, there is a 00 button to make entering round monetary amounts into the system quicker and easier. Shop tills also often have buttons for common currency denominations. These shortcuts are based on the frequency of digits and characters that are entered in these particular settings. Context is important when designing these shortcuts: If the .com button appeared every time someone wanted to send a text message, it would be considered a waste of screen space as, for that task, the occurrence of the characters in that order would not necessarily be significantly greater than other strings. With knowledge about which digits and numbers are being used in hospitals, it might be possible to adapt the number entry interfaces on infusion devices in similar ways so as to meet the needs of the specific task. This approach has been used in the design of some opticians' software that is used for entering the prescription of lenses to be ordered (Figure 5.10). The user is provided with a standard number pad with some added buttons that represent the set of four possible end digits that a lens prescription may have (.00, .25, .50, and .75).

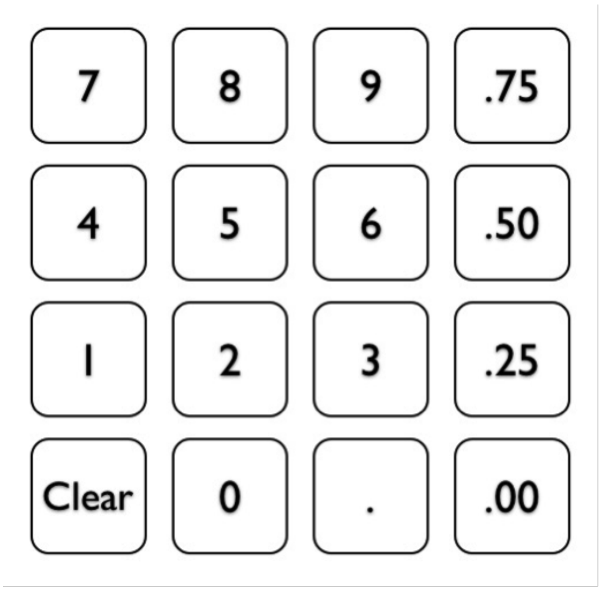


Figure 5.10: Illustrative example of a number entry interface for optician software. The interface is adapted to the task of entering a lens prescription.

One of the requirements for developing interfaces that are adapted to the task they are used for is to understand which numbers are commonly entered. In order to adapt the interface of an infusion pump, the findings from the previous log analysis, and the associated heuristics are used.

With such patterns emerging in medical numbers, it is clear that medical number entry interfaces could be tailored towards the task of medical infusions, by making it easier to enter the most common values. To evaluate this idea, three existing interfaces are tailored (by adding keys or altering the way in which button presses affect the number being entered). These redesigned interfaces are evaluated in an experiment in which participants are asked to enter a series of medical prescriptions, based on the numbers collected from the infusion pump log analysis. The study aims to understand how tailoring affects the speed and accuracy of the interface, as well as looking at user preference for tailoring.

Hypothesis

It is hypothesised that, firstly, the tailored interfaces will be faster to use than their standard counterparts. By reducing the number of button presses needed to enter the medical data, the speed of interaction will increase.

The second hypothesis is that the number of errors made on the tailored devices will be reduced. Again, by allowing the participants to have fewer interactions with the device, the chances for errors occurring reduces similarly.

5.4.2 Method

Participants

Thirty participants (13 female) from the UCL Psychology subject pool took part in the study. The age of participants ranged from 18 to 43 (M = 26.4 years, SD = 7.0 years). None of the participants were experienced in using medical devices.

Design

A mixed 3x2 (interface type x tailoring) design was used, where interface type was the between-subjects factor. The three interface types that were evaluated were: keypad, chevron and dial; these are common interface designs found on infusion pumps (Oladimeji et al., 2013). For each interface, participants used a standard and a tailored version (an explanation of the tailoring is given in the Materials section).

The dependent variables in the experiment included the time taken to enter a number and complete each trial, along with timing data for interkey interval timings (where appropriate). Keypress logs allowed for error data to be collected about the frequency of error. User preference was recorded at the end of the trials.

Materials

Participants interacted with a physical prototype of a representative infusion pump (Figure 5.11). The use of a prototype, rather than a computer interface, made the experiment more ecologically valid; the tactile feedback of a computer keyboard is very different to that of an infusion pump soft key interface. The interface of the prototype could be easily changed between a number pad, a chevron design and a dial.

The number pad interface (Figure 5.12a) has input buttons for the numbers 0-9 along with buttons for a clear function and a decimal point. This interface resembles that which is common on calculators and mobile phones, with each press of a digit key adding that digit to the display. Moreover, this entry interface can be found on many infusion pumps.

The chevron interface (Figure 5.13) has four buttons for incrementing and decrementing a number shown on a separate screen. Pressing the upwards-pointing chevrons would increase the number being displayed. This increments by a large amount if the double chevron is pressed, and by a small amount if the single chevron is pressed. Interacting in the opposite way would decrease the number. The chevron design is also common on infusion pumps.

For the dial interface, the participant turned a knob to increase (turning clockwise) or decrease (turning anti-clockwise) a number on the screen. Depressing the dial before turning allowed for smaller increments and decrements to be made; depending upon the configuration of the pump this small increment may either be smaller integers or numbers after the decimal place. While this interface is used less often on medical devices, it is nonetheless a familiar interface used in many other contexts, for instance, to control the volume of a music system.

As well as exploring different types of input device, the style of interface was also explored: standard vs. tailored. The tailored keypad interface had additional buttons on the interface according to heuristic 4 in the previous section.

The chevron and dial interfaces were adapted by changing the value of the increments and decrements. For the standard interface of the chevron and dial, the large increment and decrement value was set at 10. On the tailored interface, the large increment and decrement value was set to 25 in order to better match the types of numbers being inputted; the log analysis showed that 61.25% of numbers were divisible by 25. These two interfaces looked the same in both the standard and tailored condition, but had different effects upon the number on screen when in the tailored arrangement.

The prototype was mounted to a stand at a height that allowed participants to operate the device from a seated position. The numbers that the participants were to enter were presented in pairs as a prescription of volume and rate. These were displayed on a nearby laptop computer to better mimic a hospital setting, where medical workers copy numbers from a separate location (Back et al., 2012). During the experiment, participants were able to clear the number they had entered and begin again if they wished, for instance, if they had made an error.

The numbers that participants had to enter into the device were representative of the numbers collected during the log analysis. The sample maintained the ratio of "common" and

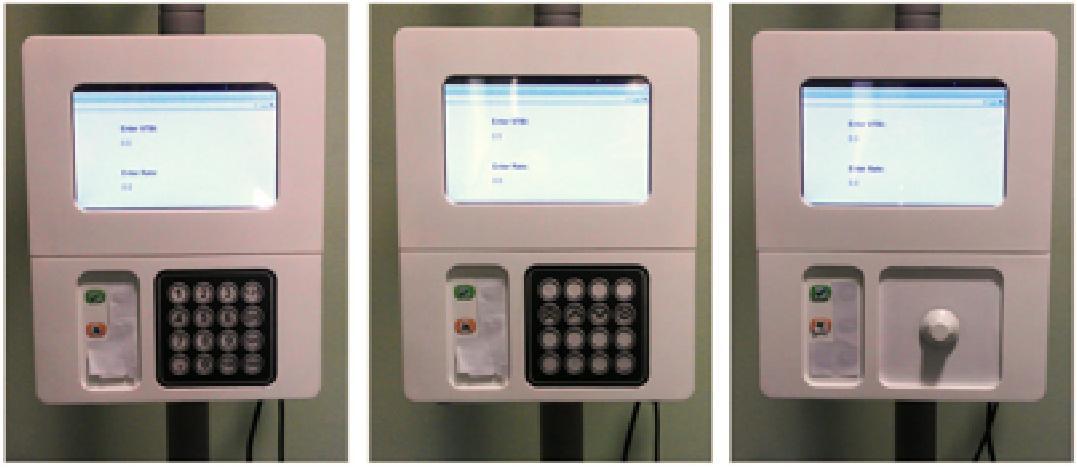
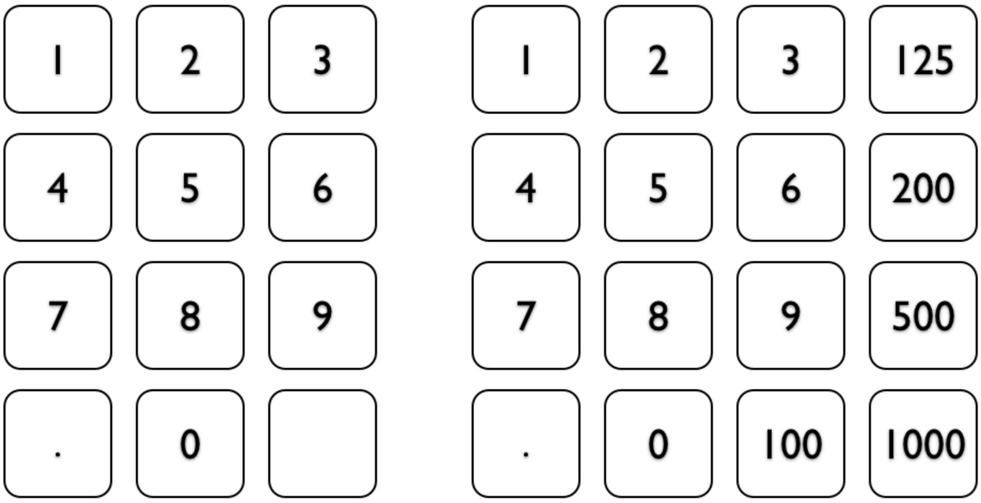


Figure 5.11: Prototype infusion pump set up with each of the interfaces



(a) Standard Numberpad

(b) Tailored Numberpad

Figure 5.12: Numberpad layouts used in the experiment

"uncommon" numbers to ensure that the tailoring did not cause the interface to become unusable in edge cases.

Procedure

Participants were randomly assigned to one of the interface designs, either keypad, chevron or dial. Participants were given an initial training period in which they entered 10 numbers to allow them to become familiar with the interface. In the main experimental phase, participants entered 60 pairs of numbers on both the standard and the tailored version of their assigned interface (the order of which was counterbalanced across participants). After each pair of numbers (volume and rate) had been entered, a new prescription was displayed on the laptop screen. After participants had entered each of the 60 number pairs they were

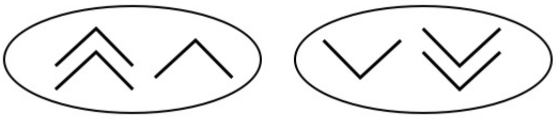


Figure 5.13: Chevron interface used in the experiment

given a 5 minute break whilst the interface was changed from standard to tailored, or vice versa. Critically, the same number set was used in both the standard and tailored condition, with the order of numbers being randomised within each condition. Participants were told to either enter the numbers quickly or accurately (half in each condition). Participants were also told that if they made an error and keyed the wrong digit that they could correct it by using a clear button before entering it.

5.4.3 Results

The time to completion, the number of keypresses and number of errors for the six interfaces are analysed here. For statistical analysis, effects were judged significant at a .05 significance level.

Key Presses

To better understand how participants were interacting with each of the interfaces, the total number of keypresses needed to enter a given number is an important consideration. Both hypotheses made prior to the experiment rely upon tailoring resulting in fewer keypresses. It is important, therefore, to check that this was indeed the case and that participants were able to make use of the tailored interfaces effectively. (The term 'keypress' is used to describe all interfaces, on the dial interface this relates to one click of the wheel when turned in either direction.)

Results show that participants made significantly fewer keypresses when entering a number using the keypad interface (M = 3.3, SD = 0.1) than when using the chevron interface (M = 12.6, SD = 0.7) or the dial interface (M = 15.4, SD = 0.8), $F(2, 27) = 1105.24$, $p<.001$. Participants also made significantly fewer keypresses when entering a number using the tailored interface (M = 8.7, SD = 4.4) compared to the standard interface (M = 12.2, SD = 6.2), $F(1, 27) = 420.37$, $p<.001$. The interface type x tailoring interaction was also significant, $F(2, 27) = 62.41$, $p<.001$. A follow up test of this interaction shows that the benefit provided by tailoring the interface, in terms of fewer keypresses being needed to enter a number, extended to all interface types (all p's <0.05).

Going beyond these data, it is also possible to determine how well the participants adapted to the tailored interfaces. This is done by comparing the keypresses made by participants to how many keypresses or dial rotations would be made on each of the interfaces using "optimum" strategies. This was calculated by finding the minimum number of keypresses required to enter each of the numbers in the target number set used in this experiment. Comparing the participants' keypresses to these estimates of the optimal strategy determines how well participants had learned to use the interface in its standard and tailored form. Table 5.6 shows the optimal performance compared to the actual performance on each of the interfaces. It can be seen that participants were performing close to optimal on the keypad, but were making many more keypresses than were necessary for the chevron and dial interface in both their standard and tailored forms.

	Standard Interface		**Tailored Interface**	
	Actual *M (SD)*	*Optimal*	*Actual* *M (SD)*	*Optimal*
Keypad	3.7 (0.1)	3.7	2.9 (0.1)	2.9
Chevron	15.0 (0.7)	14.3	10.2 (0.7)	9.5
Dial	17.8 (1.1)	14.3	13.1 (1.1)	9.5

Table 5.6: Mean number of keypresses needed to enter the numbers using an "optimal" strategy compared with the mean number of keypresses used by participants in the experiment

Time

The time taken to enter each pair of numbers (i.e., to complete the prescription task) on each of the interfaces is considered, the data can be seen in Figure 5.14. Results show that participants were significantly faster at entering (pairs of) numbers using the keypad interface (M = 5.57s, SD = 1.45s) than when using the dial interface (M = 7.81s, SD = 0.84s) or the chevron interface (M = 12.37s, SD = 2.12s), $F(2, 27) = 49.41$, $p<.001$. Participants were also significantly faster at entering numbers using the tailored interface (M = 8.14s, SD = 2.90s) compared to the standard interface (M = 9.03s, SD = 3.71s), F(1, 27) = 20.43, $p<.001$. However, the interface type x tailoring interaction was also significant, $F(2, 27) = 14.56$, $p<.001$. Follow up tests of this significant interaction show that the speed benefit provided by tailoring the interface was limited to the chevron interface, $F(1, 27) = 47.30$, $p<.001$; there was no significant benefit of tailoring the interface for either the keypad, $p = .17$, or the dial interface, $p = .64$.

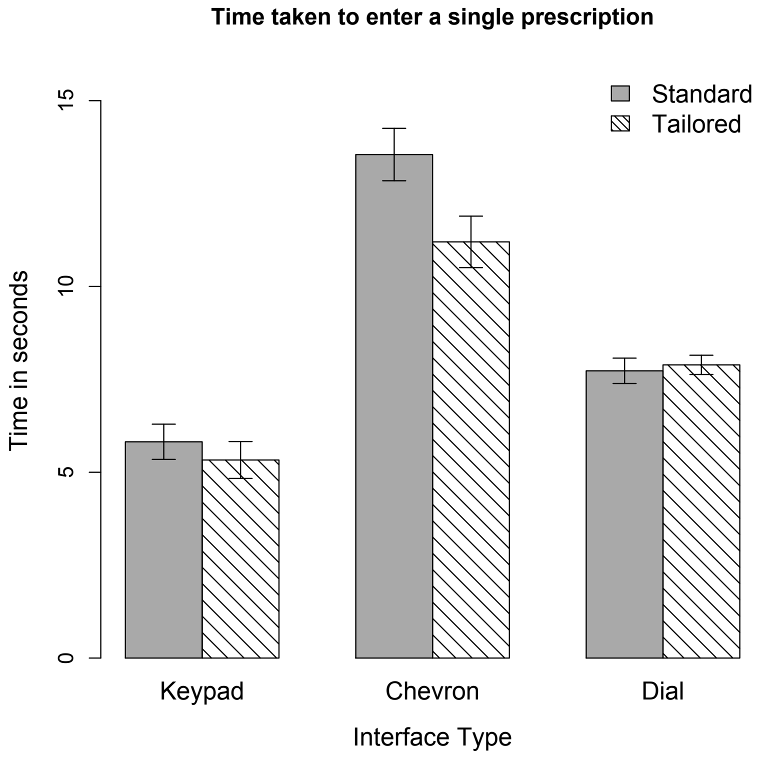


Figure 5.14: Graph showing average time taken on each interface to enter a single prescription

Errors

The number of errors made on each interface was also a key measurement during this experiment. Errors were defined as any number that was entered which did not match the target number on the prescription. In general participants made few errors; the mean error-rate across all conditions was 0.99% (*SD* = 1.07%). Error-rates were marginally higher when participants used the keypad interface (M = 1.42%, *SD* = 1.52%) compared to the chevron (M = 0.87%, *SD* = 0.67%) or the dial interface (M = 0.67%, *SD* = 0.77%). However, this difference was not statistically significant, p=0.89. There was also no difference in error-rate between the standard (M = 1.00%, *SD* = 1.36%) and tailored interface (M = 0.97%, *SD* = 1.05%), p=0.28. The interaction was also non-significant, p=.21.

User Preference

Finally, participants were asked for their reaction to the tailoring of interfaces; the feedback was varied. Tailoring was preferred for the keypad (8 out of 10 participants preferred the tailored version), but was widely rejected for the dial interface (9 out of 10 participants preferred the standard version). For the chevron interface, participants did not have a clear preference (6 out of 10 participants preferred the tailored interface).

Participants were also asked how familiar they were with the interface that they had used. Nearly all participants in the keypad interface condition and the dial interface condition had experience using similar interfaces. However, participants in the chevron condition had the highest number of participants (30%) with no previous experience of that interface.

5.4.4 Discussion

This work was conducted in order to find out whether tailoring interfaces, by making them better suited to the tasks they are used for, could make interaction with them faster and less error prone. Here the key hypotheses are evaluated to understand the success of the tailoring approach.

Timing

The tailoring successfully reduced the time needed to program the chevron interface, but had no significant effect on the time taken to program the keypad and dial interfaces. This shows that tailoring, in some circumstances, can have a beneficial effect upon speed. However, this benefit was only seen on one of the three interfaces.

This was an unexpected result, as all tailored interfaces showed significant improvements in terms of keypresses required to enter a prescription. A reduction in number of keypress or dial turning actions would likely result in a reduction in interaction time. Since this was not the case, it is likely that another issue was causing interaction times to remain the same.

One explanation for this effect would be the time taken for participants to stop and think about the next action they would need to take. All participants were familiar with the standard keypad and dial interface and so only a small amount of time was required to think about which actions to take. When switching to the tailored version of each interface, the participants would need time to learn about the new features of each interface, thus increasing the time taken to interact with them, despite reducing the number of keypresses. This issue was not seen in the chevron interface, as fewer participants had experience with the standard interface, meaning the same amount of cognitive processing was required to interact with both, meaning a reduction in keypresses would result in a reduction in time taken to use the interface.

It is possible that, with more training, participants would become more familiar with the interfaces and would learn the new features or adapt their behaviour, thus reducing the time needed for cognitive decisions and potentially reducing the time taken to interact with the interface. Indeed, the analysis of optimal keypresses shows that for the chevron and dial interfaces, more optimisation is possible in terms of number of keypresses.

Error Rate

It was shown that the keypresses were successfully reduced on the tailored interfaces, however, the analysis shows that this did not result in a significant reduction in error rates as hypothesised. Error rates during the experiment were low in all interface conditions. Further work might investigate how the tailored and standard interfaces perform in multitasking or high stress situations where errors might be more likely (Wiseman et al., 2011). This result does suggest that the timing gains for the tailored chevron interface were not due to a speed/accuracy trade-off, that is, the faster interaction did not result in more errors.

Previous research has been conducted into the error rates associated with the number pad and chevron interfaces. Oladimeji et al. (2011) asked users to enter numbers on a simulated pump displayed on a computer screen. Using the mouse, participants would enter numbers on both a number pad and chevron display. Throughout the trials, errors were artificially inserted into the displays. The results from this work showed that significantly more uncorrected errors went unnoticed whilst using the number pad interface compared to the chevron interface. This result was not replicated here. There are a number of possible reasons for this; one being the style of interaction. It may be that the difference between pressing buttons on a real device, as in this experiment, is different to clicking with a mouse on screen, as in Oladimeji et al. (2011), errors in one situation may occur less frequently in the other. Another issue that prevents direct comparison was the insertion of errors into the display of Oladimeji's work; the uncorrected errors may not have been the result of a slip made by the participants in the experiment. This supports the Oladimeji's hypothesis that it was the extra checking that participants were doing when using the chevron display that resulted in reduced uncorrected error rates, rather than the chevron interface lowering the number of errors made during the input task.

User Preference

Prior to the experiment, no hypothesis was made as to how participants would react to the tailored interfaces. This was explored using a simple question post-experiment to understand which of the two interfaces they preferred. The results showed that the tailored interface was preferred for the number pad, and the standard for the dial, but no clear distinction was made for the chevron interface.

It is hypothesised here, that this may be due to the participants' familiarity with the devices and the types of tailoring. For the purpose of discussion, the tailored interfaces are split into two types: visual and functional. The number pad was visually tailored, with extra buttons being placed on the interface. The chevron and dial interfaces, however, were functionally tailored, in that they looked the same, but the way that they functioned had changed.

Participants with experience of the number pad interface were easily able to identify how it had been altered from its standard state, as the tailoring was visually salient. Not only that, but if they wished, they could continue to use the interface as a standard number pad. For this reason, the tailored interface was preferred, participants could understand both the tailoring and the benefits that the tailoring produced.

For the dial interface however, experienced participants were unable to see how the interface had changed from their expectations. Only with interaction was the tailoring clear. This may have resulted in frustration when the participants, with expectations of a standard interface, found out that the device acted in a seemingly unpredictable way. This combination of participant prior knowledge and hidden, functional tailoring, resulted in participants preferring the standard layout.

Participants had least experience with the chevron interface, and were therefore expecting neither the standard, nor tailored version of it. This meant they were not surprised, or shocked when using the tailored chevron interface, the standard interface was equally new to them. Participants with no prior experience of the interface were not alarmed by the functional tailoring, because they had no previous knowledge of the interface to deviate from.

This result provides an interesting lesson for any type of interface tailoring: if the tailoring is functional, and not obvious upon visual inspection, or if the standard functionality has changed, then participants with previous experience of the interface may find it frustrating and difficult to use. A similar result was noted by Google when managing the transition of the Google Drive interface, one method employed for making the transition easy for the users was to maintain the previous functionality for as long as possible (Sedley & Müller, 2013).

5.4.5 Limitations

One limitation of this study was the participants' inexperience of the medical domain. Although the task used an infusion pump-like device and realistic data, participants were not medical professionals and, therefore, did not have the same level of prior experience. It has been shown in Chapter 4 that a user's familiarity with the numbers they transcribe has an effect upon various typing metrics, meaning this data is not likely to be representative of how a medical worker may have completed the task. However, there is no reason to believe that comparisons between interfaces would not still hold, if not their absolute values. Any speed or accuracy improvement gained by having familiarity with the numbers being entered should affect each interface equally, thus maintaining the relative differences between the designs.

It may also be useful to investigate the effects of manipulating motivation in this study in the future. There were no obvious penalties for entering numbers incorrectly in this experiment, which is obviously at odds with the real world application of this to the medical domain. Penalties for incorrect number entry may have encouraged the participants to work more slowly and accurately.

5.4.6 Conclusion

Interfaces can be effectively tailored to meet the needs of the task with benefits including reduced interaction time and fewer keypresses without effecting error rates. Tailoring interfaces does not, however, appear to offer a solution to reducing user error. A different approach for reducing error in the number entry domain should be taken.

An unexpected result of this experiment was the analysis of user preference. User preference for interface tailoring is affected by the familiarity of the original interface. When tailoring a familiar interface, users prefer adaptations to be visually salient. In the future, tailoring may make it possible to create an interface that users prefer, is faster, and reduces keypresses and thus the chance for error.

This experiment used an understanding of the numbers found in the medical domain to tailor number entry interfaces and thus make them better adapted for their task. In the present experiment this involved increasing speed of interaction. However, the results of this experiment also showed that this approach had no improvement upon error rates. Therefore, alternative approaches are required. In addition to the high levels of frequency of some numbers that were identified by the analysis of the numbers used in this context, the analysis also revealed that the numbers have a special relationship with each other. As stated in Study 7, there are three numbers used to specify each infusion and two of the numbers can be used to calculate the third. This feature provides a natural “checksum” that can be used to identify error. In the following experiment, this feature of the numbers used to program infusion devices is utilised to design an interface that can check for number entry error.

5.5 Study 10: Checking for Number Entry Errors in the Medical Domain

The following experiment was conducted by Sarah O'Carroll, a Masters student at UCL. The coding and running of the experiment was done by Sarah O'Carroll, preliminary analyses are presented in her Masters thesis "Mistakes with Numbers: How interface design can influence number entry errors". My contribution during this Masters dissertation was the design of the experiment. In this Ph.D Thesis, a more thorough analysis of the results is presented.

The results from this chapter were presented at the CHI 2013 conference in Paris, France. The work can be found in the paper Wiseman, Cox, Brumby, Gould, & O'Carroll (2013c).

5.5.1 Introduction

It has been shown how critical accuracy is when entering numbers in a medical domain. The results of errors can be catastrophic, leading to patient harm or death. This thesis has highlighted a number of ways to attempt to rectify this issue. The first being the importance of testing with ecologically valid numbers. A more direct attempt was made in the previous study, where number entry interfaces were tailored to better match the data being entered. However, although tailoring was shown to increase the speed of entry, it did not affect the participants' error rates.

Preventing error is a difficult task. Thimbleby & Cairns (2010) have shown that through design, it is possible to prevent incorrectly formatted numbers from being entered into a system, for example, by recognising particularly error prone infusions (as taken from Institute for Safe Medication Practices (2006)). This error blocking helps prevent systems from acting in unpredictable ways. In the paper, the authors highlight the various ways that a machine could interpret the entry sequence $\boxed{1}\,\boxed{\bullet}\,\boxed{2}\,\boxed{\bullet}\,\boxed{3}$ (where each box represents a separate key press on the machine and thus equates to the non-legal number string 1.2.3). Different interfaces interpret this in a number of different ways, some as 1.23, 1.2, 1.3 or even 1.5. Preventing users from entering ambiguous number strings, such as 1.2.3, can help prevent unpredictable errors. This paper shows that with well-defined syntactical rules, some errors can be caught before entry. However, this accounts for only a subset of number entry errors and cannot help in preventing legally formed numbers from being entered into a system.

Another approach to preventing a wider range of number entry errors is to encourage users to check for errors themselves. This system would catch more errors than the previous.

One suggestion for achieving this is to use purposefully unfamiliar number entry interfaces (Oladimeji et al., 2011). This research shows that users spend more time looking at the display when using the chevron interface, compared to the number pad. The reasons behind this are two-fold, firstly the number pad provides the user with direct feedback about the number they have entered through the button the user presses, meaning they do not have to look at the display. Such feedback is not given when using the chevron interface, as the keys increment the display, rather than adding a predefined, static number to the display, meaning users need to check the display to know what number they have entered. The second reason that users looked at the display more often when using the chevron interface was that the participants lacked experience with that interface and were therefore double checking their actions more often. Whereas, participants were confident when using the number pad, as it is a more common interface to find in every day life. This confidence translated to the users checking the number display less frequently.

This approach led to a reduction of errors when compared to error rate for the standard number pad interface. However, some errors still went unnoticed. Additionally, this research was conducted on participants unfamiliar with the interface. There is no way of knowing if this effect would still be true for expert users who would be familiar with the interface and thus more confident, potentially leading them to check for error less frequently.

This previous research shows that it is possible to reduce error. That no clear solutions to the problem have been discovered suggests that more research is required and that solving the problem is a particularly difficult task. Any attempt to reduce problems arising from number entry errors in a given practically important domain needs to take into account the specific conditions that prevail when numbers are entered and the other related measures that are taken in that context (see The Health Foundation (2012) for an extensive survey of these issues in the context of prescribing medication). In this section, the aim is not to solve a particular practical problem but to initiate research on an approach that may ultimately prove to be applicable in a variety of contexts.

Motivation

The process of users entering data into a machine can be likened to the process of sending a signal over a noisy channel. Both involve the transfer of data between two locations and both are potentially subject to errors occurring during that process. To catch errors in the case of signal processing, error-correcting codes are used. This involves the sender providing redundant information, along with their message. This redundant information can be used to detect if any errors have been made during transmission and can allow for error correction. It is possible to apply this method of error detection to cases where a user has to enter data.

More specifically, the key idea is that of a checksum: an additional (redundant) number that is related to the to-be-entered numbers in such a way that it is sufficient to verify the correctness of the checksum. The benefit to this solution is that it asks the user to view only one number, and does not require the checking of each of the entered numbers, which has been shown not to be a successful strategy (Olsen, 2008). As was suggested by Thimbleby (2008) if a checksum is presented along with the numbers to be entered and the checksum is also computed on the basis of the numbers that the user has actually entered, it is possible to check the correctness of the numbers simply by comparing the two checksums.

Sometimes checksums occur naturally, this is the case with infusion pump programming in the medical domain. Infusion pumps are used to administer a drug to a patient intravenously over a period of time. Three values specify each infusion: time, volume to be infused (VTBI) and rate of infusion. The relationship between these values is such that knowing two of them allows the third to be calculated:

$$VTBI = Time \times Rate$$

As a result, infusion pumps often require only two of these values to be entered; some may compute and display the third value. Hence the third value can serve as a naturally occurring checksum.

There are two ways in which a checksum can be used for verification: (a) The user is prompted to compare the original checksum with the computed checksum; or (b) the user is required to enter the checksum along with the other numbers and the computer does the comparison. (This latter method is somewhat similar to the use of cross-validation in form design where the same piece of information is requested in two different questions to ensure it is correct (Chen, Hellerstein, & Parikh, 2010)) In this study, both of these methods are tested, using a 2-number and 3-number entry interface (an explanation of each is found in the Materials section).

Hypothesis

The hypothesis is that participants will complete number entry tasks with the 2-number interface more quickly (because fewer key presses are required); but that entry will be more accurate with the 3-number interface, which makes it almost impossible not to notice an error that has been made and which does not rely on the correctness of visual comparison of numbers by the participants.

5.5.2 Method

Participants

Twenty-four participants (13 male) were recruited from the university subject pool. The participants had no previous experience with medical devices. The mean age of participants was 34.7 years (range 21-59, *SD* = 10.8). All were experienced computer users who either had normal or corrected to normal vision.

Design

The experiment was a within-subjects design and the order in which the interfaces were presented was counterbalanced. The independent variables were the interface type and the instruction. There were two levels of interface type: the 2-number and the 3-number interface.

The 2-number interface required participants to enter two numbers (e.g., VTBI and Rate). It then calculated the third number (e.g., time) and prompted the participant to compare this "checksum" with the third of the presented numbers. The 3-number interface asked the participant to enter all three numbers; the computer then performed a calculation with the first two numbers to test if the result was the third. The computer displayed the feedback "correct" or "incorrect" depending on whether the expected result and inputted number matched. Screen shots of both interfaces can be seen in Appendix C.

The instructions given to the participants were either to perform the task as quickly as possible or as accurately as possible. This decision was made in order to ensure that speed accuracy trade-offs were accounted for (Brumby et al., 2011). The dependent variables for this experiment were the number of errors and the time taken to complete the trials.

Materials

The experiment was conducted using a desktop PC. The participants were provided with a standard computer keyboard, and were asked to use the numeric keypad at the right of the keyboard when entering the numbers. Navigation through the interface was done using the tab and arrow keys.

With both interfaces, if a participant noticed that they had made an error, they could correct it by using the backspace key before moving on.

Each participant would have to copy numbers from the screen into the number entry fields. Data to be input was presented in columns of VTBI, Rate and Time. Each row represented one ‘prescription’ that would have to be entered before submitting (see Figure 1). The prescriptions were displayed in black Arial 12pt font. The numbers were taken from the numbers collected in the Log Analysis work in Study 7.

There were two phases in the experiment. In the no-choice phase, participants used both the 2-number and 3-number interface. For each interface, participants were required to enter six sets of 20 prescriptions. Each prescription set required the entry of 40 numbers when using the 2-number interface and 60 numbers when using the 3-number interface.

After each prescription entry on the 2-number interface (which required the participant to enter a VTBI and Rate), the display calculated what the expected third (Time) number was. The participants were asked to compare this calculated number to the expected Time displayed in the prescription list. Participants then had to select either the ‘yes’ or ‘no’ option to say whether the numbers matched and continue with the rest of the list of prescriptions.

On the 3-number entry interface, participants were required to enter all three numbers that comprised the prescription. The computer then performed the expected calculation and if any discrepancies were found, the user was instructed that they had entered a number incorrectly.

On neither interface was the user penalised for entering a number incorrectly, but half of participants were instructed to work as accurately as possible.

The second phase was the choice phase, during this phase, participants were asked to choose which interface they wished to enter numbers on. Participants were asked to choose an interface that would allow them to work as quickly as possible, and were then asked to choose an interface that would allow them to work as accurately as possible. For each of these tasks, the participants entered one set of 20 prescriptions.

Procedure

Participants were brought into a lab setting to complete the experiment. The task was to enter a series of numbers, either as quickly or as accurately as possible (this variation was counterbalanced throughout, with participants using each interface with each condition).

Participants were given instructions for the task and then completed a training session, which consisted of entering 10 prescriptions (5 for each interface). Firstly they completed the no-choice, then the choice phases of the experiment. Participants were offered a break before the choice phase began. The experiment took an hour to complete. After completion of all tasks, participants were debriefed.

5.5.3 Results

The measures taken during this experiment were the time taken to enter each prescription and the proportion of initial number entry errors made. For statistical analysis, a dependent t-test was used with a .05 significance level. Data from one participant was corrupted and so was removed from the analysis. Two participants did not complete the choice section of the experiment due to time constraints. It was found that the instruction given to the participants to either perform the task as quickly or as accurately as possible during the no-choice sections had no significant effect on error rate or time.

No-Choice Phase: Performance Profiles

Time The mean time taken for each participant to complete the task was significantly shorter for the 2-number interface (M=7.6 s, SD=2.6 s) than for the 3-number interface (*M* = 11.1 s, *SD* = 4.1 s), $t(22)$=5.3, p<.001. The increase in time of 46% for the 3-number interface is understandable in that it requires the participant to enter 3 rather than 2 numbers (though the 2-number interface calls for an additional visual comparison operation).

Number Entry Errors A single entry error was counted when the participant entered at least one of the (two or three) numbers of a prescription incorrectly. Reporting error rate per prescription, rather than per number allows direct comparison between the two interfaces. The percentages of entry errors in the 2-number interface condition (*M* = 5.14%, *SD* = 3.98%) and the 3-number interface condition (*M* = 6.67%, *SD* = 2.93%) were not significantly different, $t(22)$=1.565, p=.132.

After considering the occurrence of errors, the next analysis looks at the likelihood of participants noticing these errors. To do this, the percentage of corrected errors is calculated, which is the number of entry errors that were corrected after they were made, divided by the total number of entry errors made. With the 2-number interface, participants noticed 36.18% (*SD* = 34.58%) of the entry errors when prompted to check for them. With the 3-number interface, 100% of all entry errors were flagged by the system after its checksum computation and the participants corrected all of these.

Choice Phase: Participant Preference

For the final two sets of numbers, which were presented under the "fast" and "accurate" conditions, respectively, in that order, participants were given a choice of interface to use. In 37 of the total of 42 choice trials that were completed, participants preferred the 2-number interface; there was no reliable difference in choices between the “fast” and “accurate” instructions.

5.5.4 Discussion

Timing

The hypothesis that the 2-number entry interface would allow for faster interactions was supported during this experiment. This was perhaps an unsurprising result as, although in the 2-number interface, participants had to complete a visual check step, the 3-number entry interface required participants to enter in 50% more numbers. Participants were acutely aware that this interface was quicker and were most likely to choose it when told to complete the task as quickly as possible.

Increasing speed of interaction is a common aim when designing interfaces; we often strive to give users a quick and error-free experience. Based on the timing result, it appears that the 2-number entry interface is the most suitable way to use checksums. However, in a hospital domain, speed cannot be the only factor against which an interface is assessed. It is important to understand how this speed affects error rate. It may not be that the 2-number entry interface is the best choice.

Error Rate and Error Checking

The error rates for both interfaces were comparable; interface design had no significant effect upon the number of errors made during the experiments. Ordinarily an analysis of errors would stop at this point: using a simple error rate analysis shows neither interface to be superior to the other. However, an analysis of the error checking showed that participants using the 2-number entry interface were less likely to notice their errors.

The 2-number entry interface did encourage some error checking behaviour. However, participants failed to detect 64% of the entry errors that they had made, even though the checksum presentation reduced the verification step to a comparison of two new and not previously typed numbers. This result suggests that users are not thorough enough when checking for errors.

When asked to choose an interface that would allow them to enter numbers as accurately as possible, nearly all participants picked the 2-number interface, despite its poorer performance. This suggests that either the participants were unable to determine the accuracy of the two interfaces they used, or that they were not honestly making a choice based upon accuracy, and were choosing the interface that they perceived as easiest to use.

That humans are poor at performing checks on their own work is not a new finding. This issue is seen in real world examples of error checking. In the hospital, often nurses will double check each other's calculations before programming infusion pumps. However, this checking is sometimes unsuccessful if the medical workers do not thoroughly perform the check. This was seen in the Fluorouracil overdose case (Institute for Safe Medication Practices Canada, 2007), where a nurse's checking calculations failed to notice an error made during a calculation completed by another nurse.

This result suggests that, in order to use the checksum method to achieve the most accurate and safe results, a 3-number interface should be used, despite the time cost. In reality, it is likely that a combination of speed and accuracy requirements will need to be carefully balanced, in some circumstances speed may be of essence.

The results from this experiment present an initial exploration of the checksum method and show that the approach can have significant effects upon the number of errors committed into a system by the user. The results suggest that further analysis of this approach would be beneficial. When checksums are to be compared by the user (as in our 2-number interface), it should be possible to make this comparison easier than it was in the current interface. For example, the checksum could be displayed in a consistent, distinctive way both in the original prescription and in the computer's feedback. Additionally, alternative forms of displaying the checksum could be investigated, for instance, the values could be used to produce distinctive simple patterns.

5.5.5 Limitations

One factor affecting these results might be the experimental setting: participants were not penalised in any meaningful way for entering inaccurate information. In situations where the results of incorrect number entry were more costly, it is possible that users may pay more attention to the information provided by the checksum in the 2-number interface; or be more willing to use the slower 3-number interface.

The real world application of this system would also require changes to be made in the domains where this checksum system was used. Currently, only two numbers are provided on prescriptions (Back, Iacovides, Furniss, Vincent, Cox, & Blandford, 2013). For either of the two interfaces tested in this experiment, all three pieces of information used to define a prescription would need to be available to the medical worker.

One concern with this system is that it is possible, when using the 3-number check system, to enter incorrect numbers and still be given a positive result by the system. One easily imaginable way is for a user to mix up the Time and Rate values. Multiplication of the two would still provide the correct VTBI, regardless of whether the Time was entered into the Time or Rate fields. The system could also fail to catch an error if the user entered two of the three numbers incorrectly. For instance, Table 5.7 shows how the user adding an extra zero, and omitting a zero could result in a correct result for the system, but an unexpected result for the user, who would have entered a faster infusion than they intended. Despite the possibility for these errors to still be submitted to the system, they make up only a fraction of the errors possible, the rest of which the system would catch.

Measure	**Prescription Values**	**Entered Values**
Rate	50	500*
Time	20	2*
VTBI	1000	1000

Table 5.7: Example of how two number entry errors (marked with an asterisk) can result in a valid calculation.

Currently, the corrected error rates can only be compared between the two interfaces tested in the experiment. The use of a control condition, with no encouragement to check for errors, would have shown how the two interfaces improved upon the base rate checking behaviour.

5.5.6 Conclusion

This experiment illustrates the strengths and limitations of the two ways in which number entry errors can be caught with the help of a verification process based on a checksum. Application of either system to a real world domain would require a thorough consideration of when accuracy outweighs the need for a quick interaction, and vice versa.

The work presented here introduces a new approach to encouraging error checking behaviour. The approach highlights how a knowledge of the numbers used in a specific context can aid the design of novel number entry interfaces.

5.6 Discussion

The work in this chapter explored the importance of understanding domain-specific number patterns. Previous research into number distributions has been limited, yet work into understanding digit distributions has shown the robustness of patterns in various domains (for instance, the pattern of terminal digits in blood pressure measurements (Wen et al., 1993)). Study 7 has shown that it is not just possible to see patterns within digits used in a specific context, but also to see patterns in full numbers. This finding has not only been useful for the experiments within this chapter, but can also influence other research into medical number entry error, which often uses randomly generated numbers (Thimbleby & Cairns, 2010; Cauchi et al., 2012; Oladimeji et al., 2011; Wiseman et al., 2011).

The implications of this result go beyond the medical domain. There is no reason to believe that these patterns will exist solely in hospitals, but may occur in other contexts where number entry is important. In some cases, it may be that every number is equally likely and in these situations randomly generated numbers will suffice for testing and evaluation. However, in some cases there may be patterns in the numbers used and so it is important that number entry researchers use numbers that are representative of the task or can justify their choice to use random numbers.

The work in this chapter also highlighted how understanding more about the numbers that are used in a specific context can have implications for devices used in that setting. Using this knowledge, two different approaches to changing number entry interface design were tested. The first finding of this application to design was that it is possible to adapt an interface to suit common tasks, whilst still allowing for variation in input. Without, firstly, the knowledge that some numbers are more likely than others and, secondly, what those common numbers are, such an adaptation would not have been possible.

The specialist knowledge about the numbers being entered into these devices also allowed a key feature of medical numbers to be taken advantage of: the natural checksum. This aspect of the numbers that are used to program infusion pumps provided a robust solution to error checking, as reported in Study 10. It was only through exploration of the numbers entered into devices that this design feature could be developed and tested.

5.7 Limitations

One important limitation throughout the two experiments in this chapter was the lack of medical experience of the participants. It was important in these experiments that the numbers used when evaluating the devices were taken from the set of data logs in Study 7. These numbers, although highly familiar for users programming infusion pumps, are not familiar in more general contexts. This meant that the participants in the experiments, who had never programmed an infusion pump, would not find the numbers familiar. The finding of the work in Chapter 4 was that people enter familiar numbers more quickly than non-familiar and, therefore, evaluations of number entry interfaces should be tested with numbers familiar to the participant. The experiments conducted in this chapter break that rule. By using non-trained participants during these experiments, both constraints could not be met. Either the users could enter numbers familiar to them, or the devices could be tested with ecological valid number sets. The solution to this issue is to use medical workers when testing these design interventions; this is the next logical step for future research in this domain.

5.8 Summary

This chapter has highlighted a variety of benefits associated with understanding the numbers that are used in a specific context. Firstly, it has shown that the theoretical findings in Chapter 4 have implications for the way that we study number entry: users in the real world, especially in the medical domain, do have sets of numbers that are used frequently and are thus likely to be familiar.

Secondly, the knowledge of which numbers are used when programming infusion pumps has had serious implications for the design and evaluation of medical devices. Initially, through log analyses, it was shown that there are clear and distinct patterns in the numbers used to program infusion devices (Study 7). By summarising these findings into a short set of heuristics, a heuristic analysis of existing number entry interfaces showed that current infusion pumps are not necessarily well-designed for the task (Study 8). Using these heuristics to redesign current interfaces shows how tailoring can help make interaction with infusion devices faster (Study 9). However, this did not significantly reduce the errors made. A further experiment showed that using a unique feature of the numbers used in the medical domain (the validation between Time, Rate and VTBI), interfaces can be designed to reduce the chances of errors being committed when programming an infusion pump (Study 10).

In summary, this chapter has shown that number analysis can have serious benefits on evaluation and design of number entry interfaces. It is important for number entry researchers to consider number distributions when designing interface solutions.

Chapter 6

General Discussion

6.1 Introduction

In this chapter, the results of all studies and experiments contained within this thesis will be summarised. Following this, the implications of these results are explored and the key contributions to the areas of text transcription and number entry are enumerated. The chapter will also consider the limitations of the research presented in this thesis, and will outline directions for future research.

6.2 Research Summary

The key hypotheses and findings of each study are presented here; these findings are presented as a way to summarise the work completed in this thesis, along with presenting the reasoning and justification behind the studies that were conducted.

6.2.1 Study 1: Gathering a set of familiar numbers

The research presented in Chapter 2 highlighted that the familiarity of a number can potentially affect the transcription process. In order to explore this idea empirically, it was necessary to create a list of familiar numbers. As familiarity is a subjective measure, varying from person to person, it was important that this list was not created by the researcher.

In an online study, participants were first asked to rate 50 numbers as familiar or random, and then to generate their own familiar numbers. These were then rated by subsequent participants. The results of this study showed that participants were able to distinguish between the concepts of familiar and random with regard to numbers. Another outcome from this study was that a list of 73 familiar numbers was created. This familiar numbers list contained numbers that were rated by 4 or more participants, and were rated familiar at least 75% of the time.

In addition, the time it took for participants to respond to each number in the familiarity decision task was collected. The analysis of this response time data showed that participants were significantly faster when rating a familiar number than when rating a random number. This finding replicates similar results from lexical decision tasks, where words elicit faster decision times that non-words.

6.2.2 Study 2: What makes a number familiar?

The first step in understanding the effects of number familiarity on the number transcription task was to create a list of familiar numbers. However, no information was gathered about the *reasons* behind the numbers' familiarity. This would be important for studies with different groups or populations of users, as some culturally significant numbers may be unfamiliar. In order to do this, it was important to establish the reasons for familiarity, so that new lists could be generated. Previous work had highlighted some possible causes (Cohen et al., 1994; Delazer & Girelli, 1997), but had not provided enough information to be usable.

A card-sort was conducted with participants to establish possible reasons for number familiarity. Participants were asked to arrange the 73 numbers taken from Study 1 into categories. The total number of categories produced was 62, with several re-used categories between participants. These categories were then grouped, to find patterns in the reasons for familiarity. In total five groups were formed: "Culturally significant", 'Mathematical", "Pattern", "Number format" and "Personal significance". These groups can be used either to generate new familiar number lists, or as an aid to begin future replications of Study 1.

6.2.3 Study 3 and 4: Eye-Hand Span

The next step in the thesis was to replicate the Salthouse phenomena experiments. These experiments were conducted using words and non-words, and familiar and non-familiar numbers. There were two main aims in this experiment. Firstly to investigate whether there were differences in typing metrics between words and numbers, and secondly if there were differences between familiar and non-familiar numbers.

The first span tested was the eye-hand span. This measures how large the preview window of text being copied needs to be in order for the typist to type at full speed. Table 6.1 summarises the results. The span-related hypotheses were partially supported; eye-hand span was larger for words than numbers, but familiarity only had an effect on words, not numbers. The eye-hand span results were different to those found in previous research (Salthouse, 1986), suggesting there was an issue with the current method.

Comparison	**Significance**
Familiar v Non-familiar*	✓
All Words v All Numbers*	✓
Familiar Numbers v Non-familiar Numbers*	✗
Words v Non-words*	✓

Table 6.1: Results of statistical analyses for the eye-hand span measure in (Study 3) Starred conditions had a larger span.

Table 6.2 presents the results from the interkey interval (IKI) measure. This measures the speed of typing. All timing-related hypotheses were supported in this experiment; IKI was significantly different for words and numbers, and was significantly different for familiar and non-familiar numbers.

Comparison	**Significance**
Familiar v Non-familiar*	✓
All Words v All Numbers*	✓
Familiar Numbers v Non-familiar Numbers*	✓
Words v Non-words*	✓

Table 6.2: Results of statistical analyses for the IKI measure in (Study 3) Starred conditions were faster.

There was a limitation with Study 3 in terms of the nature of the non-words used. Pronounceable non-words were used where in previous studies, non-pronounceable non-words were used. This may have accounted for the differences in the results of this experiment compared with previous studies.

Due to this methodological problem, the study was repeated, and the non-words switched to be non-pronounceable (Study 4). Table 6.3 presents the result from the second eye-hand span study. The span-related hypotheses were partially supported. Words and numbers had different eye-hand spans, but no difference was found between familiar and non-familiar numbers. The results of the spans for words and non-words were closer to those found in previous experiments (Salthouse, 1986), suggesting that changing to non-pronounceable non-words made the experiment closer to the original in terms of methods.

Comparison	**Significance**
Familiar v Non-familiar*	✗
All Words v All Numbers*	✓
Familiar Numbers v Non-familiar Numbers	✗
Words v Non-words*	✓

Table 6.3: Results of statistical analyses for the eye-hand span measure in (Study 4) Starred conditions had a larger span.

The results of the IKI analysis (shown in Table 6.4) confirmed previous results and hypotheses. An additional finding was that familiar numbers were not only typed faster than non-familiar numbers, but non-words also.

Comparison	**Significance**
Familiar v Non-familiar*	✓
All Words v All Numbers*	✓
Familiar Numbers v Non-familiar Numbers*	✓
Words v Non-words*	✓

Table 6.4: Results of statistical analyses for the IKI measure in (Study 4) Starred conditions were faster.

The results of these studies suggest that familiar numbers have a stronger representation in long term memory, this is supported by the faster typing speeds reported for familiar numbers compared to non-familiar numbers. However, the lack of significant difference for the eye-hand span measurement does not support this theory.

6.2.4 Study 5: Replacement Span

The replacement span study aimed to test similar hypotheses, that the replacement span would be larger for familiar targets. The replacement span measures the point at which typists notice changes to the characters they are typing. Table 6.5 reports the effects of condition on replacement span. As with the eye-hand span the hypotheses were supported partially, with the replacement span being larger for words than numbers, and larger for words than non-words. But there was no effect of familiar and non-familiar numbers.

Comparison	**Significance**
Familiar v Non-familiar*	✓
All Words v All Numbers*	✗
*Familiar Numbers v Non-familiar Numbers**	✗
Words v Non-words*	✓

Table 6.5: Results of statistical analyses for the replacement span measure in (Study 5) Starred conditions had a larger span.

The results of the effects on IKI can be seen in Table 6.6. These results replicate the previous results, supporting the hypotheses that familiarity affects how a number is represented in memory, which in turn affects the typing speed for numbers.

Comparison	**Significance**
Familiar v Non-familiar*	✓
All Words v All Numbers*	✓
Familiar Numbers v Non-familiar Numbers*	✓
Words v Non-words*	✓

Table 6.6: Results of statistical analyses for the IKI measure in (Study 5) Starred conditions were faster.

6.2.5 Study 6: Copy Span

The final of the Salthouse phenomena-based span experiments tested the hypothesis that the copy span would be larger for both familiar numbers and words. The copy span measures how much of the text a typist is able to recall when the text they are copying from disappears from view. Table 6.7 summarises the results. As in the replacement span experiment, the copy span was not significantly larger for words than number, but was larger for familiar targets than non-familiar targets.

Comparison	**Significance**
Familiar v Non-familiar*	✓
All Words v All Numbers*	✗
*Familiar Numbers v Non-familiar Numbers**	✗
Words v Non-words*	✓

Table 6.7: Results of statistical analyses for the copy span measure in (Study 6) Starred conditions had a larger span.

This experiment replicated the results of previous experiments in this thesis and supported the hypotheses that IKI would be faster for words than numbers, and would also be faster for familiar numbers compared with non-familiar numbers. These results can be seen in Table 6.8.

Comparison	**Significance**
Familiar v Non-familiar*	✓
All Words v All Numbers*	✓
Familiar Numbers v Non-familiar Numbers*	✓
Words v Non-words*	✓

Table 6.8: Results of statistical analyses for the IKI measure in (Study 6) Starred conditions were faster.

The results of these typing experiments supported the hypothesis that familiar numbers would be typed faster than non-familiar numbers. This suggests that familiar numbers have a stronger representation in memory than non-familiar numbers, which facilitates the transcription process when performing a number entry task.

6.2.6 Study 7: Numbers used in infusion pump programming

The results of the Salthouse phenomena-based experiments in Studies 3 to 6 suggest that familiarity can have an effect upon the speed that a number is typed. This results is important because it suggests that familiar numbers are represented more strongly in memory than random numbers. To understand the significance of this finding for research into the design of number entry interfaces, it was necessary to know whether familiarity was an important factor in tasks in applied domains where numbers are entered. This was tested in the medical domain by investigating whether some numbers were used more frequently than others. A log analysis of numbers used when programming infusion pumps was performed and novel patterns of numbers were identified. This result did not replicate any previous research into number distributions. This finding verifies the results of the previous chapter, by demonstrating that number patterns in an applied domain can cause some numbers to be familiar. These number patterns are likely to apply to contexts other than the medical domain, which means the results of the previous chapter may apply to a range of other tasks. This will affect the way that number entry devices in such domains are evaluated in the future, by suggesting that random number entry will not accurately model user behaviour. In addition to this, this study also suggests that this knowledge of the numbers used when programming infusion pumps could be used to inform medical device design.

6.2.7 Study 8: Generation of design heuristics

Based on the results of Study 7, a set of design heuristics were arrived at to guide the design and evaluation of number entry interfaces for infusion pumps. These heuristics allowed the evaluation of existing number entry interfaces using the knowledge gained about the real world numbers. This analysis highlighted that there are no current infusion pump interfaces that are ideally suited to the task of programming common infusions.

6.2.8 Study 9: Tailoring number entry interfaces

This experiment aimed to test whether devices that were adapted to better suit the task of programming infusions by use of the heuristics could reduce the chances of number entry errors. Three interfaces were tested in Study 9 and it was found that tailoring interfaces to match the distributions of numbers that are commonly input was able to increase interaction speed on one interface without causing a higher number of errors. For the other interfaces, although no increase in speed was found, it was shown that the tailored interfaces required fewer keypresses, and with training could also become faster to use. This study highlights how an understanding of the number patterns used in a task can beneficially impact and inform the design of a number entry interface. However, this interface adaptation did not result in a reduction in number entry errors.

6.2.9 Study 10: Checking for number entry errors in the medical domain

The analysis of numbers used to program infusions highlighted a special property: the numbers used to program infusion pump devices contained some redundancy, in that for each infusion, two of the required numbers could be used to calculate the third. By requiring participants to enter all three of the numbers, a natural checksum could be used to ensure that there were no errors during the number entry task. The result of this experiment showed that by using this method, number entry error rates could be significantly reduced. In fact, in the case of this particular experiment, no errors went unnoticed. This increase in accuracy however, came at a cost to the time required to complete the task: because three numbers had to be entered, rather than just two. These results suggest that, by utilising the knowledge of the numbers that are used in a particular context, interface design can be improved by reducing error rates. However, it is not clear currently how this would be implemented, as it is important to balance both increased accuracy with decreased speed.

6.3 Contributions

The findings and results presented in this thesis have implications for both our basic understanding of the cognitive processes involved in the transcription process, as well as providing insights that can inform the design of number entry interfaces. These contributions are explored in more detail in the following section.

6.3.1 Contribution to Transcription Typing Research

Application of alphabetic transcription to numeric transcription

Previous research in the number entry domain has assumed that text transcription metrics (Salthouse, 1986) can be applied to the number transcription task (Lin & Wu, 2011) without adaptation. This is in part due to the lack of relevant numeric transcription typing metrics previously available. It was suggested in Chapter 2 that words and numbers were read using different processes, due to aphasic patients' varying impairments when reading the two types of text (Cipolotti & Butterworth, 1995).

A number of findings within this thesis suggest that a direct application of previous text transcription research to number transcription cannot be made without adjustment. Studies 3-6 consistently showed a significant difference in transcription typing speed between words and numbers. Additionally, the eye-hand span was found to be significantly larger for words than for numbers. When comparing familiar words to familiar numbers, the replacement span and copy span were also larger.

Applying a chunking theory of information processing during a transcription task (Salthouse, 1986; John, 1996), these results suggest that words are processed using larger chunks than numbers. The faster typing speeds for words implies that the chunking strategies when reading text are more efficient for words than numbers. This is supported by the fact that a typist appears to require more information to be available in a preview window when reading words compared with numbers. This suggests that the chunks being formed are larger for words than numbers. The smaller preview windows caused the words to be broken, meaning they could not be stored as a single chunk. This lead to the typists processing them at the syllable level, resulting in a reduction in the amount of information being processed, as shown by the slower typing speed. The typing speed for numbers was not affected until an even smaller preview window size, which suggests that information flow was not affected by the full numbers being broken by the smaller preview window size. The lengths of numbers and words were matched, implying that words are being stored more often as whole-word chunks, and numbers are more often split into smaller chunks. If typing metrics are to be used to inform a model of number transcription, this result suggests that further research is required into specifically understanding the task of number transcription and cannot re-use results from alphabetic text transcription.

The importance of familiarity of number

Previous research on number entry, both in the form of empirical studies of the transcription process (Wiseman et al., 2011; Oladimeji et al., 2011; Lin & Wu, 2011) and computer simulations (Thimbleby & Cairns, 2010; Cauchi et al., 2012) has tended to use randomly generated numbers. The results of this thesis, however, show that this approach is problematic, and that more care should be taken to sample numbers from the task domain of interest.

Previous research with aphasic patents suggested that familiarity of a number can affect how it is read, as the representation in memory facilitates the reading of more familiar numbers

(Cohen et al., 1994; Delazer & Girelli, 1997). This finding highlights that the underlying assumption that transcribing random numbers is satisfactory for simulation of number entry tasks may be incorrect and require further investigation.

The results in this thesis were consistent with these previous findings. The work in Study 1 replicated this outcome, as participants were able to categorise a number as 'familiar' faster than they could categorise it as 'random'. This finding also replicated results from the word reading domain, which has found that words are faster to categorise than non-words (Meyer & Schvaneveldt, 1971). This similarity suggests that the relationship between familiar and non-familiar numbers may be likened to that between words and non-words.

The results of Studies 3-6 also support the theory that familiar and non-familiar numbers are read differently, but in the context of a typing study. In each of the experiments, familiar numbers were significantly faster to transcribe than non-familiar numbers. Again, this mirrors results from alphabetic text transcription as words are typed faster than non-words (Salthouse, 1986).

These results suggest that familiar numbers have representations in long term memory, and that non-familiar numbers do not. These findings suggest that long term memory (LTM) representations of number can facilitate the reading, and subsequently, typing process. This representation in LTM means that during the transcription typing process, any recognisably familiar number can be read as a full chunk, as the 'spelling' of the number can be recalled from memory, rather than having to be memorised during the transcription process. This allows for faster typing speeds.

This result however, was not reflected in the typing span results, as familiarity of a number did not have a significant effect upon the eye-hand, replacement, or copy spans. This may have been as a result of the purity of the familiar and non-familiar number sets; it was likely that not every number within the familiar number set was familiar to each participant, meaning at some points they would have been effectively typing non-familiar numbers. This issue may have diminished the effect of familiarity of number on each of the spans. It may also suggest that the span experiments are not suitable for the task of understanding the effect that familiarity can have upon number transcription.

Familiar numbers and their causes

The previous research, which highlighted the importance of familiarity in number reading, did not elaborate on which numbers exactly were found to be familiar, nor did they exhaustively report the reasons behind that familiarity (Cohen et al., 1994; Delazer & Girelli, 1997), stating only that historical years and car brand names were important.

This thesis has presented a comprehensive list of numbers that have been rated as familiar by a large population of mainly British participants (Study 1), as well as providing a means to generate further lists of familiar numbers (Study 2). The categories generated can provide a means to generate familiar number lists for use in experiments with other participant groups, thus saving future researchers the additional data collection that was required in Study 1.

6.3.2 Contribution to Number Entry Research

Infusion pump evaluation guidelines

The analysis of logs presented in Study 7 has shown that there are clear patterns in the numbers entered when programming infusion pumps; 50% of infusions in the sample set used one of three numbers when determining the rate of infusion. This information is important when testing interfaces for infusion pumps. Rather than modelling the task using random numbers, for stronger ecological validity the distribution of numbers discovered in Study 7 should be used. This approach has since been used in the evaluation of infusion pump interfaces (Cauchi et al., 2013). The new approach allowed for more realistic error rate modelling, and provided a different recommendation for how one feature of the pump should be set-up.

The heuristics generated from the distribution of numbers in Study 8 provide a simple interface for interacting with the real numbers used to program infusion pumps. The heuristics can help to direct the future evaluation of infusion pump interfaces by highlighting key functionality that the interface should allow.

The work in Study 9, which investigated different number entry interfaces, highlights the importance of understanding error causes when evaluating devices. Previous research had suggested that the chevron interface was safest in terms of error rates, as it encouraged error checking behaviour (Oladimeji et al., 2011). However, the results from Study 9 did not support this finding, as there were no clear differences in error rates between the chevron interface, and the other interfaces tested. This discrepancy between results can be attributed to different methodological set-ups (virtual versus physical pump, natural versus artificial errors) and makes clear the importance of considering multiple different experimental approaches before making claims about the benefits of one interface over another.

The implications of understanding the numbers used in the task

In Chapter 2, the current practice in the text entry domain of using context dependent phrase sets was highlighted. These domain-specific phrase sets (such as those for children (Kano et al., 2006), or to test SMS messaging (How & Kan, 2005)) were important to ensure external validity of the evaluation. Equivalent phrase sets for numbers were not available. The work in Study 7 has highlighted that numeric phrase sets may also be necessary when studying number transcription. Just as children's text entry could not be simulated from taking language from financial newspapers (Kano et al., 2006), it has been shown in this thesis that infusion pump programming cannot be simulated by randomly generating numbers.

This thesis has also highlighted the benefits that understanding domain-specific numbers can have upon design. By using knowledge about the task, two different approaches to solving number entry design problems have been used. Firstly, Study 9 showed that existing interfaces can be adapted to make frequent number entry tasks easier by making them faster and reducing keypresses. Although the time increase was not universal across all interfaces, the keypress data suggests that with training this benefit could be seen for all the interfaces.

Secondly, an alternative approach used the knowledge of numbers within the medical domain to create a checksum interface that was capable of preventing errors from being entered into a simulated medical device (Study 10). Without the information about the features of the numbers used in the domain, this design solution could not have been explored.

The interfaces used in the experiments within this thesis were specific to the medical domain, particularly for numbers used to program infusion devices. However, the principles of this approach are applicable to other domains. From this research, it has been shown that an understanding of the numbers used in the task can provide unique design solutions for number entry interfaces.

6.4 Limitations

The work presented in this thesis is intended as an exploration into the effects of number familiarity and frequency on transcription typing in both the lab and the real world. However, necessarily, some aspects of the task could not be accurately simulated and for that reason, the research has limitations.

6.4.1 Ecological validity of the task

One potential issue with regard to the ecological validity of the experiments presented in the Chapter 5 is that number transcription in the real world rarely involves copying from a line of numbers, indeed often only one or two numbers are copied at a time. For instance in the medical domain, the infusion pump device requires a user to enter two values from a prescription. Dialling a telephone number from a separate source only involves entering one number. In domains such as the financial sector, or any job involving data entry, multiple numbers are often entered in succession, however it is likely these are presented in vertical list form.

The experiments in Chapter 4 of this thesis presented numbers in a similar fashion to text: as lines of characters. This decision was made so that comparisons could be made between how words and numbers are typed. It also allowed comparisons with various text transcription studies, which invariably involve copying from a standard prose layout. Laying out the numbers in a vertical manner would have made comparisons difficult to perform. As yet, there is no initial work on how this layout might affect the Salthouse spans, and so there would have been no previous results to compare with.

When considering how this issue could be addressed, it is not clear in an ideal situation what layout should necessarily be used. In an investigation of dosage layouts on prescription charts, it was found that the numbers were not laid out consistently, and that this layout could impact on the procedural errors made (Back et al., 2013). The paper suggests that there is much more work to be done in terms of understanding how number layout can affect the order of steps taken when completing a task. This suggests that in some situations, it is unclear how numbers should be laid out, which makes investigating a generic number presentation difficult.

An additional issue with ecological validity can be found in Studies 9 and 10, where participants were asked to complete medical infusion pump programming tasks despite not having any medical training. This meant that the numbers that the participants were entering were likely to be non-familiar, whereas a medical worker entering the numbers would likely treat them as familiar. The results in this thesis have shown this to be an issue.

The importance of accuracy in the medical domain was also difficult to simulate in laboratory settings. Despite participants being told the importance of striving for high levels of accuracy, the costs of errors in the laboratory were not as high as in a real world medical context. The error rates found in these experiments may therefore be higher than would be expected in a medical setting. However, in a medical setting workers may be required to complete numerous tasks in quick succession. The tasks that participants completed in Studies 9 and 10 were just number entry tasks; there were no distractions for the participants or additional tasks to complete. Again this did not simulate a true medical setting and so may have produced a lower error rate in the experiments than would be expected in a medical domain. It is unclear how this artificiality may have affected error rates, it is possible that both issues worked against each other to result in a representative error rate, however, it is also possible that one could have affected the participants more than the other, resulting in a non-representative error rate.

6.4.2 Appropriateness of spans when applied to numbers

The results of the three span experiments appear contradictory when taken together with the results for the speed of typing. The span measurements suggest that there is no difference between familiar and non-familiar numbers. Whereas the timing data suggests that there is some difference between the two, which affects how they are typed.

The span phenomena were initially introduced purely to measure and understand alphabetic text transcription and have been particularly effective in highlighting the differences between how words and non-words are transcribed. However, as suggested in the previous section, the differences between familiar and non-familiar numbers may not be as great.

For this reason, the traditional spans may not be best suited to elucidating differences in the task of number transcription. This research suggests that the IKI measurement is more sensitive. Alternative tests may also need to take into account the differing ways that numbers can be presented when used in a transcription task.

6.4.3 Defining familiarity

The subjective nature of "familiarity" of numbers, compared to the more binary definition of words and non-words, has imposed limitations in the work of this thesis. In experiments where corpuses of words and numbers were required, the generation of words and non-words was a trivial task. This process is not as easy when applied to numbers.

This meant that a corpus had to be created using a surveying technique. This technique was effective and sets of familiar and non-familiar numbers were collected. However, an issue with this technique is that the numbers used in the experiments could not be guaranteed to be perfectly familiar or non-familiar to participants. It may have been that some participants did not know that 1048 is an exponential of 2 and, therefore, did not consider that number to be familiar. Equally it may have been the case that one participant saw the apparently random number 3552 and recognised it as their PIN. The familiarity of a number cannot be as certain as familiarity with a word, meaning the measurements for familiar numbers may have also included some non-familiar numbers.

Additionally, the use of corpuses of familiar and non-familiar numbers has limited the scope of the experiments that could have been run. Numbers could not be repeated during the experiments, as this would have artificially inflated their familiarity, or may have made non-familiar numbers appear more familiar. For this reason, the length of the typing experiments in Chapter 4 was limited by the amount of numbers in each corpus. With a larger set of numbers, the spans could have been tested more times to ensure more consistency within the results.

6.5 Future Research

As stated previously, the work in this thesis aims to expand the possible avenues of research within the number entry domain, by linking the work to the wider text entry domain and exploring how current number entry research can utilise real world number distributions. There are many routes that future work could take.

6.5.1 Modelling

One clear avenue for future research is to produce a computational model capable of simulating a number transcription typing task. Work is already being completed in this area (Lin & Wu, 2011) but, as discussed previously in the Motivation chapter, the model is in no way comprehensive enough at this point. The model only considers hearing and typing numbers, and does not consider the reading element of the task. Additionally, the model simulates the transcription of random numbers and so does not take into account familiarity. This leaves the model open to improvement.

The creators of the number transcription model have also previously successfully created an alphabetic transcription typing model that simulates the Salthouse typing phenomena. Using this methodology, it is possible to see how a model of number transcription could be created, but more knowledge would be needed with regard to the numeric typing spans. The spans discovered in this thesis are necessary but not sufficient for adapting this model to numeric transcription and further work would require investigation into other numeric typing phenomena, such as information about error rates.

6.5.2 Number Value Representation

Another property of the numbers used when programming infusion pumps is that each number represents a real amount in the world. This could be utilised in design, by giving magnitude feedback to users to help them reflect upon the size of numbers they are entering. Such research has been done previously, but was aimed at helping users to calculate dosage values by using graphs (Gould, Cox, & Brumby, 2013). This graphical aid made participants faster and more accurate when solving calculations.

Similar solutions could be applied to the entry of numbers into interfaces. The aim would be to help users understand the magnitude of the numbers they were entering, in order to prevent errors where the number entered was 10 times larger than intended, a common error as suggested in Thimbleby & Cairns (2010). This might involve abstract solutions such as assigning shapes to numbers, or could be more realistically represented, in the medical domain for instance, as varying bag sizes. These values could be displayed on devices, and would give users more information about the numbers they have entered, as it was shown in Study 10 that users do not always accurately check values when they are represented as numbers.

6.5.3 Number Error Research

This thesis briefly explored the errors made during the studies, but it did not discuss the types of errors that were made when transcribing numbers. One reason for investigating error types is that errors can provide information about what is happening when we transcribe text. In the paper Wiseman et al. (2011), the possible causes of error are elaborated, including errors occurring at a perceptual (if a user commonly types 1 instead of 7 for example we might assume that the perceptual element of transcription is not performing accurately), cognitive (a user reading the number 4224 and misremembering it as 2442) and motor (a user typing 5 instead of 4 due to a slip whilst typing) level.

Future studies could probe into this issue and further our understanding of what types of errors are associated with which part of the typing task, for instance, by isolating elements of the transcription task to discover when certain errors occur. Asking users to enter numbers using a speech-based interface might help us understand which errors are attributed to the input and parsing phase of typing, and not the motor phase for instance.

6.5.4 Domain Specific Number Familiarity

The typing experiments in this thesis used a general definition of "familiarity" that would be applicable to any participants that were recruited for the study. Future work might consider investigating context-sensitive familiarity. For instance, this could involve asking medical workers to classify the numbers gathered from the log analysis (Study 7) as familiar or random, to see if the frequency of number translates to its familiarity.

6.6 Summary

This thesis represents an important contribution, to the theory of both number reading and text transcription, and to practical aspects of number entry interface design. An additional contribution of the work is to situate number entry research as a key research area within the text entry domain. Previous work on number entry has largely approached the problem from a more applied angle, and has not drawn upon the theoretical results found in text transcription research. Equally, text entry research has largely focused on the process of transcribing alphabetic text, and had largely overlooked the task of number entry.

The work in this thesis shows that familiar numbers have a stronger representation in memory, and as such they are faster to type in a numeric transcription task, and can be recognised faster during a decision task. The implication of this is that users type numbers they are familiar with faster than those they are unfamiliar with. This has implications in the real world, where number patterns specific to a domain can result in users becoming more familiar with numbers. In evaluations of number entry interfaces, therefore, it may not be acceptable to test using randomly generated numbers.

Secondly, this thesis has shown how understanding the landscape of numbers used in a particular task can have implications upon the way that a number entry interface is designed. Without this knowledge, generic number entry interfaces are often the only designs considered. However, applying this understanding of number patterns to design can result in interfaces that are faster and more accurate to use.

This thesis argues that these contributions are important, in both the theoretical and applied domain. Future work in the number entry domain will have stronger ecologically validity as a result of findings within this thesis, and new approaches to number entry design can be explored.

References

Adam, R. (2011). NHS Blood and Transplant SUI Summary Report. Tech. Rep. December 2010, NHS.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the Environment in Memory. *Psychological Science*, *2*(6), 396–408.

Back, J., Cox, A., & Brumby, D. (2012). Choosing to Interleave: Human Error and Information Access Cost. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems*, 69, (pp. 1651–1654).

Back, J., Iacovides, I., Furniss, D., Vincent, C., Cox, A. L., & Blandford, A. (2013). Designing Better Prescription Charts: Why we can't just ask the nurses. In *MediCHI-2013, the Workshop on Safer Interaction in Medical Devices, at CHI-2013*.

Baljko, M., & Tam, A. (2006). Indirect text entry using one or two keys. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, (pp. 18–25). ACM.

BBC News (2012a). Baby Maisie Waters died after fed day's food in an hour. http://www.bbc.co.uk/news/uk-wales-south-east-wales-20323497.

BBC News (2012b). Stafford twins' death: Failings in care, says coroner. http://www.bbc.co.uk/news/uk-england-stoke-staffordshire-18172188.

Benford, F. (1938). The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, *78*(4), 551–572.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, *59*(11), 977–984.

Bosman, E. (1993). Age-related differences in the motoric aspects of transcription typing skill. *Psychology and Aging*, *8*(1), 87–102.

Brumby, D. P., Davies, S. C., Janssen, C. P., & Grace, J. J. (2011). Fast or safe? How performance objectives determine modality output choices while interacting on the move. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, (pp. 473–482).

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological review*, *62*(3), 193–217.

Cairns, P., Pandab, P., & Power, C. (2014). The influence of emotion on number entry errors. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, (pp. 2293–2296). New York, NY, USA: ACM.

Card, S. K., Newell, A., & Moran, T. P. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc.

Cauchi, A., Gimblett, A., Thimbleby, H., Curzon, P., & Masci, P. (2012). Safer 5-key number entry user interfaces using Differential Formal Analysis. In *Proceedings of the 2012 BCS Conference on Human-Computer Interaction*, 77, (pp. 29–38).

Cauchi, A., Thimbleby, H., Oladimeji, P., & Harrison, M. (2013). Using Medical Device Logs for Improving Medical Device Design. *2013 IEEE International Conference on Healthcare Informatics*, (pp. 56–65).

Chen, K., Hellerstein, J., & Parikh, T. (2010). Designing adaptive feedback for improving data entry accuracy. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, (pp. 239–248).

Cipolotti, L., & Butterworth, B. (1995). Toward a Multiroute Model of Number Procesing: Impaired Number Transcoding With Preserved Calculation Skills. *Journal of Experimental Psychology: General*, *124*(4), 375–390.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Cohen, L., Dehaene, S., & Verstichel, P. (1994). Number words and number non-words: A case of deep dyslexia extending to arabic numerals. *Brain*, *117*, 267–279.

Cook, R., Nemeth, C., & Dekker, S. (2008). What went wrong at the Beatson Oncology Centre. *Resilience engineering*, *1*, 225–35.

Damasio, A. R. (1992). Aphasia. *New England Journal of Medicine*, *326*(8), 531–539.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*, 1–29.

Delazer, M., & Bartha, L. (2001). Transcoding and calculation in aphasia. *Aphasiology*, *15*(7), 649–679.

Delazer, M., & Girelli, L. (1997). When 'Alfa Romeo' facilitates 164 : Semantic Effects in Verbal Number Production. *Neurocase*, *3*(6), 37–41.

Dunlop, M. D., Komninos, A., & Durga, N. (2014). Towards high quality text entry on smartwatches. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI '14, (pp. 2365–2370). New York, NY, USA: ACM Press.

Dunlop, M. D., & Levine, J. (2012). Multidimensional pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, (pp. 2669–2678). New York, NY, USA: ACM Press.

Dunlop, M. D., & Masters, M. M. (2009). Pickup usability dominates: a brief history of mobile text entry research and adoption. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, *1*(1), 42–59.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. *Journal of Forensic Accounting*, *5*, 17–34.

Dvorak, A., Merrick, N. L., Dealey, W. L., & Ford, G. C. (1936). *Typewriting behavior*. New York: American Book Company.

Gentner, D., & Larochelle, S. (1988). Lexical, sublexical, and peripheral effects in skilled typewriting. *Cognitive Psychology*, *20*(4), 524–548.

Gould, S. J., Cox, A. L., & Brumby, D. P. (2013). Using graphical representations to support the calculation of infusion parameters. In *Human-Computer Interaction–INTERACT 2013*, (pp. 721–728). Springer.

Hennessy, O. (2012). *Can task performance be improved by adapting number entry interfaces to fit the task?*. Master's thesis, University College London.

Hershman, R., & Hillix, W. (1965). Data processing in typing: Typing rate as a function of kind of material and amount exposed. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *7*(5), 483–492.

Holmes, C., Jefferson, P., & Ball, D. (2009). Difficulties implementing the guidelines for the management of severe local anaesthetic toxicity. *Journal of the Association of Anaesthetists of Great Britain and Ireland*, *64*, 573.

How, Y., & Kan, M.-Y. (2005). Optimizing predictive text entry for short message service on mobile phones. In *Proceedings of HCII*, vol. 5.

Inhoff, A. W. (1991). Word frequency during copytyping. *Journal of experimental psychology. Human perception and performance*, *17*(2), 478–87.

Institute for Safe Medication Practices (2006). List of error-prone abbreviations, symbols and dose designations. Tech. rep., Institute for Safe Medication Practices.

Institute for Safe Medication Practices Canada (2007). Fluorouracil Incident Root Cause Analysis. Tech. rep., Institute for Safe Medication Practices Canada.

Isokoski, P., & Linden, T. (2004). Effect of Foreign Language on Text Transcription Performance : Finns Writing English. In *Proceedings of the third Nordic conference on Human-computer interaction*, (pp. 109–112).

John, B. (1996). TYPIST: A theory of performance in skilled typing. *Human-Computer Interaction*, *11*(4), 321–355.

John, B. E. (1993). A quantitative model of expert transcription typing. Tech. Rep. CMU-CS-93-120, Carnegie Mellon University, School of Computer Science.

Jones, E., Alexander, J., Andreou, A., Irani, P., & Subramanian, S. (2010). Gestext: Accelerometer-based gestural text-entry systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, (pp. 2173–2182). New York, NY, USA: ACM.

Kano, A., Read, J. C., & Dix, A. (2006). Children's phrase set for text input method evaluations. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, (pp. 449–452). ACM.

Kim, S., Son, J., Lee, G., Kim, H., & Lee, W. (2013). Tapboard: Making a touch screen keyboard more touchable. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, (pp. 553–562). New York, NY, USA: ACM.

Kohn, L., Corrigan, J., & Donaldson, M. (2000). *To err is human: building a safer health system*. November. The National Academies Press.

Kristensson, P. O. (2009). Five challenges for intelligent text entry methods. *AI Magazine*, *30*(4), 85.

Kristensson, P.-O., & Vertanen, K. (2012). Performance comparisons of phrase sets and presentation styles for text entry evaluations. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, (pp. 29–32). New York, NY, USA: ACM.

Ley, E. (1996). On the Peculiar Distribution of the U.S. Stock Indexes' Digits. *The American Statisticiain*, *50*(4), 311–313.

Li, I., Dey, A., & Forlizzi, J. (2010). A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 557–566). ACM.

Lin, C.-J., & Wu, C. (2011). Factors affecting numerical typing performance of young adults in a hear-and-type task. *Ergonomics*, *54*(12), 1159–74.

Logan, F. (1999). Errors in Copy Typewriting. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1760–1773.

MacKenzie, I. S., & Soukoreff, R. W. (2003). Phrase sets for evaluating text entry techniques. In *CHI '03 extended abstracts on Human factors in computer systems*, CHI '03, (p. 754). New York, NY, USA: ACM Press.

McCurry, J. (2005). Too fat, too fast. The £1.6bn finger. http://www.theguardian.com/business/2005/dec/09/japan.internationalnews.

Meyer, D., & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words : Evidence of a dependence between retrieval operations. *Journal of experimental psychology*, *90*(2), 227–234.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA, USA: Harvard University Press.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, (pp. 249–256). New York, NY, USA: ACM.

Nietert, P. J., Wessell, A. M., Feifer, C., & Ornstein, S. M. (2006). Effect of terminal digit preference on blood pressure measurement and treatment in primary care. *American Journal of Hypertension*, *19*(2), 147–52.

Noel, M.-P., & Seron, X. (1993). Arabic number reading deficit: A single case study or when 236 is read (2306) and judged superior to 1258. *Cognitive Neuropsychology*, *10*(4), 317–339.

NPSA (2010). *Design for patient safety. A guide to the design of electronic infusion devices*. NHS, 1 ed.

O'Carroll, S. (2011). *Mistakes with Numbers: How interface design can influence number entry errors*. Master's thesis, University College London.

Oladimeji, P., Thimbleby, H., & Cox, A. (2011). Number entry interfaces and their effects on error detection. In *Human-Computer Interaction*, INTERACT 2011, (pp. 178–185). Springer.

Oladimeji, P., Thimbleby, H., & Cox, A. L. (2013). A Performance Review of Number Entry Interfaces. In *Human-Computer Interaction*, INTERACT 2013, (pp. 365–382). Springer.

Olsen, K. A. (2008). The $100,000 Keying Error. *Computer*, *41*(4), 108–107.

Paap, K. R., Newsome, S. L., McDonald, J. E., & Schvaneveldt, R. W. (1982). An activation–verification model for letter and word recognition: the word-superiority effect. *Psychological review*, *89*(5), 573–94.

Patterson, K. E., & Marcel, A. J. (1977). Aphasia, dyslexia and the phonological coding of written words. *The Quarterly journal of experimental psychology*, *29*(2), 307–18.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological review*, *111*(1), 159–182.

Rieger, M. (2004). Automatic keypress activation in skilled typing. *Journal of experimental psychology. Human perception and performance*, *30*(3), 555–65.

Rothkopf, E. Z. (1980). Copying Span as a Measure of the Information Burden In Written Language. *Journal of Verbal Learning and Verbal Behavior*, *19*, 562–572.

Rubenstein, H., Lewis, S., & Rubenstein, M. (1971). Homographic Entries in the Internal Lexicon : Effects of Systematicity and Relative Frequency of Meanings. *Journal of Verbal Learning and Verbal Behaviour*, *62*, 57–62.

Rudchenko, D., Paek, T., & Badger, E. (2011). Text text revolution: a game that improves text entry on mobile touchscreen keyboards. In *Pervasive Computing*, (pp. 206–213). Springer.

Salthouse, T. A. (1984). Effects of age and skill in typing. *Journal of Experimental Psychology: General*, *113*(3), 345–71.

Salthouse, T. A. (1985). Anticipatory processing in transcription typing. *The Journal of applied psychology*, *70*(2), 264–71.

Salthouse, T. A. (1986). Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological bulletin*, *99*(3), 303–19.

Salthouse, T. A., & Saults, J. S. (1987). Multiple spans in transcription typing. *The Journal of Applied Psychology*, *72*(2), 187–96.

Sandron, F. (2002). Do populations conform to the law of anomalous numbers? *Population*, *57*(4), 755–761.

Sedley, A., & Müller, H. (2013). Minimizing change aversion for the google drive launch. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, (pp. 2351–2354). New York, NY, USA: ACM.

Smith, R. (2012). Millions of GP prescriptions contain dangerous errors: research. http://www.telegraph.co.uk/health/healthnews/9238524/Millions-of-GP-prescriptions-contain-dangerous-errors-research.html.

Swift, E. J. (1904). The acquisition of skill in type-writing; a contribution to the psychology of learning. *The Psychological Bulletin*, *1*(9), 295–305.

The Health Foundation (2012). Evidence Scan: Reducing prescribing errors. Tech. Rep. March, The Health Foundation, London.

Thimbleby, H. (2008). Ignorance of Interaction Programming is Killing People. *interactions*, *15*(5), 52–57.

Thimbleby, H., & Cairns, P. (2010). Reducing number entry errors: solving a widespread, serious problem. *Journal of the Royal Society, Interface*, *7*(51), 1429–39.

Thioux, M., Pillon, A., Samson, D., de Partz, M.-P., Noël, M.-P., & Seron, X. (1998). The Isolation of Numerals at the Semantic Level. *Neurocase*, *4*(4-5), 371–389.

Truitt, F., Clifton, C., Pollatsek, A., & Rayner, K. (1997). The perceptual span and the eye-hand span in sight reading music. *Visual Cognition*, *4*(2), 143–161.

Vertanen, K., & Kristensson, P. O. (2011). A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, (pp. 295–298). ACM.

Vertanen, K., & Kristensson, P. O. (2014). Complementing Text Entry Evaluations with a Composition Task. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *21*(2).

Vicente, K. J., Kada-Bekhaled, K., Hillel, G., Cassano, A., & Orser, B. A. (2003). Programming errors contribute to death from patient-controlled analgesia: case report and estimate of probability. *Canadian Journal of Anaesthesia*, *50*(4), 328–32.

Wells, F. L. (1916). On the psychomotor mechanisms of typewriting. *The American Journal of Psychology*, *27*, 47–70.

Wen, S. W., Kramer, M. S., Hoey, J., Hanley, J. A., & Usher, R. H. (1993). Terminal digit preference, random error, and bias in routine clinical measurement of blood pressure. *Journal of Clinical Epidemiology*, *46*(10), 1187 – 1193.

Wiseman, S., Brumby, D. P., Cox, A., & Hennessy, O. (2013a). Tailoring number entry interfaces to the task of programming medical infusion pumps. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 57, (pp. 683–687). SAGE Publications.

Wiseman, S., Cairns, P., & Cox, A. (2011). A taxonomy of number entry error. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, (pp. 187–196). British Computer Society.

Wiseman, S., Cox, A. L., & Brumby, D. P. (2013b). Designing Devices With the Task in Mind: Which Numbers Are Really Used in Hospitals? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*(1), 61–74.

Wiseman, S., Cox, A. L., Brumby, D. P., Gould, S. J. J., & O'Carroll, S. (2013c). Using Checksums to Detect Number Entry Error. In *CHI 2013*.

Wobbrock, J. O., Myers, B. A., & Kembel, J. A. (2003). EdgeWrite: A Stylus-Based Text Entry Method Designed for High Accuracy and Stability of Motion. In *Proceedings of the 16th annual ACM symposium on User interface software and technology - UIST '03*, (pp. 61–70). New York, New York, USA: ACM Press.

Wu, C., & Liu, Y. (2008). Queuing network modeling of transcription typing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *15*(1), 6.

Zhai, S., & Kristensson, P.-O. (2003). Shorthand Writing on Stylus Keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 97–104).

Zhai, S., Kristensson, P.-O., Gong, P., & Greiner, M. (2009). Shapewriter on the iPhone: from the laboratory to the real world. *Proc. CHI*, (pp. 2667–2670).

Appendix A

Familiar numbers generated in Study 1

Number	Total Ratings	Familiarity Ratio
1990	17	1.00
123	8	1.00
1984	7	1.00
888	5	1.00
1800	4	1.00
1850	4	1.00
1877	4	1.00
1962	4	1.00
12345	4	1.00
2011	29	0.97
100	56	0.95
365	18	0.94
2015	18	0.94
2009	17	0.94
1991	16	0.94
911	47	0.94
1992	14	0.93
2000	27	0.93
2012	35	0.91
2013	46	0.91
1918	23	0.91
2001	33	0.91
999	50	0.90
101	30	0.90
1996	10	0.90
1	29	0.90
24	19	0.89
2003	19	0.89
2020	18	0.89
666	60	0.88
2016	17	0.88

Table A.1: All familiar numbers generated in Study 1. Familiar numbers were those rated 4 or more times, with a familiarity ratio of 0.75 or above.

Number	Total Ratings	Familiarity Ratio
0	33	0.88
3	24	0.88
1994	16	0.88
10	29	0.86
1024	29	0.86
1998	36	0.86
2002	14	0.86
2010	14	0.86
808	7	0.86
1914	27	0.85
2006	20	0.85
16	26	0.85
21	26	0.85
1985	13	0.85
7	58	0.84
2014	19	0.84
25	25	0.84
1945	54	0.83
5	6	0.83
50	6	0.83
900	6	0.83
800	11	0.82
1999	27	0.81
404	21	0.81
2007	26	0.81
42	40	0.80
1993	15	0.80
2004	15	0.80
128	10	0.80
1000	5	0.80
1901	5	0.80
1986	5	0.80
99	29	0.79
200	14	0.79
44	9	0.78
555	13	0.77
30	17	0.76
1066	62	0.76
64	40	0.75
1969	12	0.75
10000	8	0.75
118118	8	0.75

Table A.2: All familiar numbers generated in Study 1. Familiar numbers were those rated 4 or more times, with a familiarity ratio of 0.75 or above.

Appendix B

Recruitment strategies for Study 3

B.1 Introduction

During Study 3, new methods of recruitment were explored to investigate the possibility of running HCI experiments online. The approach will be discussed here, and the results of the studies will be compared. It was found that the results gathered from each of the different recruitment conditions did not vary significantly, and so it was possible to compare the results across conditions within the analysis of Study 3.

The first method used was the Citizen Science approach. This approach taps into the publics desire to help in scientific research, and there are many projects making use of this crowd sourcing (see (Bonney, Cooper, Dickinson, Kelling, Phillips, Rosenberg, & Shirk, 2009) for a review). The tasks that volunteers are given to complete often involve using them as data gatherers, or processors, rather than as the data themselves, as is required in many HCI experiments.

The second recruitment method used was the Quantified Self approach. Unlike the citizen science approach, the personal informatics, or "Quantified Self" movement is however concerned with data about the human. This field is interested in the way that people log data about themselves, from simple fitness apps, to taking minute by minute photographs (see (Li, Dey, & Forlizzi, 2010) for a review). People are getting more interested in gathering data about themselves in order to find out more about themselves and allow for more informed reflection. This form of participant incentive has been used in HCI research, (see the Lab

In The Wild[1] and Test My Brain[2] projects) which host a series of experiments online that participants can take part in, in exchange for finding out new information about themselves.

These methods were used in conjunction with the standard approach to experiment recruiting, where participants were paid for their time to either come to the lab, or complete the study online via the University subject pool site.

B.2 Method

Table B.1 summarises the recruitment methods for the four conditions. People who took part in the Citizen Science condition were asked if they could take part in an experiment, and were told they would receive no monetary reimbursement. Participants in the Quantified Self condition were provided with an analysis of their performance in the experiment, including their raw data and graphs, which were generated online immediately after they had completed the experiment.

Condition	**Participants**	**Mean Age (SD)**	**Reimbursement Type**	**Location**
Lab Reimbursed	14	29.4 (8.1)	£3	Lab
Online Reimbursed	14	24.4 (9.8)	£3	Online
Citizen Science	44	24 (6.7)	No reimbursement	Online
Quantified Self	45	24.5 (10.6)	Personal typing data	Online

Table B.1: Recruitment methods used in Study 3

B.3 Results

The results of the experiment in each condition were analysed using an ANOVA with a significance level of .05. Statistical analyses were performed to test the effect of target type (words × numbers) and preview window size on typing speed, measured via interkey interval (IKI).

Lab Reimbursed

Here, the results of the lab reimbursed study are reported as a base line. The further experiments will be compared to the results presented here. The analysed data comprised only of words and numbers that were typed correctly.

[1] www.labinthewild.org

[2] www.testmybrain.org

The mean length of time taken per character for numbers was 403.55 ms, and for words was 356.32 ms. An ANOVA reveals that this difference was significant, $F(1,12) = 4.9474$, $p < .05$.

The effect of preview size was also found to be significant ($F(7,84) = 167.69$, $p < .001$). The larger the preview size, the faster participants were able to type. There were no significant effects of interactions between conditions.

Online Reimbursed

The average keypress length per character for numbers was 442.67 ms. The average keypress length per character for words was 396.75ms.

The effect of preview size was found to be significant ($F(7,84) = 83.034$, $p < .001$) which replicated the finding in the lab reimbursed condition. However, there was no significant effect of target typed on typing speed ($F(1,12) = 2.2096$, $p = 0.1629$). This partially replicated the results in the first experiment.

Citizen Science

Numbers took on average 375.65 ms per keypress and numbers took on average 295.30 ms.

There were two main effects on typing speed found in the study: that of target type ($F(1,40) = 49.83$, $p < 0.001$) and preview size ($F(7,280) = 491.5$, $p < .001$). This replicates the findings of experiment 1.

Quantified Self

The average speed per key press per character for numbers was 347.83 ms and for words was 295.71 ms.

There was a significant effect of target type on typing speed ($F(1,38) = 31.346$, $p < .001$) and the preview size also had a significant effect ($F(7,266) = 296.37$, $p < .001$).

B.4 Discussion

The timing data from all conditions were comparable, the average IKI for numbers was 392.43 ($SD = 40.49$) and for words it was 336.02 ($SD = 49.61$). Words were consistently typed faster than numbers, this difference was significant in all conditions except for online reimbursed.

In each condition, preview size had a significant affect upon typing speed, with smaller preview window sizes resulting in slower typing speeds

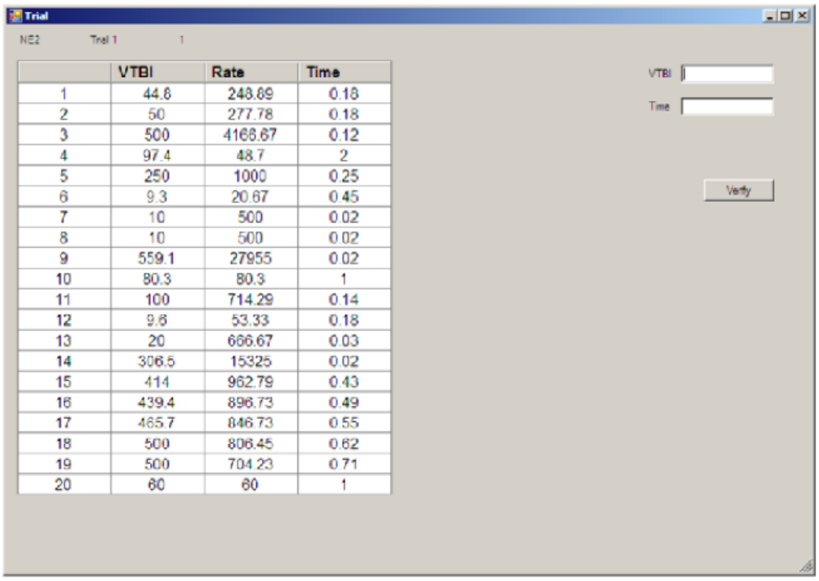
B.5 Conclusion

The results throughout each study were similar, with preview size affecting typing speed and words consistently being typed faster than numbers. In three cases this comparison was significant. In the online reimbursed condition, this difference was not significant, but the results did not contradict those found in other conditions

For these reasons, it was decided that the results in each of the conditions could be grouped together and analysed as a singular data set.

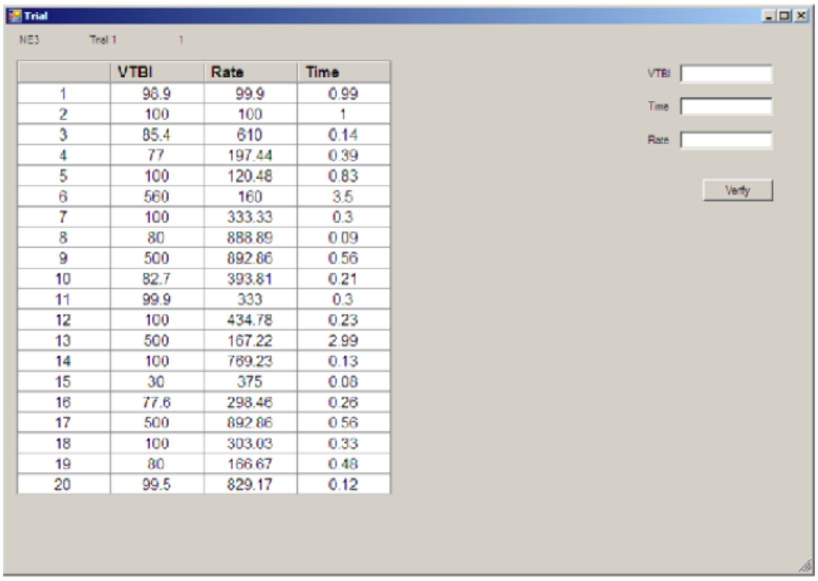
Appendix C

The two interfaces used in Study 10



	VTBI	Rate	Time
1	44.8	248.89	0.18
2	50	277.78	0.18
3	500	4166.67	0.12
4	97.4	48.7	2
5	250	1000	0.25
6	9.3	20.67	0.45
7	10	500	0.02
8	10	500	0.02
9	559.1	27955	0.02
10	80.3	80.3	1
11	100	714.29	0.14
12	9.6	53.33	0.18
13	20	666.67	0.03
14	306.5	15325	0.02
15	414	962.79	0.43
16	439.4	896.73	0.49
17	465.7	846.73	0.55
18	500	806.45	0.62
19	500	704.23	0.71
20	60	60	1

Figure C.1: Screenshot from the 2-number interface used in Study 10



	VTBI	Rate	Time
1	98.9	99.9	0.99
2	100	100	1
3	85.4	610	0.14
4	77	197.44	0.39
5	100	120.48	0.83
6	560	160	3.5
7	100	333.33	0.3
8	80	888.89	0.09
9	500	892.86	0.56
10	82.7	393.81	0.21
11	99.9	333	0.3
12	100	434.78	0.23
13	500	167.22	2.99
14	100	769.23	0.13
15	30	375	0.08
16	77.6	298.46	0.26
17	500	892.86	0.56
18	100	303.03	0.33
19	80	166.67	0.48
20	99.5	829.17	0.12

Figure C.2: Screenshot from the 3-number interface used in Study 10