The Effect of Interface Elements on Transcription Tasks to Reduce Number-Entry Errors

Frank Soboczenski

Doctor of Philosophy

University of York
Computer Science

June 2014

Abstract

Many tasks in daily life require transcribing information accurately from one medium to the other. However, humans make errors frequently. While most of the errors in daily life are little more than an inconvenience in safety-critical domains, such as healthcare, a small error like typing the wrong number when programming a medical device can have grave consequences. Despite potentially fatal consequences of errors, little is known about the errors people make using medical devices, such as infusion pumps, and how the devices themselves influence the errors made. This thesis reports ten studies looking at different interface design features in the context of medical devices and their potential influence on reducing errors.

The first three studies empirically evaluated particular design features proposed by previous work. These studies did not produce the predicted error reduction although there was a recurrent low rate of errors in all the studies. Therefore such claims made in earlier work could not be supported. The studies however showed that it is hard to see the impact of particular features due to the nature of the research area and due to the robustness of the evaluated interface. Additionally, interesting insights into how people use such interfaces to enter numbers need to be taken into account.

Inspired by results from cognitive psychology, which suggest that representing information in a poorer quality format increases the likelihood of memorising the information more accurately, a further set of seven experiments are presented in this work evaluating the effect of such an approach on number transcription tasks. Results showed that people made significantly fewer errors when transcribing less visible numbers as well as text. Furthermore, the studies also confirmed that the source displaying the number and not the entry display is responsible for this counter-intuitive approach to reduce errors. This aligns with previous work in human-computer interaction and psychology. Moreover, the robustness of the discovered effect was investigated once levels of audio distractions were present. Even in a distracting environment the effect led to a significant decrease in errors.

The potential impact of this work could be a valuable contribution for domains where accuracy is of great importance.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deepest appreciation to all those wonderful people who provided me with the possibility to complete this work and I sincerely hope that I did not forget anyone to mention here.

First and foremost, I would like to express my immense gratitude to a fantastic person and a truly brilliant mind: my supervisor, Dr. Paul Cairns. Thank you Paul for all your help, guidance and support that you offered me throughout as well as beyond my studies. Paul, thank you for your trust and patience which made this work possible. I could not have reached this stage without your inspiration, your brilliance and your enthusiasm. Thank you for believing in me and for believing that I could achieve this. I am eternally grateful for your thoughts and for everything you taught me. You are a fantastic teacher, supervisor and mentor. From the bottom of my heart: Thank you so very much Paul!

I would also like to thank my second supervisor Professor Richard Paige who always had an open door for my requests, who often provided me with a most valuable second view on my work and for giving me constructive feedback on all my strange questions. Thank you Richard!

I would also like to express my gratitude to Professor Ana Cavalcanti for her support in making it possible to work with such fantastic people. I am deeply grateful for that, Ana and I consider myself privileged having had the chance to work with such brilliant minds.

I am immensely grateful to Dr. Anna Cox for taking the time to look at my work, for her assistance and for the fantastic collaboration between us. Thank you so much Anna! I would also like to thank Dr. Iain Bate for his support, advice and for the constructive feedback he gave me. Thank you Iain, completing this work would not have been possible without your support. Additionally, I'd like to thank Dr. Lilian Blot, Dr. Silke Goebel, Dr. Jan Boehnke, Dr. Leandro Indrusiak, Dr. Jenny Roe, Dr. Frank Zeyda for all the fruitful research discussions and for all the fantastic feedback!

I would also like to thank Professor Harold Thimbleby and the whole *CHI+MED* team for their support, feedback and for giving me the chance to present and discuss my work with so many of you! I would like to thank in particular, Patrick Oladimeji, Sarah Wiseman, Andy Gimblett, Jonathan Day, Dr. Paolo Masci, Dr. Paul Curzon, Dr. Jo Iacovides, Dr. Dominic Furniss, Dr. Jonathan Back, Dr. Huawei Tu and Abigail Cauchi for all the countless discussions and the feedback on your and my work. In addition, I would like to thank you Professor Daniel Kahneman and Dr. Daniel Oppenheimer. You inspired me to take this direction for my research.

It is also important for me to express my gratitude to my fellow research students many of whom I am very happy to call my friends. Thank you Imran Nordin, Matthew Hudson, Joe Pugh, Sorachai Kamollimsakul, Matt Kamal, Burcu Can, Malihe Tabatabaie, Thomas Lampert, Andre Pimenta Freire, Michael Banks, Nazatul Naquiah, Rasha Salah El-Din, Dina Mostafa, Ipek Caliskanelli, Judith Borghouts, Yvonne Eichhorn, Chris Poskitt, James Williams, Aniket Datta and Simon Greiner. I enjoyed it very much when we discussed our research as well as our successes and our utter failures with each other. Thank you for all your help in finding participants and constructively critiquing my work!

I would also like to thank the entire Department of Computer Science for the support that made my research possible. This includes, but is not restricted to: Chrysle Hudson, Pauline Greenhough and Marysia Koc for their administrative support, the entire hardware and software support staff for their hard work. You guys are fantastic! Thank you Alex Cooper for all the technical support. Thank you Swaroopa Mucheli for always having an open door for my crazy ideas to build strange devices. I am also grateful to all the numerous people who took part in my studies and provided valuable feedback as well as suggestions.

My gratitude belongs to my dear friends Neil Ge Zhengbo, Pedro Ribeiro, Nikki Chen, Jade Davis, Vipata Rujtikumporn, Amy Gao, Saranya Chensukitphanit, Luis Carlos Rodrigues, Siavash Mishamadi, Shayan Mortazavi, Theodora Lee, Narisa Jitpraneechai, George Margoutidis, Wataru Yamaguchi, Mustafa Aydin, William Tsu and Margaret Chen, Takako Nonaka, Anocha Rugchatjaroen and so many others, who I could not list here for always being there helping and supporting me in so many ways and so many many times, even without noticing. You are all so wonderful people!

Additionally, I would also like to thank a very special couple, Kurt and Ann Strauss, who I had the extraordinary privilege and opportunity to get to know during my time in York and who will always have a special place in my heart. Thank you so much for being there during the “ups” and “downs” of my studies. Moreover, I am very grateful to Dr. Steve Kelly, for having the time to discuss my research. Thank you, Steve for all your thoughts and great advice.

It is also very important to me to express my deepest thanks to a very dear and special friend and his wife: Richard and Rosa Ribeiro. Thank you both for every moment I could spend with you, for being there when times where extremely difficult and for always having an open door. Richard, this work would not have been possible without you! I can not express how much you helped and how grateful I am for what you have done. Thank you so much my dear friend!

I would also like to thank many dear friends who supported me from the distance, especially those who passed away before I had the chance to see them again. You'll be missed.

I must express my deepest gratitude to my family, who provided me with invaluable support throughout the years. Especially, my dearest brother Andre Soboczenski who always listened and supported me in so many ways as well as my uncle Kurt and my aunt Martina who encouraged me to pursue this PhD.

Additionally, I wanted to thank the *"Great Bird of the Galaxy"*: Gene Roddenberry, who's vision inspired me throughout my life and who is ultimately responsible for sending me on this path. Thank you Gene!

At last but not least I want to express my deepest gratitude to my wonderful love Victoria Han, who always stood by me and provided me with the support and strength during the most difficult times of my studies.

Declaration

I declare that the work presented in this thesis, if not otherwise stated, is my own. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. Some of the material presented within this thesis has previously been published in the following journals, conference papers and posters:

1. Soboczenski, F., Hudson M. & Cairns P., (2015). The Effects of Perceptual Interference on Number-Entry Errors. Interacting with Computers (IwC), *Accepted and to appear in IwC Special Issue.*

2. Borghouts, J., Soboczenski, F., Cairns P. & Brumby, D., (2015). Visualising Magnitude: Graphical Number Representations Help Users Detect Large Number Entry Errors. Accepted and to appear in proceedings of the Human Factors and Ergonomics Society (HFES) 2015 International Annual Meeting, Los Angeles.

3. Soboczenski, F., Hudson M. & Cairns P., (2014). The Effects of Perceptual Interference on Number-Entry Errors. Proceedings of the first Int. conf. on Interaction Design and Human Factors, Kochi Japan.

4. Soboczenski, F., Cairns, P., & Cox, A. L. (2013). Increasing Accuracy by Decreasing Presentation Quality in Transcription Tasks. In Human-Computer Interaction Proc. Interact 2013, Springer LNCS 8118 (pp. 380-394).

5. Soboczenski, F. (2013). Reducing Number-Entry Errors in Healthcare. Poster presented at the BCS Health Informatics Conference Scotland 2013.

6. Soboczenski, F., Datta, A. & Cairns, P. (2013). Reducing Number-Entry Errors in Medical Systems: A Tale of Two Studies. In proceedings of the 6th York Doctoral Symposium on Computer Science and Electronics 2013 (p. 94).

Ethics

All experimental studies, the design of the experiments and the tasks created for participants followed the ethical principles of *"Do No Harm"*, *"Data Confidentiality"* & *"Informed Consent"*.

Do No Harm:
None of the participants of the individual studies were put in any harmful situations. All experimental tasks were designed in such way so that participants would not be subjected to any risk. Participants had the right to abandon the study at any time in case they felt threatened in any way.

Data Confidentiality:
All data was stored on password protected systems; in reporting results no participant could be identified individually as all data was reported in aggregate.

Informed Consent:
All participants recruited for the experiments were informed about the specific experimental design and tasks that they were going to undertake. Participants were well informed with the appropriate information before a study commenced with a briefing session and after a study with a debriefing session. After the individual experimental briefing they were requested to sign a consent statement, which is available in the Appendix.

In the course of the conduct of this thesis the University governance of ethics has developed. Originally, the studies in this thesis were considered ethical because of the general clearance by the HCI group to understand that they would only raise ethical concerns when they felt that there was an ethical risk. The studies presented here were judged by the supervisor as the authority behind the ethical conduct of this thesis.

To my grandmother Hermine Spranger

&

my aunt Thea Zimmer

Chapter 1

Introduction

"To kill an error is as good a service as, and sometimes even better than, the establishing of a new truth or fact."
Charles Darwin

1.1 Background

People commit errors in many situations in daily life. Errors are unavoidable and can occur during everyday activities or during the most complicated tasks. While in some cases errors are not of much concern, errors in safety-critical systems, such as aircraft, nuclear power plants or hospitals can have catastrophic results. Errors can occur in several forms such as slips, lapses, violations or mistakes (Reason, 1990). Previous research has shown that human errors are the source for more than 50% of incidents in most domains, and in aviation the percentage was even higher (90%) (Zhang, Shortliffe, Patel, Freed, & Remington, 2000) a number which has not changed much in recent years (Thimbleby & Cairns, 2010). Unfortunately, we often only pay great attention to errors when human life is threatened, when disaster is imminent or after serious incidents. As a result of all the accidents and near misses (many of which were reported by the media and the research community), many safety precautions, mechanisms and procedures have been established to prevent future adverse effects. Some examples for such safety precautions are standards such as the *NASA Human System Standard*, which addresses human factors in space flight (NASA, 2011). Another example is the *NHS Saving Lives in Surgery Guide*, which implements surgical safety check-lists to prevent harmful events to patients during treatment (NHS, 2009). In aviation, it is known that pilots are committing errors every day and some of the errors are even critical (FAA, 2013). Their intense training, known as crew resource management (CRM), and continuously updated procedures are part of the reasons that they are able to address these issues in time so that subsequent disaster can be avoided. Such actions, to ameliorate errors, evolved from the fact that aviation safety is a well established field where human factors are the centre of attention (Dekker, 2006). Some cases go back as far as World War II, where Fitts and Jones (1947) evaluated cockpit features. They found, for example, that pilots confused flap and gear handles on multiple occasions as both handles looked and felt the same. In modern days, many items regarding safety, such as check-lists, have been adopted in other domains including healthcare (NHS, 2009; Gawande, 2011). But what about the mundane errors? What about the small tasks such as entering numbers into digital devices? What about the seemingly simple tasks that apparently require no intense training? People still make errors despite their training and those errors which are not intended are still committed. An interesting question is what can be done to make, for example, number-entry safer? In safety-critical domains such as aviation and healthcare a wrong number can lead to severe consequences for peoples' health e.g. when medical staff have to program drug infusion pumps. Mistyping a single digit while transcribing the volume to be infused (VTBI) from the drug prescription can quickly turn a routine task into a serious situation. Research surrounding the body of human error has grown substantially over the last decades (Norman, 1980, 1983; Reason, 1990; Hollnagel & Amalberti, 2001; Reason, 2008) but the field around number-entry errors specifically focused on healthcare systems has only developed in recent years. In order to address these gaps in research, we need to build a better understanding of human errors and their causes as well as empirically evaluate factors such as design elements for interfaces which could help reduce errors.

1.2 Motivation

Healthcare is a domain where a small error can often quickly lead to fatal consequences, as many published cases show (Vicente, Kada-Bekhaled, Hillel, Cassano, & Orser, 2003; ISMP, 2007). Yet it embodies many mundane tasks such as transcribing information from one medium to another. For example, when nurses or doctors enter numbers in infusion pumps, a single incorrect digit can have serious effects on the well-being of patients (Grissinger, 2011). Despite the potentially fatal consequences of errors, little is known about the errors people make while using these devices and how the devices themselves influence the errors made. Previous work has shown that devices do surprisingly little to help prevent or mitigate errors (Thimbleby & Cairns, 2010). What is known is that misprogrammed infusion pumps are the second most frequent cause of medication errors, making it imperative to eliminate as many errors as possible (Smetzer & Cohen, 2006). Moreover, according to Lesar (2002), there is a gap in strategies to address these kind of errors. Additionally, work by Obradovich and Woods (1996) presented that medical devices do suffer from design issues. There is a need to help reduce errors, and indeed there have been efforts to classify the different errors made while using number-entry devices (Wiseman, Cairns, & Cox, 2011). Researchers started to look at specific design features for such device interfaces to evaluate their influence on errors. For example, Cauchi, Gimblett, Thimbleby, Curzon, and Masci (2012) focused on a formal mathematical approach to evaluate different design features for number-entry devices and their effect on errors. They proposed that, analytically, some features have greater influence on error reduction than others. However, there is a call for studying the impact of such features on an empirical basis. This thesis addresses this gap and presents empirical studies on specific design features for healthcare number-entry interfaces.

1.3 Research Approach

This thesis is heavily motivated by safety-critical domains such as healthcare and aviation. The long-term aim of this research is to make transcription tasks in safety-critical domains i.e. number-entry in healthcare more reliable. The goal of the thesis is to better understand how number-entry processes work and how errors might therefore be mitigated or prevented when interacting with technology.

The research body of number-entry in the context of medical devices is a relatively recent development in human-computer interaction. Whilst there is a great amount of previous work in psychology on how people process numbers, studies in this area focus on how humans perceive numbers and whether they have an area in the brain that is responsible for number processing (McCloskey, Caramazza, & Basili, 1985). Others try to understand how people make sense of numbers, such as Dehaene (1997) who state that people store numbers in three different ways in their mind (verbally, as an Arabic notation or as a quantity representation). Additionally, researchers looked at the number length and its influence on memory (Miller, 1956) or simply looked at what brain areas are stimulated when people process numbers (Goebel & Rushworth, 2004). In the domain of Human-Computer Interaction (HCI), research has focused amongst other factors on how the design of number-entry devices can affect errors (Blandford, 2010; Furniss, Blandford, Rajkomar, Vincent, & Mayer, 2011; Vincent & Blandford, 2011). To this end studies have looked at design elements for medical devices such as interface layout or key arrangements, as well as key sizes and their influence on reducing errors (Colle & Hiszem, 2004; Lee & Zhai, 2009). Others have investigated different types of interfaces such as serial, i.e. calculator based compared to chevron or dial interfaces (Oladimeji, Thimbleby, & Cox, 2013). Chapter Two will describe the different interfaces in detail. The initial approach for this thesis was to look at design features for infusion pumps, which might have a significant impact on reducing errors. This led to the first research question, which was motivated by a view on a specific set of interface design features proposed by earlier work of Cauchi et al. (2012). To this end, studies were conducted which look at what are the errors that happen when people enter numbers. The first research question is:

(RQ1) ***What design features can be used to ameliorate the effects of human error in entering numbers into digital healthcare devices?***

The studies which address this research question are described in Chapter Three. Despite the claims by other researchers no significant result in reducing errors could be detected by the evaluated design features bringing such claims into question. As none of the studies showed a significant effect in the evaluated design features and, in light of the limitations of device specific features on reducing errors, a further review of the literature was conducted. It was at this time, that a theory from psychology was taken into account for further research. Therefore, it was decided to move away from a rather technical to a theoretical approach to reduce errors in a medical device context. The discovered theory describes peoples' mind as being separated into two different systems or in other words into two different ways of thinking. Psychologists describe them as *System 1 & System 2* (Kahneman, 2011). The former system epitomises the fast, effortless and automatic thinking whereas the latter describes slow and effortful thinking. A further and more detailed description of these systems and the theory surrounding them is given in the literature review (see Chapter Two). Further inspiration came from results of yet another very interesting experiment by Diemand-Yauman, Oppenheimer, and Vaughan (2011), which led to the idea that indeed there could be a way of directly stimulating *System 2* when people enter numbers. In their work Diemand-Yauman et al. (2011) show that people who read information in a less readable form can memorise this information better than people who read the same information in plain text. This effect, in line with the theory from Stanovich (1999), Stanovich and West (2002), Kahneman (2008) and Kahneman (2011), is what is defined in the context of this thesis as the ***Diemand-Yauman, Oppenheimer and Vaughan - effect (DOV)***. A detailed discussion surrounding the *DOV* effect is given in the literature review (see Chapter Two). The body of research around the *DOV* effect has motivated the following additional three research questions:

(RQ2) ***Can the DOV be used to ameliorate the effects of human error by reducing the error-rate in transcription tasks?***

(RQ3) ***How does the DOV work with a view of moving to real-world applications?***

(RQ4) ***What are the implications of the DOV effect in the context of transcription tasks?***

This thesis will present a series of studies investigating the dynamics surrounding the *DOV* effect, which has never been done before in an interactive context. Chapter Four describes studies that discuss core factors surrounding the *DOV* effect such as how can the effect be stimulated in transcription tasks. Chapter Five focuses on studies, which go one step further by looking at the robustness of the *DOV* effect and how this effect could behave in a real-world environment related to attention and distraction. A high-level review of the research questions will be presented at the end of the literature review in Chapter Two.

1.4 Methodology

1.4.1 Data Collection Methods

The research was primarily based on a series of quantitative controlled laboratory studies in order to address the research questions as this thesis aimed to empirically evaluate interface design features. The collected quantitative data was initially stored on hard copies and in later experiments in form of text log files or MySQL databases. The data was then compiled into a Microsoft (MS) Excel sheet and imported into RStudio for the statistical analysis. RStudio was the primary tool which was used for all statistical analyses. The complete anonymised data sets for all studies are available on the following website:

`http://www.cs.york.ac.uk/~feynman/thesis.htm`.

The research approach that was chosen here was not to establish ecological validity but to use controlled laboratory studies to see whether it was possible to get clear answers in *principle* whether the studied effects were able to be used.

1.4.2 Research Instruments

A number of simulations of interfaces designed after real-world infusion pumps (see Chapter Three, Chapter Four and Chapter Five) were used as research instruments to collect the data in the individual studies. Notably, the evaluated interfaces are very common in multiple domains not just healthcare. The simulations seem to be the optimal approach as actual infusion pump devices were not available and as previously mentioned the aim was not to establish ecological validity but to investigate and to test the principles surrounding the *DOV* effect. These interfaces were programmed in JavaScript, PHP or Objective-C and used a text log or a MySQL database running in the background to store the data. It can be argued that by creating these simulations of real-world device interfaces enabled the possibility of a “pure view” i.e. isolating the number-entry task rather than having other factors such as turn-on the device or steps to prepare the device for number-entry potentially influencing the results.

1.4.3 Demographic Questionnaires

Throughout the studies a series of qualitative questions where used to collect demographic data. These questions were presented on paper and, in later experiments, as an on-line Google form. The structure of the questions were 'yes', 'no' or multiple choice questions (see Figure 1.1 and 1.2). In later experiments a set of five-point likert scale questions was attached to measure the emotional state of participants (see Figure 1.3). This set of questions is part of the Positive and Negative Affect Schedule (PANAS) questionnaire (Watson, Clark, & Tellegen, 1988). The full set of the questionnaires can be found in the individual appendices.

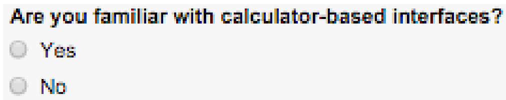


Figure 1.1: An example of a yes or no question

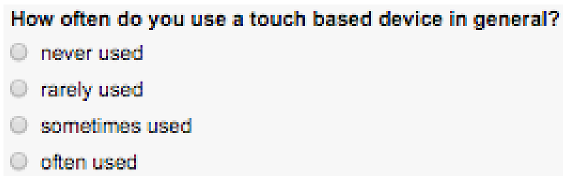


Figure 1.2: An example of a multiple choice question

Figure 1.3: An example question to measure the emotional state of participants

1.4.4 Pilot Studies

All studies were piloted to ensure that there were no procedural errors or flaws in the intended experimental design. In some cases the amount of numbers or text that participants had to enter in the given time-frame was too short which required modifying the available time (see Studies One, Two and Four). Other cases revealed some flaws in the research instruments which had to be corrected. For example, there was an error in the programming code of the interface simulation provided for the first study. Without exception, no participant of a pilot study shown in this thesis was allowed to take part in the following main experiment.

1.4.5 Systematic Errors

One option to discuss and interpret the data in experiments is to evaluate the errors against their systematicity (Byrne & Bovair, 1997). Systematicity is the error rate or as Li, Blandford, Cairns, and Young (2008) (p. 319) mentioned: *"the ratio of number of occurrences to the number of opportunities for that error."* The idea of all the experiments is to achieve systematicity. Researchers have different views on systematicity as for example Li et al. (2008) suggest only to analyse errors which achieved a systematicity of at least 5%. However, all the experiments reported in this thesis show a very low but consistent error rate which therefore suggests that this is the systematicity of this domain.

1.4.6 Statistical Analysis

This thesis presents experiments that look at error data. Error data does not fit or follow a normal distribution and is therefore non-parametric. It is empirically founded that non-parametric error data tends to be strongly positively skewed and more accurately follows an exponential or Poisson distribution. Therefore, the analyses in this thesis use primarily non-parametric tests (*Mann-Whitney*, *Wilcoxon* and *Kruskal-Wallis*).

Some of the experiments (Study Three, Seven, Eight, Nine and Ten) are two-way designs and therefore an Analysis of Variance (ANOVA) is used in this cases. ANOVA is the dominant and well understood analysis technique to study interaction. However, ANOVA is a parametric technique which can be seen as inappropriate for the observed distributions leading to inaccurate p-values. For this reason all main effects of an ANOVA are confirmed with a non-parametric version and interaction effects are confirmed with the non-parametric *Aligned Rank Transformation* (ART) test (Sawilowsky, 1990). The ART is non-standard and can be complex to interpret therefore it was decided not to report it as primary test.

1.5 Thesis Scope

Transcription tasks such as when people enter numbers are an essential part when interacting with technology and are also a critical element in domains where accuracy is of great importance. The scope of this work is primarily motivated around such number-entry processes in context of the interaction with infusion pump devices. This thesis is motivated by such devices used in the healthcare domain and did not look at devices used in other safety-critical domains such as aviation or nuclear power plants.

However, as the work developed it became clear that the findings of the studies presented here are potentially far more relevant than initially expected. While the initial motivation and focus lie on number transcription in medical device interfaces, later studies showed a novel approach which has the potential to reduce transcription errors in multiple domains or maybe even in any transcription task regardless of a domain.

The caveat is that the studies presented in this thesis are all laboratory experiments in a controlled environment not actual field studies or hospital trials. Whilst several promising results of the studies suggest future field studies, it was too early to conduct them at this stage where the importance of the underlying principles and the dynamics of the discovered effect to reduce errors was unknown. All studies used simulated medical device interfaces designed after real-world devices following the primary motivation. The use of actual devices would be a move towards field studies but is also out of scope of this thesis.

It should be mentioned that it is not the purpose of this work to propose a unique all-in-one design for number-entry interfaces. Additionally, this thesis is also not focused on challenging traditional design principles but to empirically evaluate both the design features proposed by previous work and the introduced design features which lead to a novel approach on reducing errors in number-entry.

Chapter 2

Literature Review

"Errare humanum est..." (Engl."To err is human...")

Cicero, I^{st} century BC

"...to understand the reasons why humans err is science."

(Hollnagel, 1993)

The literature review will first direct the attention to the field of *human error*, the history surrounding it, different classifications of human error as well as current research in the field. This is followed by a discussion of human error in the context of the healthcare domain, which will include research limitations of this field and a focus on number-entry interfaces as well as number-entry errors. Next, current HCI research will be presented to discuss how domains where accuracy is of great importance can benefit from research in this field. Afterwards the attention is directed towards transcription tasks in general, which will elaborate text as well as number transcription related to this thesis. The review will also present a series of theories from psychology, which could have valuable applications as well as implications on number transcription tasks to reduce errors. The discussion of the before mentioned theories will then lead up to the introduction of the *DOV* effect. The literature review will conclude with a high level review of the research questions.

2.1 To Err is Human

Errors are part of human nature and constantly appear in daily tasks. For example, whether people accidentally pour orange juice instead of milk in their flakes, put the wrong coloured clothes in the washing machine or simply make typos. People make errors in all kinds of situations even in tasks with the lowest level of complexity (Kantowitz & Sorkin, 1983). The on-line repository *errordiary*[1] is a collection of such everyday errors. *Errordiary*, which can be contributed to by using the hashtag *#errordiary* on Twitter, was created to increase the discussion about human error and its causes (Wiseman, Cox, Gould, & Furniss, 2012). Some examples of many trivial errors can be seen in figure 2.1 where people posted cases of errors while making coffee.

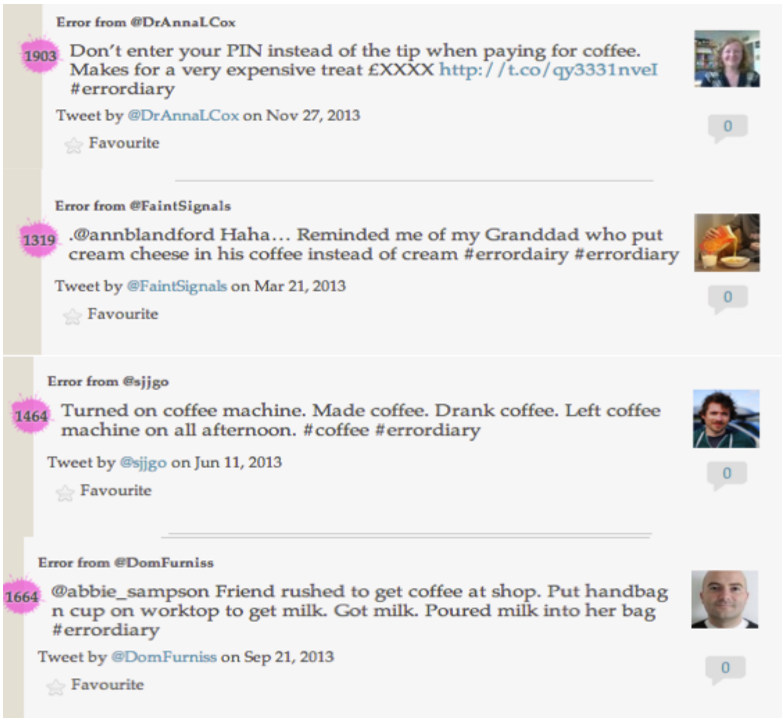


Figure 2.1: Errordiary example posts

Although many of the errors are little more than an inconvenience, in safety-critical domains such as aviation, nuclear systems or healthcare, a small error can have severe consequences. Human errors were the reason for more than 50% of incidents in most domains (Zhang et al., 2000) which were even higher in aviation. According to a report by the United States Federal Aviation Administration (FAA) and also mentioned by Martins, Martins, Soares, and da Silva (2013), an analysis of the flight recorders

[1] http://www.errordiary.org

revealed that 70% to 80% of accidents are based on human error or based on a chain of failures in relation to the human factor (FAA, 2010). A reason why the percentage of human error in some domains is that high could be a feature of reporting. For example, aviation has a culture of reporting human error whereas in healthcare people struggle to report such cases (Section 2.2.1 will elaborate further).

The literature provides us with numerous cases of human error, some resulting in tragedies (Casey, 1998, 2006; Kostopoulou & Delaney, 2007). One example, where a small error led to terrible consequences, was a marine tragedy that occurred in Chesapeake Bay in 1978. The U.S. Coastguard ship *Cuyahoga* collided with the freighter *Santa Cruz II.* The accident happened on a clear calm night and both ships saw each other on the radar and visually. What happened next was that the Cuyahoga made a turn in front of the freighter which resulted in the collision where eleven coastguard men died. There was no equipment malfunction or other technical failure. The later investigation showed that it was solely human error. The errors that could be identified were that the captain of the coastguard ship misinterpreted the configuration of the running lights of the freighter. He miscalculated its size and heading and at the moment when he ordered the turn he thought that he would be clear of the other ship. In addition, the crew did not react although realising what was happening. The crew thought the captain would have everything under control and knew what he was doing. As a result the crew did not intervene and so the tragedy started to take its course (Rothblum, Wheal, Withington, Shappell, & Wiegmann, 2002).

Such incidents are by no means isolated cases. Tragedies where a trivial error lead to disaster happen frequently. Sometimes the reason for an error can be attributed to a mistake made by individuals or a group of people. Unfortunately, many times someone is wrongly blamed as a scapegoat when human error is named as the reason for an accident. However, there are cases where not the people are required to change but rather the system or the design of a system has to be altered to avoid further harm. The following air traffic incident is an example of a case where multiple factors led to catastrophe and where a better number-entry design feature could have saved 87 lives.

In 1992, *Air Inter Flight 148* crashed into the Vosges Mountains at an altitude of 800 meters while executing their landing manoeuvre at Strasbourg airport. At first the landing seemed to be an ordinary procedure for the pilots, who did not have substantial experience on the new Airbus A320 at that time. Both the pilot and copilot were experienced in flying with over 12000 hours flying time between them but the A320 was relatively new at that time and equipped with new digital cockpit devices and instruments instead of analogue displays. One of these instruments was the autopilot interface which can be seen in figure 2.2. Air Inter flights were usually business flights aimed to have a short turnaround rate. This means that the crew aimed to travel as fast as possible to their designated point of arrival. The crew's salary was directly attached to that business model resulting in better wages for faster turnaround rates. At the start, before the crew even took off from Lyon airport they programmed the autopilot to land on runway 23 in Strasbourg. However, when *Air Inter Flight 148* approached Strasbourg airport, there was a sudden change to their landing approach due to high winds and poor weather. The captain was hoping to land on runway 23 which had the best navigational fix for the autopilot but then he was suddenly faced with landing on runway 05 which is exactly the opposite runway. Additionally, runway 05 did not have a precise navigational assistance for the autopilot as the signal was blocked by the Vosges Mountains on that side. The approach on runway 05 would result in pilots not receiving any electronic guidance on their altitude and only on their horizontal position. The flight crew performed landings at Strasbourg many times and was aware of that issue. This resulted in the crew being faced with a sudden peak in their workload to adjust their systems as the captain had to calculate the rate of the decent manually. The recovered cockpit voice recorder revealed a state of emotional tension of the captain at that moment. He calculated a descent of -3.3 degree. A descent angle of -3.3 is considered a normal flight angle which would provide a good approach for landing. Additionally, there was a poor guidance and an improper communication between the controller in the Strasbourg tower who directed the plane to the "right" but from the pilots perspective the airport was to the "left". This added to the pilots confusion and directed the plane closer to the mountains. Directing a plane to the "left" or to the "right" is considered the wrong terminology. This was an issue among many others that the investigation discovered. When the pilot entered the flight path angle of -3.3 he accidentally entered the number -33 in the vertical speed setting, which resulted in an abnormally high vertical speed as 33 in this mode is the abbreviated form of 3300 feet per minute. The captain did not notice that the display modes where the other way around as both of them were displayed in the same overlapping display window and controlled by the same knob. The investigation also discovered that had the crew noticed the interface issue then they could have avoided the crash. A detailed study revealed that pilots made this mistake several times during training exercises. Moreover, there was a little known feature in the autopilots design, which detects if the plane needs to change direction quickly then the autopilot would register this as an emergency situation and reverse the planes direction at twice the programmed rate. In the exact moment as the captain programmed the autopilot there was a turbulence, which caused the plane to ascend slightly and therefore resulted in a high vertical speed of 6600 feet per minute. The captain was concerned about the high speed but his attention was refocused by the copilot on the planes horizontal position to the runway. Shortly after this incident another Air Inter flight had the same issue with the autopilot and the two overlapping modes but this flight could recover in time as they were at a higher altitude and did not have a mountain in front of them. Another issue was that *Air Inter Flight 148* did not have a Ground Proximity Warning System (GPWS) as the airline management did not decide to install such as system due to the false alarms some planes experienced at that time. This procedure was legal at that point in time but because of this decision *Air Inter Flight 148* lost another safety layer. A GPWS is a radar system that points to the ground and sets off a visual and audio alarm in the cockpit in case a plane would fly at a too low altitude (BEA, 1993; FAA, 2014).

Figure 2.2: Original A320 Autopilot design and display modes (source: FAA, 2014)

The investigation concluded that this incident occurred because of many separate elements in a chain of unfortunate events and not solely due to a single error committed by the pilot. Yes, the pilot entered the wrong number but multiple factors of the system, of which he was part of, made it impossible for him or the copilot to resolve the situation in time. The lack of training in the A320 and the sudden increase in workload surely had their weight in this case as well. As a result of this accident more than 35 changes for the aviation industry were proposed by the investigative bodies (Johnson & Pritchett, 1995). One of these changes was a new design for the autopilot interface, which was altered to display a four digit number for the vertical speed setting (see figure 2.3).

The first example in this section shows an error that can be traced back to the mistakes of the captain of the coastguard ship and the trust of his crewmen in his abilities to do the right thing. However, the case of *Air Inter Flight 148* shows, sometimes errors occur due to bad design features and due to multiple factors. This is in line with earlier work by Kohn, Corrigan, and Donaldson (2000) in the healthcare domain. Kohn et al. (2000) mentioned in their work, by examining the healthcare system at that time, that the majority of errors was not caused by people acting careless but by erroneous systems and procedures which lead people to commit errors. This is also one of the reasons why *Errordiary* was created: to raise awareness that sometimes system designs are a key factor in the field of human error (Wiseman et al., 2012). The findings of this thesis, presented in later sections, are also discussed in relation to the *Air Inter* example highlighting its importance.

Figure 2.3: Revised Airbus A320 autopilot interface design (source: FAA, 2014)

Researchers have long been working towards a better understanding of this highly interdisciplinary research area by identifying causes of human error as well as means to prevent them. One of the reasons why research is focused to understand human error is trying to understand the attribution i.e. what was the cause of the error. The philosophy of causation is an incredibly deep field. However, the psychology of causation (identifying the cause of an effect i.e. the variable that causes a change in other variables) is making good progress towards people understand what is behind an error (Dekker, 2006). This thesis acknowledges previous work done to understand human error but specifically focuses on human error in transcription tasks. More explicitly, the thesis presents a novel approach to ameliorate transcription errors in an interactive context.

2.1.1 A Brief Historical Review of Human Error

Whilst there are studies on human error going back into the 1940s by Fitts and Jones (1947), it can be seen that early studies on human error originate from psychology and aimed on aspects such as human judgement and human performance in decision making (Brunswik, 1955; Hammond & Summers, 1972). Later studies focused on errors in human performance as part of a system (Norman, 1980). In the mid-1970s developments in cognitive psychology, a series of tragic incidents such as the *Three Mile Island* and the *Challenger* disaster helped to establish human error as an independent research body (Reason, 1990). Today the field is a highly multidisciplinary research area which embodies domains such as cognitive psychology, computer science, aviation and numerous others.

The research community described human error with different views namely, the old and the new view (Dekker, 2006). For example in the old view of human error it can be seen as *the cause of a mishap.* This means that human error is seen as a cause of incidents. The earlier example of the U.S. Coastguard ship can be seen as an example for the old view. Whereas in the new view human error can be seen as *the symptom for deeper trouble.* In the new view, human error is not seen as the sole cause of an accident but rather an indicator or a part of a larger system that failed which makes it necessary to discover the circumstances which led to the adverse effect. The example of *Air Inter Flight 148* would be an example for the new view of human error as so many different factors contributed to the plane crashing into the mountains.

The literature provides us with several resources where researchers were even discussing the nature and meaning of the term *human error* itself (Hollnagel & Amalberti, 2001). While some researchers argue the term is inappropriate due to its lack of explanatory power by only explaining that humans are involved in an incident (Park, 2003; Henriksen, Dayton, Keyes, Carayon, & Hughes, 2008) others valued the term as it reflects the impact of human actions and responses to many major incidents as well as ordinary less disastrous events (Rouse & Rouse, 1983; Reason, 1990). Generally speaking, there is no solid definition of human error. Rasmussen, O., Mancini, Carnino, Griffon, and Gagnolet (1981) (p. 313) provided an early attempt:

"Frequently they [human errors] are identified after the fact: If a system performs less satisfactorily than it normally does - due to a human act or to a disturbance which could have been counteracted by a reasonable human act - the cause will very likely be identified as a human error. This is probably because the analyst will not have the information - or psychological background - which is necessary to trace through the human performance in the explanatory causal backtracking process to find a possible causal input."

To understand the field of human error, researchers started to create taxonomies and to define the phenomena. Clear definitions can help to better understand as well as facilitate communication about a topic. However, as with many other aspects in research a single clear definition is not always possible and this is also the case for *human error*. An alternative approach is to classify *human error*.

2.1.2 Classifications and Models of Human Error

In order to develop a better understanding of *human error* and to distinguish between different kinds of errors, over time researchers have developed a number of taxonomies, classifications and models. Early research in this area was mainly focused on aeronautics and nuclear power systems (Rasmussen et al., 1981). One early classification by

Swain and Guttman (1983) classifies errors as errors of omission and errors of commission. The first category are errors that occurred simply by omitting a complete task or a step in a task. The latter category is further divided into four subcategories: selection errors (caused by selecting the wrong control, making improper connections or issuing the wrong command), errors of sequence (caused by poor performance of one or more tasks in a sequence for a correct system performance), time errors (some action committed too early or too late) and qualitative errors (too little or too much). The following figure 2.4 illustrates this classification.

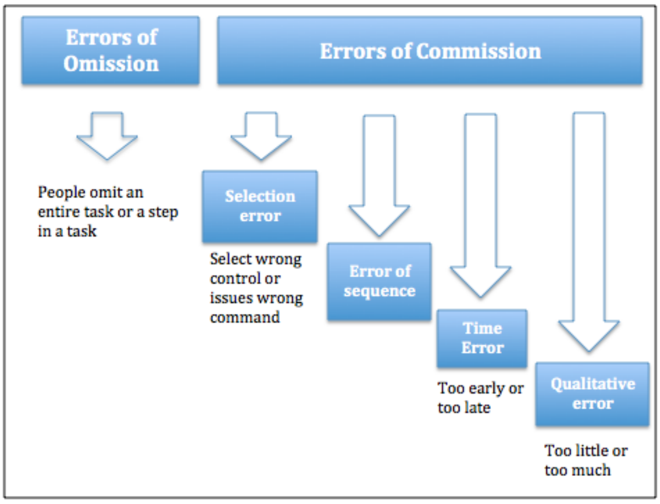


Figure 2.4: Error classification according to Swain and Guttman (1983)

Norman (1980, 1981) sees human errors divided into 'slips' and 'mistakes' and while mistakes are intentions that are considered inappropriate, slips are regarded as an unintentional error. Norman (1980) and others have approached slip errors with an information processing view. This means researchers argued that errors mainly occur due to problems in information processing (Zhang et al., 2000). Rasmussen (1983) also adapted the information processing view. He introduced three categories of human behaviour: skill-based, rule-based and knowledge-based. Additionally, Rasmussen (1983) further categorises different modes of error, mechanisms and causes. Rouse and Rouse (1983) presented an early information-based classification, that follows the information processing view. Such an information processing view should appear during the human interaction with a system, for example, while humans overview flight control systems or medical systems. The human error classification by Rouse and Rouse (1983) can be related to Norman's stages of user activities (see figure 2.5): perception of the state of the world; interpretation of their perception according to their expectations; evaluation of their interpretation relative to what was expected; formulation of goals to achieve what is desired; formation of an intention to act to achieve the goal; specification of a sequence of actions and execution of the sequence of actions (Norman, 1990).

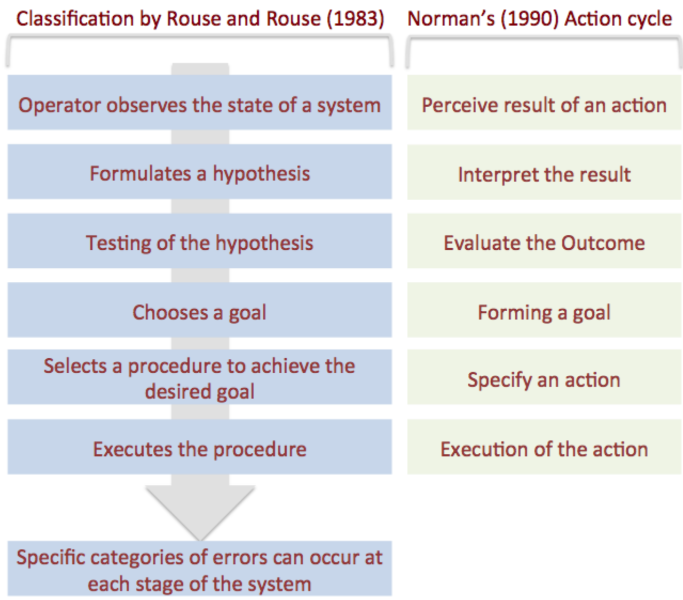


Figure 2.5: Error classification: Information Processing Model

One of the most widely accepted classifications was developed by Reason (1990). The work by Reason (1990), is based on cognitive aspects such as attention and perception and classifies errors according to intended or unintended actions. For example, lapses and slips are both classified as the result of unintended actions. Slips occur due to a lack of attention and lapses are caused by memory failure. The crew of *Air Inter Flight 148* suffered from attentional failure which lead to the Captain refocusing on the planes horizontal position rather its vertical speed. A critical loss of another safety-layer. Additionally, a mistake for example, is classified as the result of an intended action, which did not succeed in meeting the goal and can happen due to a lack of knowledge (knowledge-based mistake) or due to wrong rules applied (rule-based mistakes). This classification is also used in the International Standardisation Organisation's (ISO) standard 62366 for medical devices (see figure 2.6). This standard presents specific processes for medical device manufacturers to analyse, design, verify and validate usability in relation to safety issues and uses Reason's (1990) classification to classify the different errors that can arise (ISO, 2007). Additionally, Reason (1990) also distinguishes errors into skill-based, knowledge based and rule-based building on previous work by (Brunswik, 1955; Hammond & Summers, 1972; Rasmussen, 1983).

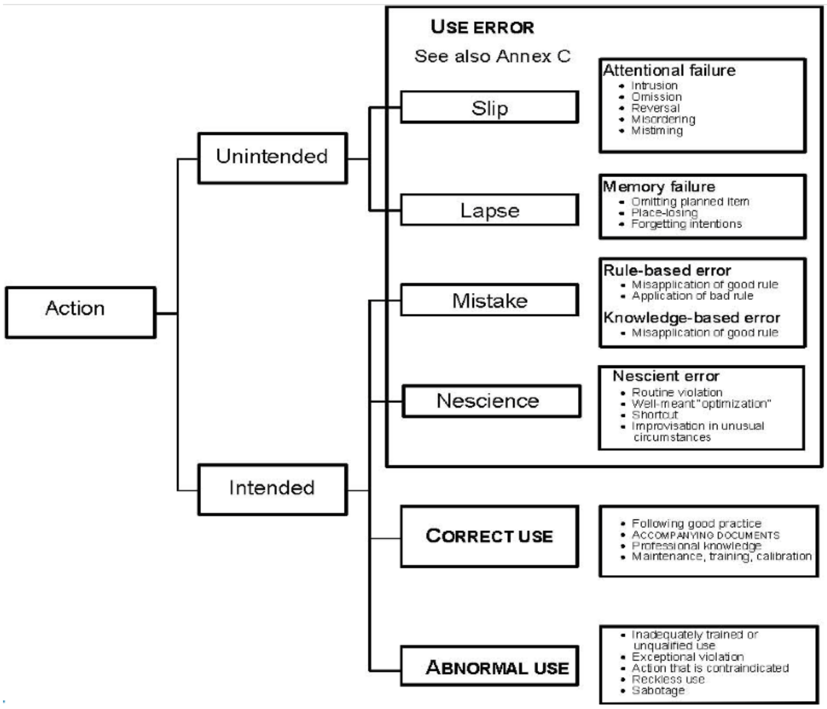


Figure 2.6: Classification of 'use errors' (source: ISO, 2007)

Reason (1990) also combines many of the earlier notions of human error in his work and developed the widely accepted Swiss cheese model. The model depicts a system consisting of a stack of slices of Swiss cheese. The holes in each slice or layer represent opportunities for a process to fail. Each slice symbolises a defensive layer in the process. An error can allow a problem to pass through a hole in a layer but in the following layer the holes are in different places, which results in the problem to be caught by this defensive layer. If all holes are aligned of the different layers this represents a flawed system where catastrophic errors can occur.

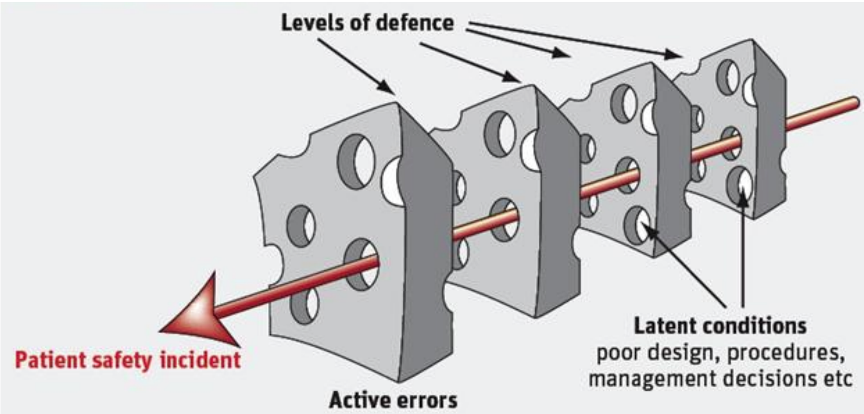


Figure 2.7: Swiss cheese model (source: Reason, 1990)

As previously discussed, in many cases *human error* does not occur as a single fault but rather than as a result or part of a chain of results that lead to an accident. According to Kohn et al. (2000), errors can happen when many factors come together to allow them to happen. The above mentioned classifications aim to better understand errors which can help to make systems safer. However, sometimes even safety mechanisms, although put in place with good intentions, can undermine other or even all layers of security. Li and Thimbleby (2012) expanded Reason's model to better understand how incidents are caused. This extended version of Reason's Swiss cheese model - the so called "Hot Cheese" model should encourage investigators to direct their attention to parts of a system that are put in place to avoid errors but then when these safety mechanisms go off they are actually causing other errors which successfully pass through all safety barriers, or in the model called, slices or layers. For example, an airport closes down one runway due to small parts of debris forcing all planes to land on the second runway as a safety precaution. However, the high demand on that single runway causes further danger and additional high wind problems which happen on the second runway actually can now present a far more dangerous problem for certain aircraft.

There are many more classifications and taxonomies in the literature, which are to some extend mostly targeted, to be applied to a specific problem area such as aviation systems, healthcare systems or other safety-critical systems (Wiegmann & Shappell, 2001; Wiseman et al., 2011). The Human Factors Analysis and Classification System (HFACS) is one such a tailored and widely accepted classification (Schmidt, Schmorrow, & Figlock, 2000; Shappell & Wiegmann, 2001). It was developed to provide safety professionals with a tool for identifying and classifying the human causes of aviation incidents and is built on Reason's (1990) human error model of latent and active failures (Shappell & Wiegmann, 2001). HFACS describes human error at each of four levels unsafe acts of operators, preconditions for unsafe acts, unsafe supervision, and organisational influences.

Although multiple classifications would probably be applicable to the studies presented in this work, the thesis primarily builds on the classification by Reason (1990) (as it is focused on cognitive aspects) and a classification for number-transcription errors by Wiseman et al. (2011).

2.2 Human Error in Healthcare

Healthcare is a complex and data intensive domain. People in healthcare are constantly interacting with a variety of systems under pressure, physical and cognitive stress. It is a highly demanding area focused on efficiency, accuracy and effectiveness. Because of the complexity of that area and the constant changes of technology there is a growing interest in studying human errors.

As with each complex system there come potential vulnerabilities to errors. Some of these vulnerabilities can be attributed to faulty systems, others to human factors or both. According to (Kohn et al., 2000), an estimate of 44000 to 98000 people die each year in the U.S. as a result of medical errors.

Whereas some safety-critical domains such as aviation greatly benefited from human error research, in healthcare this is not always the case. Many errors are not reported or simply can not be reported. Some errors seem so small so mundane and yet they can cause great harm. An example, in particular, are number-entry errors when using infusion pumps.

2.2.1 Research Limitations in the Healthcare Domain

According to Vincent, Neale, and Woloshynowych (2001), about 11% of patients in UK hospitals suffer from error incidents. From these 11% almost all of them can be prevented. Yet Healthcare has a very sensitive organisational culture. Blame is sometimes the first thing that typically follows an adverse event in healthcare. The problem is that blame, no matter in which direction self-blame, blame of the whole system or blame of others, does not help to recognise and to fix critical flaws in a system. As a result people and institutions are very careful about admitting and reporting mistakes. The literature shows many cases where people are rather tempted to blame the hospital staff or blame a scapegoat to appease the public, which is not the right approach (Aspden, Wolcott, Bootman, & Cronenwett, 2007). Additionally, how such adverse events are published is still causing concern and sometimes badly reflect on people working in the healthcare domain.

In general, studying human errors in healthcare can be challenging and requires researchers to overcome many barriers. Firstly, a researcher in healthcare need to take into account that healthcare is a highly multidisciplinary area (Atwal & Caldwell, 2006). Therefore, detecting error is sometimes not possible due to many factors such as the specific environment and how people interact with systems. For example, it is difficult to detect number-entry errors in drug infusion pumps made by nurses during surgery. Actively monitoring medical staff while entering numbers is in most cases not feasible and may introduce uncomfortableness or even fear, which can badly reflect onto patient safety. Indeed many medical devices provide logging features, however the devices do not log what the user had intended to enter but what the device interpreted. Thimbleby (2011) stated that devices which only record the final number that was entered instead of the user's detailed actions are insufficient to determine the source of a number-transcription error.

The complexity of the healthcare domain is in itself a challenge, which researchers need to take into account. This means that for example, if a study is being designed it might involve a great number of different organisational and technological levels within healthcare. A study conducted to investigate number entry errors in infusion pumps might be feasible in one department but might get much more complex if, for example, the study should not take place in paediatrics but in the intensive care unit.

2.2.2 Healthcare Interface Design & Number-Entry Errors

The healthcare domain is increasingly relying on a vast amount of technology. This includes ordinary administrative equipment or complex medical devices in the operating room such as respiratory ventilators, defibrillators or infusion pumps. As many of these devices are used in emergencies or in other life-saving situations it can be assumed that such interactive systems are well designed therefore using them would not lead to confusion or result in errors being committed. This is however by far not the case. Software issues (Wallace & Kuhn, 2001) and design issues (Wiklund, 1995) have long been studied.

Still, a poor interface design and a lack of proper testing lead to many errors which could otherwise be avoided such as mode errors as well as poor or ambiguous device feedback on operations or states (Obradovich & Woods, 1996; Zhang, Johnson, Patel, Paige, & Kubose, 2003; Wiseman et al., 2011). Mode errors are errors where users perform an action in one mode but confuse the current mode with another one. For example, using the fill all tool instead of pen tool in a photo editing program. The *Air Inter Flight 148* incident included a mode error committed by the captain as he entered the wrong number in the wrong window or mode of the autopilot system.

One area in particular, shows the high cost that can come with errors and that is drug delivery. In drug delivery medical staff, usually nurses or doctors, administer prescribed drugs via syringes and sometimes via syringe pumps or infusion pumps. To administer a drug to a patient via an infusion pump requires the nurse to program the device with the rate of the drug and the volume to be infused (VTBI). In specific cases it can be the patient him or herself who is required to program such a device. The numbers to be entered are sometimes calculated in advance or transcribed from a source for example a label on the drug or the patients' health record. There are also additional steps in between delivery which can include turning on the device, resetting the device, changing modes and others depending on the specific device. Furniss et al. (2011) describe in detail the tasks of a drug administration task in order to evaluate medical device design in the context of use. The following two examples of number-entry errors show how critical such small transcription tasks can be to patients:

The first example is the case of Denise Mellanson. She was a cancer patient being treated in a hospital. She used a portable infusion pump containing a four day dose of two chemotherapy drugs: cisplatin and 5-fluorouracil. The dosage was calculated (with a calculator) and programmed by a nurse. A second nurse was asked to confirm the calculation but could not find a calculator thus performed the calculation in her head. The second nurse did not recognise any errors and the patient was discharged. After four hours the patient noticed the infusion-pump was beeping and the drug containers were empty. She returned to the hospital where the error was recognised but it was decided that no treatment was necessary. The patient was again discharged with the advice to drink plenty of fluids to counteract the large amount of drug that has been administered. Two days later the patient complained about vomiting and nausea but was again sent home as no beds were available. Denise Mellanson died after 20 days as a result of the error (Grissinger, 2011). The subsequent investigation discovered that among a calculation error, there were flaws in the infusion-pumps design which had attributed to the error. For example, the screen to review the programmed settings did not indicate the duration of the drug delivery that had been programmed into the device and the device itself did not include any dosage limitation software meaning it was not able to detect what could be a potential dosage excess.

The second example is the case of an experienced intensive care nurse (Kimberly Hiatt) who accidentally made an error while programming an infusion pump. The nurse dispensed 1.4 grams instead of 140 milligrams of calium chloride leading to the death of eight-month old Kaia Zautner. Ms Hiatt was a nurse with 24 years of experience and although she reported the incident immediately she suffered from the following investigation. Half a year later the nurse committed suicide (Clancy, 2012).

Both cases show different aspects that can contribute to errors in drug delivery. Moreover, they also show that better medical device design could have potentially introduced another safety layer, as Reason (1990) described, which could have attributed to detecting and preventing the errors to progress any further. The United States Food and Drug Administration (FDA) and other official bodies, such as the Institute for Safe Medication Practices (ISMP) are recognising these issues and published standards and guidelines to address as many issues as possible (FDA, 2000; ISMP, 2007). For example, the ISMP published a list of error-prone abbreviations including guidelines on how dose designations should not be presented to avoid errors. For example, no trailing zeros should be used after the decimal point for doses that representing integers i.e. "whole numbers" (1 instead of 1.0). Another example is the representation of the dosage information. What should be avoided is presenting the numerical dosage and the unit of measure together without sufficient space in between (10 mg instead of 10mg) (ISMP, 2004).

Focusing on infusion pumps, there are many different kinds of such devices used in healthcare. Having such diversity of device manufacturers ultimately results in an equal diversity of interface designs for infusion pumps. Some of the most common device interface designs are *Chevron* i.e. incremental, *Dial*, *Serial* i.e. calculator-based and *5-Key* interfaces. In a *Chevron* interface users can only increment or decrement the number by using fixed actions typically in the form of a pair of "up" and "down" buttons. The *Chevron* interface shown in figure 2.8 (c) consists of two buttons each to increment or decrement numbers. The buttons with double chevrons indicate a larger amount that can be changed compared to the buttons with a single chevron. A *Dial* interface (see (d) in figure 2.8) consists of a knob that can be turned left or right in the same way as a volume knob on a radio. *Serial* interfaces (see (a) in figure 2.8) provide a calculator like interaction for users and they can be found on keypads, cash-machines or phones. *Serial* interfaces provide a direct output mapping of the numbers. However, *Serial* interfaces also vary in their layout design. While there is an argument in the literature that having the lower numbers listed in the top row (1, 2, 3) would be a better design in context of accuracy than having the higher numbers listed at the top (7, 8, 9 or sometimes 8, 9, 0) (Conrad & Hull, 1968) others suggested placing the zero key at the bottom of the interface (Marteniuk, Ivens, & Brown, 1996). *5-Key* interfaces (see (b) in figure 2.8) such as the one on the *B.Braun Infusomat ® Space* consists of left & right buttons which select digits, up & down buttons which increase or decrease digits and a fifth key which is sometimes labelled as *OK* or *Save* button to submit the entered number.

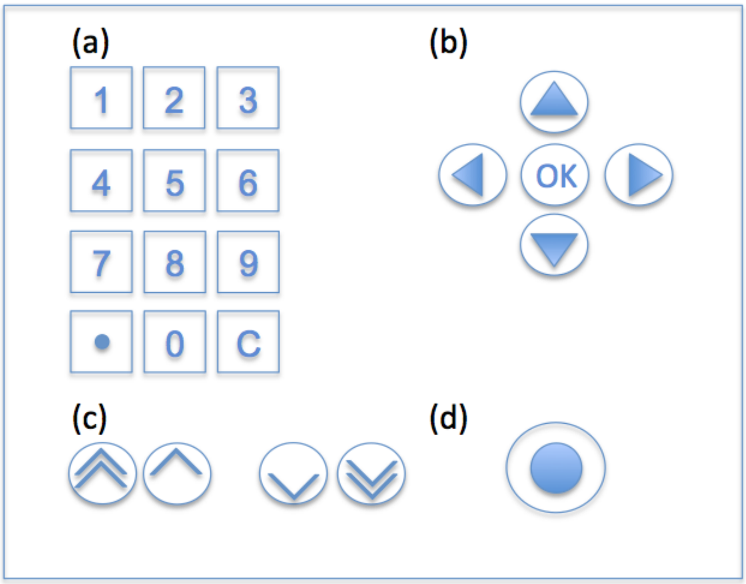


Figure 2.8: Design variations for infusion-pump interfaces

Oladimeji et al. (2013) evaluated the above listed interfaces and found that a *Serial* interface is the fastest in terms of number-entry whereas a *5-Key* interface is the most accurate one. *Dial & Chevron* interfaces are the slowest in number-entry and seem to open room for frustration when it comes to large number-entry tasks. Oladimeji et al. (2013) also mentioned that *Dial & Chevron* interfaces bear the risk of undershooting or overshooting the intended target number but have the advantage that users might focus more on the target display rather than the interface buttons or knobs and therefore are able to spot an error.

2.2.3 Human-Computer Interaction & Number-Entry Errors

With the increasing proliferation of devices used in healthcare and with an increasingly ageing population in the UK, more people will need to rely on medical devices in a multitude of environments such as in a hospital or at home (Gomes & Higginson, 2008). Blandford (2010) pointed out that reliability will become more and more an important factor for healthcare devices and that a growing automation of devices in healthcare contributes to a decrease in direct communication between patients and medical staff. Therefore it seems natural that these devices are designed to be easy to use and to avoid errors. However, earlier work by Obradovich and Woods (1996) presented that medical devices do suffer from design issues.

HCI research is an essential part in medical device design but researchers argue that traditional techniques have difficulties to cope with the complexity of device design in the healthcare domain (Thimbleby, 2007). According to Thimbleby (2007), HCI techniques such as User-Centred-Design (UCD) must be present in medical device design but they are insufficient when it comes to complex medical devices as the studies might not be able to reveal all the problems. Blandford (2010) adds, that this is especially the case for people who differ in their training. Training is of course an essential way to prevent errors but training cannot help to prevent slip errors (Byrne & Davis, 2006; Back, Brumby, & Cox, 2010). Blandford (2013) further proposes seven challenges to investigate technologies used in healthcare. Holzinger, Thimbleby, and Beale (2010) argue that HCI research should be aimed at the daily routine of medical staff and the technology which is used to improve the safety and effectiveness in healthcare.

It is known that misprogrammed medical devices are the second most frequent reason of medication errors, making it clear that it is important to reduce as many errors as possible (Smetzer & Cohen, 2006). In number-entry processes of medical devices, such as infusion-pumps, little is known about what errors people make and to what extent such devices can prevent or mitigate errors (Thimbleby & Cairns, 2010). Research around number-entry errors is a relatively young but fast expanding area and researchers have realised the importance to chart this area.

Research in number-entry has approached the phenomenon of errors in the interactive medical device context from different angles. For example, Wiseman et al. (2011) attempted to better understand what kind of errors people make while using infusion-pumps and established a taxonomy to classify the detected errors. Oladimeji et al. (2013) evaluated different interface designs for infusion-pumps, such as listed in Section 2.2.2, and concluded that some designs (while faster) are at risk of making more severe errors while by using other interfaces users require more time to enter the numbers but do not make as many errors. Others focused their attention on specific design features of medical devices (Cauchi et al., 2012).

Intuitively, double checking drug delivery information such as rate or VTBI may seem a natural way of eliminating errors. Although this technique can detect errors, it has been shown that in some cases double checking is not effective. For example, if the second check or safety check is not performed by another person than the one who performed the first number-entry task, then confirmation bias makes it difficult to detect errors (Grissinger, 2003). Additionally, even if two individuals perform the checks they may be unable to detect differences if they are distracted or do not act independently.

In case of programming an infusion device certain design features (Cauchi et al., 2012) or artefacts become important such as the way in which the to-be-transcribed information is presented (Back & Cox, 2013). Previous work by Back and Cox (2013) has shown that artefact arrangement of interactive devices can have a critical impact on the way people perform the programming task, the usability and user experience of the device as well as subsequently the errors made during the interaction. In their work they apply the *gestalt law of similarity* (a method used by the HCI community to understand how people's perception is grouping objects) which states that similar items should be grouped and presented together whereas a missing similarity is suggesting differences in the items. These differences should then be reflected in the visual grouping of such items. However, Back and Cox (2013) found that this should be done by keeping the specific task in mind and this should also be done for all items involved in the interaction with the device. For example, grouping items on the interface itself are not enough meaning items on the source or any procedural step (task) need to be considered as well.

Additionally, previous work focused on interruptions, how they affect number-entry tasks and what approach can be taken to analyse task resumptions (Rajkomar & Blandford, 2012). For example, (Back et al., 2010) demonstrated that an enforced lockout period after an interruption can be beneficial and reduce (sequence) errors by 64%.

There is also the argument that errors can be reduced by visual cues and by lowering the cognitive load of people. Byrne and Bovair (1997) for example, worked on reducing post-completion errors. A post-completion error is a procedural error that occurs after the main task has been completed. Their study, which was focused on a Star Trek game, concluded that indeed by reducing the memory load and providing visual cues at the right time that post-completion errors can be reduced.

An interesting aspect in HCI focuses on people's task performance and the strategy that they use to complete the task (Morgan & Patrick, 2010). The cost to access specific information is a central aspect in this context. Information Access Cost (IAC) is defined as: "the time, physical and mental cost of accessing information" and was introduced by (Ballard, Hayhoe, & Pelz, 1995). IAC can cause changes in people's cognitive processing strategies which in the end affects their task performance. For example, IAC in a task can be influenced by hiding a specific aspect, such as the display or the interface, in the interactive context. The cost of accessing the hidden part (low or high) can be controlled by making the step required to uncover the hidden part either easy or difficult (single key press or key combination to reveal the display). Previous research has shown that the ease with which a user can access the to-be-transcribed information can impact the strategy which they use to complete the tasks. This means, whether or not the user has to, for example, click to reveal it (Gray, Sims, Schoelles, & Fu, 2006), can lead to differences in accuracy of performance. Back, Cox, and Brumby (2012) evaluated IAC by looking at multitasking in programming of infusion pumps and found that with a high IAC, people are interleaving tasks by the use of a more memory-intensive strategy. Gray et al. (2006) specifically focused on memory and the cost to access the information explains the Soft Constraints Hypothesis, which suggests that people are more likely to use a potentially less accurate memory-intensive strategy over a more accurate perceptual-motor strategy when the cost of accessing information is higher. Whilst much research has focused on decreasing IAC, recent research has demonstrated that there may be disadvantages to this approach if the information processing strategy adopted is too passive (Morgan & Patrick, 2010).

Researchers have realised that there is a gap in HCI techniques to address the area of medical device design. Cauchi et al. (2012) started to look at specific design elements for number-entry interfaces to reduce errors which is one aspect this thesis will focus on in Chapter Three. Additionally, research surrounding IAC is touching other areas in psychology which are based on similar concepts but suggest an opposing view that when information is harder to access can be beneficial in terms of memory and accuracy (Diemand-Yauman et al., 2011). This is another aspect which this thesis will investigate in form of a series of studies in Chapter Four.

2.2.4 Attention & Distraction

The early studies of this thesis (Chapter Three), show a persistent but low error-rate regardless of the experimental design. Distractors are known to increase error-rates (Forster & Lavie, 2008; Lavie & De Fockert, 2003), a fact that later studies in this work employ. Attentional failure is one cause that lead to slip errors (Reason, 1990; ISO, 2007). Previous work focused substantially on understanding the nature of attention as well as how attention can be captured by distracting items. This gravitated towards a discussion of whether distracting items are filtered out shortly after they have been perceived (early selection) or if such distracting items are fully processed but prevented from influencing actions or memory (late selection). The perceptual load theory of attention (Lavie, 2005) states that whether early or late selection occurs depends on the perceptual load that a person is experiencing at that moment. For example, while the participant is doing the primary task in an experiment. A dual task, for example with a distracting secondary task (as used in later studies) is a common strategy used by researchers to provoke more errors.

Perceptual load has mostly been studied in visual search tasks. In case of low perceptual load, that is where there are few items to search through or items that are easily distinguished from the target item, there is a lot of spare capacity for processing distracting items therefore late selection occurs. On the contrary in case of high perceptual load, that is were there are many items or items that are easily confused with the search target, there is much less capacity for processing distracting and therefore late selection occurs. In summary, situations with high perceptual load distracting items are less distracting.

In many situations in daily life distractions are not of visual but of audio origin. Only a select number of workplaces are entirely silent and healthcare or aviation environments are no different. In fact, safety-critical domains depend on audio inputs to capture the attention in case of an emergency. It was therefore decided to chose audio distractions, to have some comparison to, although idealised, number transcription tasks in a busy safety-oriented environment.

Hearing can play an important role in monitoring a specific environment and can therefore act as an early warning system with unexpected sounds capturing the attention even if they are not relevant to the task at hand (Dalton & Lavie, 2004). Additionally, Murphy, Fraenkel, and Dalton (2013) showed that unlike the visual search task, where perceptual load can result in filtering out the distracting item, in an audio search task, distractors are still perceived regardless of the audio perceptual load. This was also the case cross-modally with the perceptual load of a visual search task having no significant effect on the perception of audio distractions (Tellinghuisen & Nowak, 2003). In particular, in the visual search task of looking for either an X or an N in a group of letters, an audio distractor of X and N either increased or decreased reaction times depending on whether the distracting item was the same or different from the target letter.

When it comes to understanding the more complex environment in which healthcare devices are used, the literature on distractions and their effect on the performance of the different tasks (particularly perceptual tasks) was employed to improve the error-rate and taken forward into Chapter Five.

2.3 Transcription Tasks

Routine data transcription tasks occur in a large number of places, for example, entering accounting data into a spreadsheet, and paying bills via an on-line banking system. However, people make errors. For example, nurses or doctors are often required to enter large quantities of sensitive data in electronic health records (EHRs) or to program drug infusion-pumps with information located elsewhere on a prescription or on a label of a drug infusion bag.

2.3.1 Text Processing

Previous research focused on the performance in typing (Rumelhart & Norman, 1982) and on understanding the processes underlying text-transcription tasks (Salthouse, 1986; Logan & Crump, 2009). Wu and Liu (2004) (p.381) describe text-transcription tasks as: *"intricate and complex interaction of perceptual, cognitive, and motor processes"*. Work by other researchers are in line with Salthouse (1986), that transcribing text includes the collection of the source text via a sequence of eye-movements, followed by the translation of the collected information into a sequence of keystrokes e.g. (Inhoff, Morris, & Calabrese, 1986).

Further to that, there is the argument in the literature that people possess a control system which is divided into two parts: an outer loop and an inner loop (Rumelhart & Norman, 1982; Wu & Liu, 2004). The outer loop can encode the to-be-transcribed content into a sequence of words. Whereas the inner loop is responsible for transforming the words into keystrokes. (Logan & Crump, 2009) add that the inner loop is also responsible for identifying finger or hand movements and suggest that the loops are encapsulated, meaning the outer loop is unaware which hand or finger types which characters and only the inner loop is aware of this.

The argument has been made in the literature that findings in text-transcription may be applicable in the domain of number-transcription tasks (Wiseman, Cox, & Brumby, 2013).

2.3.2 Number Processing

There is at the moment no clear understanding of how people process numbers as research in this area is rather scarce. However, there have been attempts to establish a body of knowledge surrounding number-transcription tasks.

Early research found that short-term memory and working memory is limited in context of numbers. In other words people struggle to memorise more than seven random digits (Miller, 1956). McCloskey et al. (1985) argued that people possess a system which processes numbers. Moreover, this system would be responsible for how people produce and understand numbers. McCloskey et al. (1985) add that, in their view, there is a mechanism which is responsible for processing numbers in Arabic notation and another one for verbal numbers. Both parts include factors which look at the syntax and semantic of a number. Other researchers expanded this view and add that there might be a another part, which stores the analogue quantity of a number in memory (Dehaene, 1997). According to Dehaene (1997), this third part is non-verbal. Once people perceive a number (hear or see) they produce an estimate of the number in their mind in form of an image but this estimate becomes increasingly imprecise once the number is increased. Dehaene (1997) describes the combination of the three parts in his work as the Triple-Code-Model (TCM).

Another view focuses on spatial representation of numbers. Dehaene, Bossini, and Giraux (1993) present the idea of the spatial numerical association of response codes (SNARC) effect which says that people possess a mental number line where they place numbers during the encoding process in memory. Previous work has shown that there are indications that people place numbers along this mental line from left to right and from small to large. However, views on this theory are divided as others found the same effect could be shown with letters (Nol, Rouselle, & Mussolin, 2005).

The issue of how people process and transcribe information is an important aspect in healthcare. This thesis presents studies which investigate the effect of interface design features on accuracy in transcription tasks which may help to get a better understanding of how people process numbers, what strategies they use or what effect specific design features can have on memory.

2.4 Dual-Processing Theories of Reasoning

For many years researchers have been discussing the theory of two distinct cognitive ways of thinking underlying reasoning (Evans, 2003; Kahneman, 2008, 2011; Frankish & Evans, 2009). Dual-processing theories of reasoning (DPT) were proposed by researchers from different scientific areas such as decision making, cognition, learning or deductive reasoning and developed largely independent from each other (Frankish, 2010). DPT was further inspired and brought to the attention of the general public when Daniel Kahneman and Amos Tversky won the Nobel Prize for their work (Kahneman & Tversky, 1982). Numerous researcher have since then focused on DPT and many concepts, models and terminologies arose over the time. Different names were used by the scientific community to describe these kinds of reasoning such as for example "implicit" and "explicit" (Reber, 1989; Evans, 2003). According to Evans (2003), the terms *implicit* and *explicit* refer to the functional differences between the two kinds of thinking which are also described as *Type 1* or *Type 2*.

However, Stanovich and West (2002) introduced a neutral and increasingly common terminology for the two different ways of reasoning namely *System 1 & System 2*. *System 1*, as described by Kahneman (2011) and Evans (2003), is a form of cognition shared between humans and animals. According to Stanovich and West (2002), it is a set of several subsystems which also include instinctive behaviours. *System 1* is hard to control, effortless as well as fast and automatic in execution and non-conscious (Kahneman, 2008; Frankish, 2010). On the contrary, *System 2* is believed to be a more recent development and considered a human-only kind of thinking (Evans, 2003). *System 2* is limited in its capabilities (i.e. working memory capacity), effortful and slow in execution. However, it is a fast learner, conscious and can be controlled. *System 2* is considered to monitor *System 1* but only on a casual basis (Kahneman, 2008). For example, when a calculation task changes from a trivial to a more effortful task *System 2* takes control. An example for such a scenario is thinking of tasks such as the question of what is the result of 3×1 or the question of what is the result of 922×1319. In the latter example, *System 2* takes the helm over the thinking process. According to the theory, people do not have an immediate number in their head for such a complex calculation so they have to compute the result. It is also assumed that cognitive biases are attributed to *System 1* whereas logical thought processes and associative processes are attributed to *System 2*, which is also believed to be a rule-based system (Frankish, 2010). There are several DPT in the literature which circle round the core aspects of *System 1 & System 2*. However this thesis is motivated by Kahneman's (2011) DPT approach.

2.4.1 Opposing Views of DPT

During the past twenty years many researchers have critically evaluated the dual-process theory (Gigerenzer & Regier, 1996; Keren & Schul, 2009; Osman, 2004; Keren, 2013). Most recent Keren (2013), specifically addresses Evans and Stanovich (2013) view of DPT in different aspects which elicited over time in the research community. One aspect are *multiple and vague definitions* by mentioning the increasing confusion about the complexity that the construct of a dual-system has developed over time. Numerous terminologies have been introduced to describe and explain the different aspects which led, in the view of Keren (2013), to a "deformation" and "ambiguity" of the dual concept. The continuous attempt of researchers that promote the DPT are in the view of Keren (2013) rather futile attempts, which only add to the growing confusion. Additionally, it seems that none of the DPT possess the foundations of well defined constructs and scientific scrutiny under investigation. DPT is hard to test and to grasp as one of the major criticism is that the theory is simply too broad and researchers have used vague terminology (Keren, 2013). Evans and Stanovich (2013) in response mention that they approve of the critics of a dual-systems model but abandon the entire list of specific characteristics that distinguishes the two systems from each other except working memory, mental simulation and cognitive coupling (Keren, 2013). Moreover, there is evidence that working memory is not a unitary concept meaning there are different views on how working memory functions (Baddeley, 2007). Additionally, there are concerns about the criterions used in the different models. Keren (2013) argues that mental simulation as measurement to define aspects of *System 2* processes is highly questionable. The literature shows that mental simulations are *System 1* processes similar to perception and motor movements Keren (2013). However, Evans and Stanovich (2013) assume that most of the criticism they received do not apply to their specific model and their newly revised DPT should silence those critical voices. Moreover, work by Osman (2004), brought to light that it might be wrong to divide higher cognitive functions into simply two systems. Nevertheless, arguments by the research community who oppose or who favour DPT are not the focus of this thesis. It is not the intention of this thesis to prove or disprove the existence of DPT but to investigate and to discuss possible effects on reducing number-entry errors which are caused by interface design elements based on the theoretical model of a DPT. DPT have strongly motivated parts of this thesis but in a very narrow and specific way.

2.4.2 Applications & Implications of DPT in Number-Entry

A very interesting question is how can transcription tasks and interaction tasks with medical devices benefit from DPT? Is such an application possible? If the human mind really operates as researchers in psychology suggest, would it be possible to somehow stimulate deeper, more effortful, cognitive processes to make people think harder? Would it be possible to somehow influence peoples' mind so that they make less errors when they transcribe numbers?

There are attempts in the literature to examine ways to stimulate *System 2*, that is the slower thinking, in certain situations. Some researchers simply try to provoke people to think harder. Diemand-Yauman et al. (2011) for example, measured the effects of hard to read fonts on the outcome of students who had to memorise information. Their results indicate that there is potential for great improvement in performance of people's memory when presenting information in a harder to read font. Kahneman (2011) described a similar process of applying blurred fonts to make *System 2* take control of peoples thinking and to improve their accuracy and consequently reducing room for errors.

Apart from studying the effects of traditional interface elements on medical devices in transcription tasks this thesis is therefore also presenting a novel approach on interface elements which are influenced by cognitive aspects based on a DPT (see Chapter Four).

2.5 The DOV - Effect

There is considerable work in psychology, which focuses on the theory that people's learning processes, for example in multimedia, would benefit from a more active and deeper processing of the information (Mayer, 2009; Kuehl, Eitel, Damnit, & Koendle, 2014). This means that, for example, students who read mixed learning material (text and pictures) would be more active and and therefore learning outcomes would increase (Kuehl et al., 2014). However, as Kuehl et al. (2014) mentioned, it is difficult for designers or educational instructors to stimulate a deeper processing of the perceived information by students. At present, the research community in psychology is divided whether this is possible or not and if the applied methods have really led to a success in learning outcomes or not.

According to Mayer (2009), one key aspect for a successful learning benefit in multimedia is that students would have to process the information actively. This means that students would have to select and structure the information actively from the pictures or text content. However, as such processes take place in working memory, working memory has only limited resources (Sweller, Ayres, & Kalyuga, 2011). Resources which could be depleted by active processing long before deeper processing is stimulated (Mayer, 2009; Kuehl et al., 2014). Therefore according to Kuehl et al. (2014) it is desired to prevent working memory (i.e. working memory capacity) being overloaded. A result that could be achieved by designing the information (in this case multimedia content: text and pictures) to avoid any unwanted excess cognitive load while processing it. Kuehl et al. (2014) points out that such unwanted cognitive excess load could be attributed to perceptual difficulties caused by poorly designed content.

However, there is an emerging theory (disfluency theory) in the research community, which opposes this view. Disfluency theory is applying perceptual difficulties with the aim to trigger a deeper cognitive processing of the perceived content (Alter, Oppenheimer, Epley, & Eyre, 2007; Alter & Oppenheimer, 2009). Disfluency theory is based on the aspects of the DPTs (Kahneman, 2011) as it distinguishes between the two systems: *System 1 & System 2*. According to Alter et al. (2007), whether *System 1 or System 2* is used might depend on how the perceived information is difficult or easy. Disfluency theory basically, states that perceiving disfluent information leads to deeper, effortful and analytic cognitive processing and therefore can lead to higher learning outcomes (Kuehl et al., 2014). Diemand-Yauman et al. (2011) were able to produce higher learning outcomes by introducing perceptual difficulties in their studies. They obscured the to-be-perceived information (e.g. text) by changing the font and font-colour of the text which led to better memorisation of that content compared to a clear presentation (e.g. clear black font and standard font-type) - an effect which is known as *perceptual-interference effect* (Nairne, 1988; Diemand-Yauman et al., 2011; Kuehl et al., 2014). Additionally, the literature also shows that there are similar results achieved by others (Sungkhasettee, Friedman, & Castel, 2011). Many researchers are motivated by the possibilities that the perpetual-interference effect offers. However, others were not able to successfully detect an advantage by obscuring the to be perceived content to introduce disfluency (Rhodes & Castel, 2008). Some researchers discovered that perceptual difficulty lead to disadvantage in memory (Yue, Castel, & Bjork, 2013) or even impairs memory (Glass, 2007). Yue et al. (2013) aimed to find a threshold of when disfluency is exactly, what they called,: a "desirable difficulty" (p. 230) which they were not able to do so. Diemand-Yauman et al. (2011) and Sungkhasettee et al. (2011) however, managed to successfully achieve such "desirable difficulty".

The phenomena, which is known in other instances as *perceptual-interference effect* (Diemand-Yauman et al., 2011; Kuehl et al., 2014) is referred to in this thesis as *Dieman-Yauman, Oppenheimer and Vaughan (DOV) - effect.* The *DOV* - effect states that by introducing disfluecy in the to be perceived information (i.e. obscuring text or numbers in a different font-type or colour) lead to effortful and deeper processing, better memory encoding as well as memory retrieval (Kahneman, 2011). The *DOV* effect can therefore lead to less errors being committed by transcribing the before perceived information. A transcription task can be in the form of text transcription or number transcription in this context. The reason for defining the perceptual-interference effect as *DOV* - effect in this thesis is to stress that the thesis was motivated and inspired by the work of Diemand-Yauman et al. (2011). Additionally, the manipulations used in the following experiments (Chapter Four and Chapter Five) are replications of those used by (Diemand-Yauman et al., 2011).

In summary, compared to traditional methods of making information as fluent and as clear as possible there seems to be evidence in the literature which suggest the counter-intuitive approach of introducing disfluency can lead to higher accuracy than the contrary. This thesis presents a series of studies investigating the *DOV* effect and contrasts the results on previous findings in the field.

2.6 High Level Review of the Research Questions

2.6.1 What design features can be used to ameliorate the effects of human error in entering numbers into digital healthcare devices?

Healthcare is a domain that houses a multitude of devices which by themselves have a diverse range of interfaces. Earlier work by Cauchi et al. (2012) proposed a range of design features which, if applied to a number-entry interface, could analytically lead to fewer errors being committed. The proposed design features included the start position of the cursor, cursor wraparound (i.e. the cursor jumps from 9 to 0), block errors (i.e. the cursor jumps from one side of the display to the other) and many others. According to their analysis some features seem to have a greater effect than others. However, Cauchi et al. (2012) did not evaluate their findings in actual laboratory studies.

The work presented in this thesis aims to provide an answer to this research question in respect to earlier work in the field (Cauchi et al., 2012). Chapter Three will focus on this research question and present three studies aimed to investigate the effect of such traditional interface elements on errors in a number-entry transcription context.

2.6.2 Can the *DOV* be used to ameliorate the effects of human error by reducing the error-rate in transcription tasks?

Slip errors are errors where training and experience is not much of help (Reason, 1990; Back, Blandford, & Curzon, 2007). Slip errors occur in all kinds of situations such as number-entry or general transcription tasks. Yet the psychology literature shows that the way cognitive processes operate can be influenced by certain stimuli to activate a deeper processes which result in better memory encoding and retrieval (Kahneman, 2011; Kuehl et al., 2014).

This thesis was in part inspired by DPT and the previous work of (Diemand-Yauman et al., 2011). Diemand-Yauman et al. (2011) provided evidence that obscuring information in a harder to read format by changing the font-type or colour could have a significant impact on how people memorise the perceived information. Their theory (e.g. cognitive disfluency) is built upon DPT which means that in theory a high level cognitive processing should be stimulated by obscuring the target information. These high-level or deeper processes would then be responsible for a better encoding of the harder-to-read information resulting in higher accuracy once the information is retrieved from memory.

While Diemand-Yauman et al. (2011) successfully demonstrated this effect by obscuring content in a text processing set-up they provided no evidence for number-processing. As previously mentioned (Section 2.3.2) little is known about how people process numbers. Additionally, Diemand-Yauman et al. (2011) also did not conduct their study in an interactive context therefore it was unknown if the same effect could be seen in such a context as number-entry.

Chapter Four shows a study, which successfully replicated Diemand-Yauman et al. (2011) results by transcribing text in a game. Moreover, additional studies are presented which successfully transfer the *DOV* effect from text to number-transcription tasks. Chapter Four and Five present studies which answer this research question and show that the *DOV* effect is an appreciable effect to successfully ameliorate human errors (such as slip errors) in transcription tasks while programming infusion devices.

2.6.3 How does the *DOV* work with a view of moving to real-world applications?

This research question will focus on how the *DOV* works in an interactive context. It will describe the dynamics of the effect based on the evidence presented by the studies in Chapter Three and Four supported by the literature. Moreover, the robustness of the DOV effect will be discussed based on the results in Chapter Five which give a first view of how the effect might behave in a real-world environment where distractions are present.

2.6.4 What are the implications of the DOV effect in the context of transcription tasks?

Transcription tasks are a part of everyone's life. Therefore the implications of the *DOV* effect (i.e. cognitive based interface elements) on such tasks could be generally enormous and also cost-efficient. Many devices or interfaces, for example infusion-pumps or autopilots sometimes provide numerous design variations. Sometimes this can lead to better usability but opens room for errors. A process that the Swiss cheese model illustrates very accurately (Reason, 1990). With continuous advances in technology it is very likely that the healthcare domain will be in a continuous stream of new innovations over the next years, which means more and more devices are flooding the healthcare sector.

The *DOV* effect presents an approach in an interactive context to successfully reduce errors by moving away from traditional device-oriented design features. This means unlike focusing on one feature for a specific device, which most likely will not have the same effect on another device, there is a way to directly influence users in a transcription task to make fewer errors. For example, simply obscuring the information would lead to higher accuracy in situations where people fill out web-forms or their on-line banking transactions. On the flight-deck the *DOV* could be introduced as additional safety-layer where information that has to be entered in certain instruments is then presented in the same obscured way as it was used in this thesis. The better encoding and retrieval of the information in memory may prevent failures of certain safety-layers as presented in the *Air Inter 148* incident. Chapter Six will discuss this research question in light of the studies presented in this work.

2.7 Summary

This thesis is very strongly motivated by healthcare but healthcare itself is a very complex area. For example, it is very difficult to get access to and to get clear understandings. However, this thesis is focusing on a small aspect in healthcare but one with a potentially high impact factor on patient safety namely number-entry and how this processes can be improved to reduce errors.

In number-entry the attribution of an error is clear in terms of cause and effects of a single error. Researchers try to understand number-entry errors in terms of the technology of the device and how people work. When people enter numbers they first form the number in their mind, then approach the device followed by unpacking the number they then use the controls to enter the number and evaluate the results and they press commit. There are factors about the device that could affect how people complete this task but also there is another approach by Diemand-Yauman et al. (2011) which shows that there might be a different way of making progress.

The following Chapter Three will now discuss studies investigating the effect of traditional interface elements on number transcription tasks in a medical device context.

Chapter 3

Feature driven studies on design elements for medical devices

"I have not failed. I have just found 10,000 ways that won't work."
Thomas Alva Edison

3.1 Introduction

5-Key, *Chevron* and *Serial* interfaces are common in healthcare and other safety-critical-domains such as aviation. Because of that reason it is imperative to minimise the errors made by interacting with such device interfaces. The following studies attempt to examine the effects of working with *5-Key* designs in particular building on the analytical work of (Cauchi et al., 2012) where they conducted an analytical formal analysis of many different design options for such interfaces by reverse engineering commercial drug infusion pumps. Cauchi et al. (2012) studied various design features in up to 28 different combinations, which analytically appeared to influence the errors made (some of them more than others). Specifically, they found that design choices such as the cursor-start position consistently affected the errors made with the interface. Moreover, they also suggest that a combination of several of the evaluated design choices could lead to a *"less error-sensitive"* design (Cauchi et al., 2012) (p.35).

This Chapter will discuss three studies which empirically investigate specific design features in the context of number-entry errors based on earlier work by (Cauchi et al., 2012). The first experiment investigated the effect of design features on a *5-Key* interface with the result that it seemed there was no significant difference in the amount of errors committed. This was however influenced by the low error rate, which resulted in an inability to perform empirical tests. However, interesting insights into people's preference for strategies in entering numbers could be detected. It appears that people who use a *5-Key* interface favour transcribing numbers from the left to the right.

The second study provided a better design that provoked more errors and therefore led to empirical tests that confirmed the results from the first study. No significance between the evaluated design features could be detected in terms of error reduction. Moreover, a very low error presentation rate (below 5%) was noted in both studies.

Study Three had the same set-up as the previous studies but took the low error presentation rate into account and included a secondary audio task to increase the workload and therefore the amount of errors made by participants. However, again no significant difference in the evaluated design features could be detected. This raised questions about claims made by previous work of (Cauchi et al., 2012). Surprisingly, again a very low error presentation of 2.3% was detected which was attributed to the secondary task either being to difficult or to the robustness of the *5-Key* design.

3.2 Experiment 1

Inspired by the previous work of (Cauchi et al., 2012) on design choices for *5-Key* interfaces, the following study was designed as the initial experiment to empirically evaluate a set of these design features. This was done to see their effect on the errors that people make while transcribing numbers. The set of design features evaluated in this study included: *Dial*, *Wrap*, *Start Position of the Cursor* and *Block Errors (underflow or overflow)* (see figure 3.1).

Dial

This feature enables each number (or cursor position) to work as an independent dial. If a number is displayed as '9' and again the *up*-key is pressed, the number will then change to '0'. The same will happen if the *down*-key is being pressed when the number displayed is a '0'. The zero will then roll over to '9' but no other cursor position would be influenced (incremented or decremented).

Wrap

The wrap feature (in this case cursor wrap) allows people to move the cursor from the leftmost position by pressing the *left*-key again to the rightmost position and vice versa.

Start Position of the Cursor

This design choice distinguishes between a *left-cursor-start* position where the cursor is presented to the far left of the number-entry display once the device is turned on and the *right-cursor-start* position, which is the exact opposite as the cursor is first presented to the far right of the display.

Block Errors (Overflow or Underflow)

Block errors interrupt the interaction of a user if he or she tries to go beyond a specific boundary. There are many different medical devices i.e. infusion-pump designs used in healthcare and while some ignore the error, once a user entered an unexpected action, (for example pressing the wrong button) and let him or her proceed, some devices block further interaction by sounding a warning alarm (Oladimeji, Thimbleby, Curzon, Iacovides, & Cox, 2012). In this case block errors are blocking an underflow or overflow. For example, an underflow can be caused if a user presses the *down*-key while a '0' is displayed the number will roll over to a '9'. It could be argued that the *Dial* and *Wrap* features are a form of block errors however it was decided to keep these design features defined separately as done in (Cauchi et al., 2012).

Arithmetic

Arithmetic is usually related to underflows or overflows For example, an underflow can be caused if a user presses the *down*-key while a '0' is displayed the number will roll over to a '9' but in this case the number in the cursor position to the left will also be decremented by the value of 1. Additionally, in case a '9' is displayed and the user presses the *up*-key then the number will roll over to '0' but additionally the number at the cursor position to the left will be incremented by '1' (overflow).

The design space available for this *5-key* interfaces is quite vast and differs in combinations of features depending on the manufacturers requirements. Cauchi et al. (2012) were mainly evaluating the features listed above. It should be noted that the "Arithmetic" feature was not included in the studies presented here as the main focus of Study One - Three was to evaluate the claim of the *Start Position of the Cursor* on reducing errors. The feature *arithmetic* would have possibly introduced confounding effects.

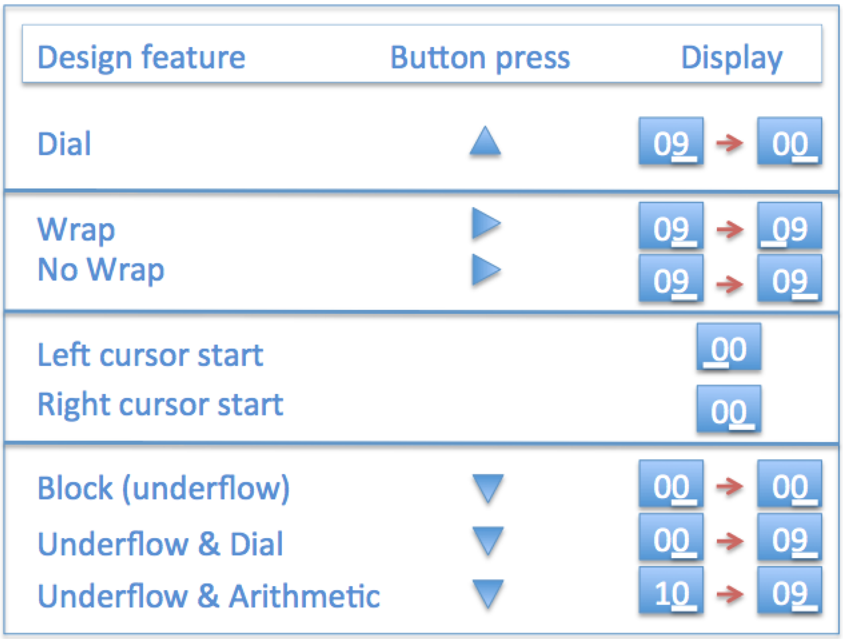


Figure 3.1: Study 1: Design features for 5-Key number-entry interfaces

3.2.1 Aims & Hypothesis

Cauchi et al. (2012) proposed in particular that a *5-Key* interface which enables the following combined features: *block errors, cursor start position to the left, no wrap, dial and block overflow* to analytically let people make less errors than an interface which does not enable this combination. According to Cauchi et al. (2012), apart from *block errors* which have the highest impact on improving *5-Key* designs in respect to errors, a *left-cursor-start* interface design is making the interface more robust to errors (second highest impact on reducing errors) than a *right-cursor-start* interface design in an interactive context. However, they did not confirm their suggestions by evaluating their results in live studies on people interacting with such an interface design. It was the aim of this study to evaluate if the combination of design features can really lead to people committing fewer errors while transcribing numbers. Special attention was given to the *Start Position of the Cursor* design feature as it proposed the second highest impact rate in Cauchi et al. (2012). Therefore the hypotheses for this experiment were:

H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.

H_1 : Errors in the *left-cursor-start* positions will be significantly lower than in the *right-cursor-start* positions. Therefore a *left-cursor-start* design would present a better design in the efforts to reduce number-entry errors as proposed by Cauchi et al. (2012).

3.2.2 Participants

There were 12 participants of whom 7 were men and 5 women. The mean age of all the participants was 27.92 with a *SD* of 6.02 making them between 21 and 33 years old. Participants were mostly students in the department of Computer Science at the University of York except one research associate and one accounts assistant. One participant came from a right-to-left writing country all others came from left-to-right writing countries. Only one participant was left-handed. None of the participants had previous experience of working in the healthcare sector nor used or seen a B.Braun infusion pump before. All participants took part in the experiment after being approached and asked in person if they would participate and each of them was rewarded with chocolate for his or her participation in the study. Participants were randomly assigned to the specific groups.

3.2.3 Number-Transcription Task

The tasks set for participants was to transcribe as many numbers out of a list of 33. These numbers are obtained from logs of syringe pumps collected by Sarah Wiseman, University College London (UCL) (via personal communication). Additionally, the numbers were presented according to the standard of the Institute of Safe Medication Practices Canada (ISMP) as nurses in hospitals would usually see the numbers in this format (ISMP, 2004).

3.2.4 Design

The experimental design was a within-participants, two group design (12 Participants each). The independent variable was the start position of the cursor at initialisation of the interface measured by *left-cursor-start* and *right-cursor-start*. The two dependent variables were the number of total errors committed with special attention to *out-by-ten* errors and the number of total errors committed which measured any deviation of the entered value from the correct value.

An out-by-ten error, as defined by (Thimbleby & Cairns, 2010), is an error where the intended number to be processed is out by the factor of ten. For example, if the intended number is 90, ending up with 9 or less or ending up with 900 or more is an out-by-ten error. Generally speaking, these kind of errors can be easily made by miskeying a zero or a decimal point. In this experimental set-up however it is not possible to miskey a decimal point as the *5-Key* number-entry interface does not provide such an option.

3.2.5 Materials

The instructions as well as an informed consent form were printed out for the participants to read and to sign. Additionally, there was also a brief demographic questionnaire which was also printed and given to the participants to fill out (see Appendix A.1.1 and A.1.2 respectively for the described documents). The computer used was an Apple Macbook Pro with a 13" size display at a 1920 × 1440 resolution. The browser used for the experiment was Google Chrome running under Apple Mac OSX 10.7 (Lion). The interface ran in the browser, the browser window was maximised and the browser controls were disabled from being shown so that only the interface area itself was visible. All other applications on the desktop were closed. Furthermore, the keyboard of the notebook was prepared so that participants could use 5 keys to enter the given numbers. Stickers were put in place on the keys to make it easier for the participants to associate the particular key. The following keys were labelled with stickers: W, A, S, D, F. The key 'W' and was used to increase a number, 'S' and was used to decrease a number, 'A' was used to move the cursor to the left, 'D' was used to move the cursor to the right and the keys 'F' was used to save a number. Each number a participant had to enter, whether it was for the training to get familiar with the controls or the main study, was printed on the front of a paper card. In addition, each card had an unique index printed on its back to be identifiable for the experimenter during the study. All of the cards except those for the pilot were laminated to support a smooth procedure.

Figure 3.2: Study 1: Experimental set-up and interface

3.2.6 5-Key Number-Transcription Interface

The visual design of the *5-Key* number entry interface was carefully created after the original B.Braun infusion pump (see figure 3.2). For the purpose of this study all but five buttons were removed from the image of the interface to avoid any possible unwanted effects on participants. The interface was implemented as a Java server page (JSP) with a logging feature which recorded all user interactions with the virtual device and wrote those interactions into a text file. The interface itself consisted of seven components. A number display, a cursor, four control buttons and an *OK* button. The cursor could be moved with the control keys to the left or to the right and the selected numbers could be increased or decreased by pressing up and down keys. Moreover, the current displayed number could be saved by pressing the *OK*-key. Saving a number cleared and reset the display.

The display itself presented a 5-digit number with two digits after a separating decimal point. Each time when the website was loaded for the first time, all the digit positions were shown as underscores simulating the device start. As soon as any control key was pressed the display showed zeros instead of underscores for the individual number position. Additionally, the logging started and recorded all interactions in a text file. The data logged in the text file had the format of: milliseconds and name of the text file, time-stamp, actual cursor position (activated and or changed), displayed number (changed or saved), button pressed, cursor wrap status message. Figure 3.3 shows an example of a log file.

```
1335182462840_file.txt,Mon Apr 23 13:01:02 BST 2012,cursorposition 7,number: 00000.00 ,keyboard: right,
1335182462840_file.txt,Mon Apr 23 13:01:02 BST 2012,cursorposition 6,number: 00000.00 ,keyboard: right,
1335182462840_file.txt,Mon Apr 23 13:01:03 BST 2012,cursorposition 5,number: 00000.00 ,keyboard: right,
1335182462840_file.txt,Mon Apr 23 13:01:03 BST 2012,cursorposition 4,number: 00000.00 ,keyboard: right,
1335182462840_file.txt,Mon Apr 23 13:01:04 BST 2012,cursorposition 3,number: 00000.00 ,keyboard: up,
1335182462840_file.txt,Mon Apr 23 13:01:04 BST 2012,cursorposition 3,number: 00001.00 ,keyboard: up,
1335182462840_file.txt,Mon Apr 23 13:01:04 BST 2012,cursorposition 3,number: 00002.00 ,keyboard: up,
1335182462840_file.txt,Mon Apr 23 13:01:04 BST 2012,cursorposition 3,number: 00003.00 ,keyboard: up,
1335182462840_file.txt,Mon Apr 23 13:01:05 BST 2012,cursorposition 3,number: 00004.00 ,keyboard: down,
1335182462840_file.txt,Mon Apr 23 13:01:05 BST 2012,cursorposition 3,number: 00003.00 ,keyboard: right,
1335182462840_file.txt,Mon Apr 23 13:01:06 BST 2012,cursorposition 2,number: 00003.00 ,keyboard: right,
1335182462840_file.txt,Mon Apr 23 13:01:07 BST 2012,cursorposition 1,number: 00003.00 ,keyboard: up,
1335182462840_file.txt,Mon Apr 23 13:01:07 BST 2012,cursorposition 1,number: 00003.01 ,keyboard: up,
1335182462840_file.txt,Mon Apr 23 13:01:07 BST 2012,cursorposition 1,number: 00003.02 ,keyboard: up,
1335182462840_file.txt,Mon Apr 23 13:01:08 BST 2012,cursorposition 1,number: 00003.03 ,keyboard: save,
```

Figure 3.3: Study 1: Sample of a log file

For the purpose of this study a flag was implemented in the source code which allowed to change the position of the cursor at initialisation of the interface. With this feature it was possible to prepare the interface for the two-groups design (*left-cursor-start* and *right-cursor-start*).

To avoid unwanted software bugs during the experiment and to exclude possible data loss a thorough test was conducted on both software and hardware before the study took place. This involved a test of all implemented design features i.e. *Wrap* and *Dial*. All key bindings were checked if the logging feature was working properly.

3.2.7 Procedure

The experiment was conducted in an empty office in the Computer Science building. All participants were run in individual sessions, which were managed by personal invitation. Each participant was first welcomed then given a copy of the informed consent form and the instructions to read and to sign. Afterwards participants were asked to become familiar with the use of the keyboard controls and the display of the interface. For this purpose they were given a small set of 6 training numbers which were separated from the main set of numbers for the experiment. The participants were asked to use their index finger only as a way to introduce experimental control and to reduce variation. After a participant became familiar with how the *5-Key* number-entry interface worked, participants were then asked to enter the given numbers, each presented on an individual and randomly ordered card, either in the *left-cursor-start* group or the *right-cursor-start* group within three minutes as quickly and as accurately as possible. Participants were carefully observed which number cards they transcribed. Moreover, the index of the cards was carefully noted down to be able to identify any errors which might occur during the experiment and to be able to detect what are the variance that people offer in relation to the numbers they were asked to transcribe. In order to introduce randomness and to avoid ordering effects with the numbers the set of cards were shuffled at the start of each run. Participants were stopped if he or she could not finish entering all the numbers within the given time limit. After the participant completed the first task he or she was then told that there would be a small change in the interface i.e. an opposite cursor-start position and that the reason for commencing with the second task was about identifying how this small change in the interface design affects peoples ability to complete the task. Each participant was then again asked to become familiar with the different design (opposite cursor-start position) and to enter another six training numbers. Each participant then proceeded to the second task to transcribe another 33 numbers in three minutes in the opposite condition. After completing the number-transcription task each participant was then asked to fill out the demographic questionnaire. The participants were then debriefed as to the goals of the study. Finally, each participant was asked if there was anything they wanted to add about the study, their experience or any general comment. All the participants received chocolate as a reward.

3.2.8 Pilot

A small pilot study was conducted to measure how many numbers a participant could enter in the given time-frame of three minutes. There were three participants and all of them were local research students in Computer Science. A small software issue was resolved. The pilot showed that the original set-up of 33 given numbers and three minutes to enter these numbers was almost too narrow. Participants almost managed to enter all the numbers in time. However, it was decided to keep the amount of 33 numbers for the participants to enter in both conditions (*left-cursor-start* and *right-cursor-start*) with the intention to put some time pressure on them to enter as many numbers as possible. Participants of the pilot were excluded from the main study.

3.2.9 Results

There were in total 518 numbers entered by participants. 284 were entered using the *left-cursor-start* design and 234 were entered using the *right-cursor-start* design. This resulted in 11 errors made in total by the participants. Five of the total errors were made by using the *left-cursor-start* design and 6 errors in total were made by using the *right-cursor-start* design.

The error presentation rate (number of errors divided by number of numbers entered by the participant) was therefore 2% which is too low for traditional measures of systematicity and too low for statistical testing (Abelson, 1995). In any case, the number of errors seems to be comparable at this stage in data gathering.

Two of the five errors in the *left-cursor-start* condition and three of the six errors in the *right-cursor-start* condition were out-by-ten errors. Notably, half of the participants (6 of the 12 participants) made at least one error suggesting that the self imposed time pressure is actually effective in provoking errors across people. Additionally, two participants entered a wrong number but noticed their mistake and entered the correct number again. These two numbers were not counted as errors. Moreover, three of the out-by-ten errors (labelled with a ‘t’ (t = teutonic) in table 3.1) were made by a German participant. This is interesting as it suggests a possible comma to decimal point translation error. Table 3.1 illustrates what kind of errors were made in which condition. The errors were classified according to Wiseman et al. (2011).

Classification	Number shown	Number entered
Wrong number entered	79.6	00000.00
Wrong digit entered	294.12	291.12
Missing digit	97.4	97.00 (occurred twice)
	444.44	444.40
	434.68	434.60
Out-by-ten error	27,955	279.55
	250	25.00
	15,325	15.32 (t)
	2,000	2.00 (t)
	13,750	13.75 (t)

Table 3.1: Study 1: Classification of discovered errors

All 6 of the participants who started with a *left-cursor-start* position entered the numbers from left-to-right. This finding supports the view that the cursor start position (either left or right) is a cue for number entry strategy. However, only one of the participants who started with a *right-cursor-start* position entered the number from right-to-left. The other five participants spotted the *Wrap* function implemented in the interface. Due to this feature they changed their strategy and moved the cursor from the *right-cursor-start* position to the other side of the display to the *left-cursor-start* position before entering each number. Additionally among these five participants who changed their strategy there was one of them who also changed his or her strategy for the following second condition. Meaning that he or she changed the strategy totally opposite of both starting conditions. The specific participant was from China. This is suggestive but not conclusive of a cultural aspect which could have had an influence on this result. Notably, one participant did not even remember that there was a *right-cursor-start* position when the experiment was explained at the end of the study.

Among the participants was one from a right-to-left writing country and one left-handed person. Apparently, the behaviour of the participant from the right-to-left writing country was the same as the behaviour of left-to-right writing participants. Additionally, the behaviour of the left-handed participant was also the same as the behaviour of the right-handed participants.

3.2.10 Discussion

The alternative hypothesis (H_1) for this experiment was that there should be a significant difference in the number of errors using the *left-cursor-start* design than using a *right-cursor-start* design while having design features such as *wrap*, *dial* and *block errors* present. There seems to be no difference in this small sample of *left-cursor-start* and *right-cursor-start* conditions. However, this experiment is limited by the unexpectedly low presentation rate of errors. Despite a large quantitative amount of number being entered the presentation rate was well below what is typically considered systematicity of 5% (Abelson, 1995).

This suggests that there is need for an improved experimental design. A design which would have a bigger sample size of participants and a design that would also have more numbers entered by participants so that the desired discernibility can be achieved.

However, interestingly half of the participants made at least one error. This suggests that even this simple task of entering numbers on a keyboard has the possibility of making errors at this low presentation rate. This suggests that a large number of people who use such devices are susceptible to making errors.

Aside from the main hypothesis, the data gathered is rich. One interesting discovery made were cues for strategy that people use to enter numbers. It seems indicative but not conclusive that people when they see the *right-cursor-start* position have a preference to use the *Wrap* to swap to the opposite start position. Additionally, the result also indicates that *left-cursor-start* does seem to match with the strategy people use to enter numbers from left-to-right. The collected data seem to point in the direction that participants do intend or prefer to keep their strategy for the *left-cursor-start* position. However if people start with a *right-cursor-start* position they continue entering the number from right-to-left unless they realise the *Wrap* feature then they proceed with the *Wrap*.

Future experiments should be designed without this feature as this would have forced participants to use either the left-start or the right-start condition. This may have presented a more clear view on the results.

An interesting aspect is that in the design of infusion pumps it is not possible to get such results as they are either designed in left- or right-cursor-start position. This might further suggest if people use a *right-cursor-start* device they work primarily from the right-to-left unless they can use a cursor *Wrap* feature.

There are also possible cultural differences which may be normal writing style but maybe have an influence on the results due to the use of commas and full-stops in numbers. No apparent difference of right-to-left writing countries as opposed to left-to-right writing countries has been detected. Though there was only one participant from a right-to-left writing country, which may suggest that in future studies this aspect needs to be taken into consideration.

In addition, the given time pressure was probably not enough on its own to provoke errors and future experiments need to be designed to further provoke errors as this experimental design was not particular troublesome for participants to make errors. By designing a task with additional high cognitive load as illustrated in Wiseman et al. (2011) would result in a higher error presentation rate. Additionally, if the small number of participants can be increased to get a larger sample (for example 40 participants) the low error presentation rate would then be acceptable.

Furthermore, a between-subjects design will be considered for future experiments. In which case participants could go for a longer particular task to increase the amount of numbers entered but only in one condition and avoid possible fatigue effects.

Today many medical devices in hospitals such as B.Braun infusion pumps have a touch screen or a semi-touch screen control interface (Braun, 2012). Although, originally aimed to provide participants with a touch screen interaction experience. Developing the *5-Key* interface for use on a Apple iPad was not possible due to the limited knowledge of Objective-C and the short time-frame to conduct the study at that point in time. Although, it was considered to use several open-source software packages[1] which would allow to run JavaScript, CSS and PHP as native application on the iPad none of them were able support Mac OSX 10.7 (Lion) and Xcode 4.3 at that time. However, for future experiments it was considered using such software to be able to provide a touch screen interaction for participants. Additionally, the use of an iPad to access the interface with the iPad browser was considered but a test run showed that, although the browser window was maximised and browser controls were set to non-visible, the interaction with the interface was very difficult. Apple's gesture support of the device made it impossible for someone under time pressure to enter numbers without accidentally zooming in, misplacing the view of the interface or simply changing the interface page to another web page.

As a result of the detected experimental design issues and the low error presentation rate this study is regarded as a trial in context of the following studies. It was the initial study to measure the effect of specific design features on medial device interfaces. The following experiment will address these issues by for example increasing the numbers participants enter for the study to at least the amount of 1000. Two percent error rate in relation to 1000 numbers would be 20 numbers. This amount would be indicative and it might be possible to start seeing patterns and this would allow statistical tests for significance.

Study One has revealed many interesting aspects in number transcription tasks but it has also shown that there is a need to collect further data. As a consequence an additional experiment was run to follow-up. Study Two built on results of Study One and focused on a *5-Key* interface design to further investigate previous claims made by (Cauchi et al., 2012) but used an improved experimental design.

[1] Phonegap http://www.phonegap.com and Appcelerator Titanium http://www.appcelerator.com

3.3 Experiment 2

The rich data gained from the first experiment was indicative but not conclusive as no statistical analysis could be performed. Therefore, Study Two implemented changes in the experimental design. Most significantly the *Wrap* feature was removed as it counteracted the efforts to clearly see what strategies people use when transcribing numbers. Additionally, the design was changed to a between-participants design to allow people in the separate conditions to enter more numbers and to avoid fatigue effects. The amount of numbers to be transcribed and the time to finish the task was also adjusted to: 40 numbers in 4 minutes.

3.3.1 Aims & Hypothesis

The aims of this experiment have not changed since the first study. This second experiment's aim is also to explore the effects of the design features proposed by (Cauchi et al., 2012) particularly if a *left-cursor-start* design for a *5-Key* interface let people commit less errors than a *right-cursor-start* design. For this reason the hypotheses for this study were identical to Study One:

> H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.
>
> H_1 : Errors in the *left-cursor-start* positions will be significantly lower than in the *right-cursor-start* position. Therefore a *left-cursor-start* design would present a better design in the efforts to reduce number-entry errors.

In this context the strategy that people use to enter numbers will also be studied as this may be one reason why errors in the *left-cursor-start* condition are fewer than errors in the *right-cursor-start* condition depending on people's preferences on how they enter the numbers.

3.3.2 Participants

There were 40 participants (14 women). Participants were between 21 and 35 years old and were mostly students in the department of Computer Science except three employees of different companies. Additionally, participants came from a variety of countries such as Japan, UK, China, Malaysia, Latvia, Portugal, Brazil, Singapore, Bulgaria, India, Germany, Costa Rica, France, Poland and Italy. Only two of the participants were left-handed all others were right-handed. All of the participants came from left-to-right writing countries. It should be noted that participants from China and Japan were one would assume that these might be right-to-left writing countries however people still write numbers from left to right from these counties. None of the participants had previous experience of working in the healthcare sector nor used or seen a B.Braun infusion pump before. Each participant was rewarded with chocolate for his or her participation in the study. All participants were randomly assigned to the specific groups.

3.3.3 Number-Transcription Task

The tasks set to participants was in principle identical to the first study as only the amount of numbers participants had to transcribe and the time frame they had available was adjusted in respect to previous results of Study One. Participants were asked to enter as many numbers out of a list of 40 in 4 minutes. The source and the presentation of the numbers used was the same as in Study One.

3.3.4 Design

The design was a between-participants, two groups design (20 participants each). One group used a *left-cursor-start* design and the second one used a *right-cursor-start* design. There were again two dependent variables: out-by-ten errors and any other errors detected. The independent variable was the start position of the cursor at initialisation of the interface as indicated above.

3.3.5 Materials & Number-Entry Interface

The instructions as well as an informed consent document were printed out on paper for the participants to read and sign (see Appendix A.2.1). Additionally, there was also a brief demographic questionnaire which was also given to the participants after finishing the task (see Appendix A.2.2).

The computer system as well as the prepared hardware was identical to the first study. The interface itself was also no different from the first study with the exception that the *Wrap* function was removed. This means that the *block errors* would now block participants from moving the cursor from the far left to the far right of the display and vice versa by pressing one key only. However, the intention to do this action would still be recorded in the log files. The *Dial* design feature was still present as in the first experiment.

Each number a participant had to enter, whether it was for the training to get familiar with the controls or the main study, was again printed on the front of a paper card. In addition, each card had an unique index on its back to make it identifiable for the later analysis. There were again two versions of the interface used in the study. One where the cursor-start position of the interface was on the left another one where the cursor-start position was on the right.

3.3.6 Procedure

All participants were run in individual sessions in the same room as Study One to avoid any environmental effects. Each participant was first welcomed then given a copy of the informed consent form and the instructions to read and to sign. Afterwards they were asked to become familiar with the use of the keyboard controls and the display of the interface. For this purpose they were given a small set of 6 training numbers which were separate from the main number set of the experiment. The participants were again asked only to use their index finger to enter the numbers. After a participant became familiar with how the *5-Key* number entry interface worked he or she was then asked to proceed to the main study. Participants were then asked to enter the given 40 numbers either in the *left-cursor-start* group or *right-cursor-start* group within 4 minutes as quickly and as accurately as possible. The participants were carefully observed which number on which card they transcribed. Moreover, the index of the cards was carefully noted down to be able to identify any errors which might occur during the experiment. Participants were made aware of the remaining time after two minutes. This was done to further increase the pressure on the participants and to increase errors. Participants were stopped if he or she could not finish transcribing all the numbers within the given time limit. Each participant was afterwards asked to fill out a small demographic questionnaire. The participants were then debriefed as to the goals of the study. Finally, each participant was asked if there is anything they wanted to add about the study, their experience or any general comment. Each participant received chocolate as a reward similar to Study One.

3.3.7 Pilot

A pilot study was conducted to measure how many numbers a participant could enter in the extended time-frame of 4 Minutes. There were two participants both of them research students in Computer Science at the University of York. The same effect as in Study One was detected in terms of the amount of numbers participants had to transcribe. It was decided to keep this design as it proposed an ideal set-up.

3.3.8 Results

The study had originally 40 participants but one participant was excluded from the statistical analysis as the person made 21 errors out of 26 entered numbers and those were systematically wrong in the same way. It can be argued that this person did not understand the interface.

There were in total 1331 numbers entered by 39 participants. 669 numbers were entered by the using the *left-cursor-start* design and 662 numbers by using the *right-cursor-start* design. The mean of the numbers entered in the left-cursor start group was 35.2 ($SD = 3.4$) whereas in the right-cursor start group 33.1 ($SD = 3.9$). A *two-tailed*

Mann-Whitney W test revealed that there was at most a marginal effect in terms of total numbers entered, $W = 256; p = 0.06$ (see figure 3.4).

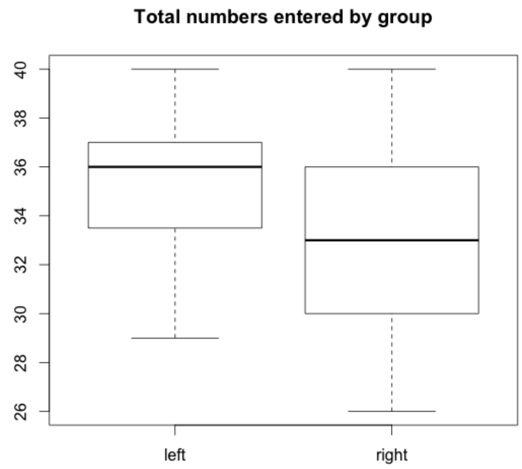


Figure 3.4: Study 2: Boxplot of total numbers entered in each condition

This resulted in 38 errors made in total by the participants. 16 of the total errors were made by using the *left-cursor-start* design and 22 errors in total were made by using the *right-cursor-start* design (see figure 3.5). The mean of the numbers entered in the left-cursor start group was 0.84 ($SD = 0.76$) whereas in the right-cursor start group 1.1 ($SD = 0.9$).

The error presentation rate in this study was 2.9%, which is low but persistent in light of the previous study. In order to get a more accurate measure of a possible difference in the groups the error-rate was also tested for significance. The mean of the error rate in the left-cursor start group was 0.02 ($SD = 0.02$) whereas in the right-cursor start group 0.03 ($SD = 0.03$).As errors are not normal distributed a non-parametric *Mann-Whitney* test was conducted on the total number of errors made in both conditions with the result of $p = 0.4; W = 136.5$ which showed no significant difference between the two designs (see figure 3.6).

Notably, 69% of the participants (27 of 39 people) made at least one error suggesting that the self-imposed time pressure is actually effective in provoking errors across

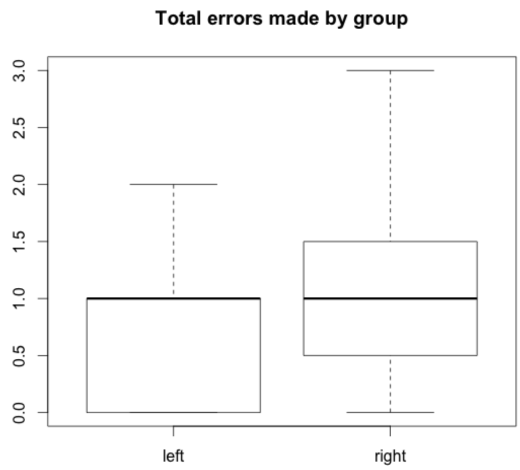


Figure 3.5: Study 2: Boxplot of total errors made in each condition

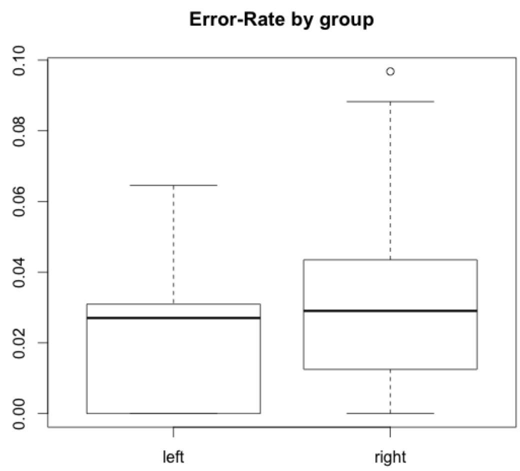


Figure 3.6: Study 2: Boxplot of the error-rate per condition

people. Additionally, three participants entered a wrong number but noticed their mistake and entered the correct number again. These numbers were not counted as errors. Moreover, three of the out-by-ten errors (labelled with a ‘c’ in table 3.5) were made by participants who came from a country where a comma is used as decimal point.

Further to out-by-ten errors, all errors were considered. Table 3.2 illustrates what kind of errors were detected and classified according to Wiseman et al. (2011). Twelve other errors were made on the left design and 16 other errors were recorded on the right design. The classification indicated with a ‘*’ matches with the taxonomy of Wiseman et al. (2011). The label ‘L’ indicated the *left-cursor-start* design and label ‘R’ indicated a *right-cursor-start* design. Additionally, ‘x2’ indicates that the error occurred twice.

All but two of the participants used a number-entry strategy that depended only on the cursor-start position. The other two participants used a *right-cursor-start* design and started entering numbers from right-to-left then changed their strategy by moving the cursor to the left and entering the rest of the number from left-to-right. This finding supports the view that the cursor-start position is a cue for number entry strategy. As previously mentioned in Study One, an interesting aspect is that in the design of infusion pumps it is not possible to get such results as they are either designed in left or right cursor-start position. This might further suggest if people use a *left-cursor-start* device they work primarily from left-to-right.

Classification	Number shown	Number entered	Position
*Out-by-ten errors:	27,955	2795.50	R
	1111.11	11111.11	R
	0.06	0.60	L
	99.9	0.99	R
	99.9	9.90	R
	2497.5	249.75	R
	27,955	00279.55 (c)	R
	1017.67	10710.67 (c)	L
	4166.67	41660.67 (c)	L
	13750	1375.00	L
*Early termination:	3.03	0.03	R
	962.79	002.79	R
*Transposition error:	294.40	294.04	R
*Wrong digit entered:	757.58	755.58	L
	434.48	434.68 (x2)	R
	434.48	434.28	L
	434.48	432.28	R
	20.67	20.57	L
	2497.5	2497.4	L
	714.29	713.29	L
	27,955	29,955	L
	757.58	757.68	L
	2497.5	2497.1	R
	500	600	R
	0.08	0.09	R
*Extra digit:	99.9	99.99	R
*Missing digit:	962.79	962.00	R
	555.56	555.50	R
	20.67	20.00	L
	294.12	294.00	L
	892.86	892.80 (x2)	R
	294.12	294.10	L
	1071.67	1071.00	L
	3.03	3.00	L
	0.49	0.40	L
	294.12	294.10	L

Table 3.2: Study 2: List of errors found

3.3.9 Discussion

The alternative hypothesis (H_1) for this experiment was that there is a significant difference in the number of errors made using the *left-cursor-start* design than using a *right-cursor-start* design with a particular view on out-by-ten errors. There seems to be no difference in this small sample of *left-cursor-start* and *right-cursor-start* conditions. However, this experiment is also limited by the unexpectedly low presentation rate of errors. Despite a large amount of numbers being entered the presentation rate was again well below what is typically considered systematicity, namely 5%.

This suggests that the applied changes to provoke more errors did not work sufficiently. Perhaps, a secondary task which increases cognitive load and a further increase of the amount of numbers which participants need to enter would be an appropriate next step so that the desired discernibility can be achieved.

However, notably 69% of the participants (27 people) made at least one error. This suggests that even this simple task of entering numbers on a keyboard has the possibility of making errors at this low presentation rate. This suggests that similar to Study One, a large number of people who use such devices are susceptible to making errors. However, the given time pressure was probably still not enough on its own to provoke further errors.

Aside from the null hypothesis, which could not be rejected, the data gathered is again rich. One interesting discovery made is that the out-by-ten errors might indicate that participants do not know how to position the numbers in relation to the decimal point. This could be specific due to the display. Additionally, at first the list of missing digits suggested that all of them were caused by participants using the *left-cursor-start* design. If true that would show that people stopped entering numbers at some point. However, as table 3.2 shows there are also missing digit errors which are related to the *right-cursor-start* design.

Additionally, the results also indicates that the cursor position does seem to match with the strategy people use to enter numbers from left-to-right or from right-to-left.

The collected data seem to point in the direction that participants do intend or prefer to keep their strategy for both conditions however if people start with a *right-cursor-start* position they may change entering the number from left to right this is indicative but not conclusive. This is reflected in terms of the numbers people entered as it appears that they were slowing down when using a right-cursor start design (although not significantly). This would match with previous results as people's strategy to enter numbers might come in play and account for the time difference. Moreover, this can also be seen in the number of missed inputs i.e. the numbers participants were not able to enter in time. In the left cursor-start 91 numbers were missed from 760 possible numbers and in the right-cursor 138 were missed out of 800 possible ones. That is roughly 5 missed numbers in the left and 7 missed numbers in the right condition.

Although this study could not detect a significant difference between the cursor-start designs and their effect on reducing errors many interesting discoveries have been revealed. Thimbleby and Cairns (2010) already pointed out that number entry seems at first a trivial task but still people make errors frequently. This study has shown that 69% of the people participating in this experiment made at least one error while entering numbers in a *5-Key* number entry interface for a medical device. The *start position of the cursor* design feature examined in this study is just one out of many. However, it is clearly not conclusively demonstrated so there is need to conduct further research just simply on where the cursor starts. Additionally, Cauchi et al. (2012) suggested a variety of design features for number entry devices clearly favouring further studies in this field. In conclusion, reducing errors in interactions with interfaces such as the one used on the B.Braun infusion pump indeed depict a fruitful and important area of research. Even a low error rate as shown in Study One and Study Two translated into real world operations still depicts severe safety risks for domains where accuracy is of great importance. Yet the two studies presented in this Chapter were not clearly able to conclusively evaluate previous work in the field. In fact the studies show there is clearly a need to empirically explore and collect further evidence for the effect of these design features in future experiments.

3.4 Experiment 3

This experiment has been designed to further examine how users interact with a *5-Key* interface, in particular, what kind of errors people make while using such an interface. Cauchi et al. (2012) made predictions about the design features in their work, which Study One and Study Two (the studies included many of the same features) evaluated. The results reflect some of the behaviours they predict but there are differences in behaviour that were not seen. It could be that what Cauchi et al. (2012) predict is indeed a smaller effect than what they expected. Another reason why Study One and Two were unable to conclusively confirm their results could be the interface layout. For example, the arrangement of the keys in the above mentioned studies was disadvantageous as a keyboard but not a real device nor an identical touch screen was used. It could also be that the environment is responsible where one would see such errors more clearly. In many cases people who work in healthcare or aviation (nurses, pilots) are influenced by the environment in which they operate. A situation that is best described by a review of the *Air Inter Flight 148* incident. People can be confronted with an increased task-load (as did the Captain of *Air Inter 148* once he realised that he had to manually land the aircraft) or distracted (for example the Co-Pilot refocusing the attention of the Captain in a critical moment that could have saved them from disaster) during transcription tasks leading them to, for example, interleave the task and subsequently increase the number of errors (Back et al., 2012).

For this reason, the following experiment was conducted and included measures to (i) increase the error-rate, (ii) align the interface properly and (iii) provide a more closely related touch-based interaction.

3.4.1 Aims & Hypothesis

This Study has the aim to investigate if indeed the effect that Cauchi et al. (2012) propose is a smaller effect than expected. For this reason, a secondary task with the sole purpose to increase the number of errors is introduced - a technique that is commonly used when studying workload effects (Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Wu & Li, 2013). This secondary task is done in two levels (grade of difficulty

Low and *High*). Apart from the secondary task Study Three evaluated the same manipulations i.e. the cursor-start positions as the previous studies. This resulted in four experimental conditions (*Left-Low, Left-High, Right-Low and Right-High*) leading to the following hypotheses:

H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.

H_1 : Errors in the *left-cursor-start* positions will be significantly lower than in the *right-cursor-start* positions. Therefore a *left-cursor-start* design is better than a *right-cursor-start* design in reducing errors as proposed by Cauchi et al. (2012).

H_2 : The secondary task will be responsible for a significant increase of the errors made in both audio conditions *High* and *Low*.

In this context the strategy that people use to enter numbers will also be studied as this may be one reason why there are fewer *left-cursor-start* errors than *right-cursor-start* errors depending on peoples' preferences how they enter the numbers.

3.4.2 Participants

Forty participants (10 women) took part in this study (10 per condition). The mean age of all the participants was 23.95 with a *SD* of 3.69. All but four participants were students of the University of York. None of the participants was familiar with an infusion pump and only 11 stated that they were familiar with a *5-Key* interface. All participants had normal hearing and vision and none of them stated that they had never used an iPad or touch based device before. Interestingly, 17 participants stated that they spoke the numbers on their mind while entering them into the tool (five in *Left-Low*, four in *Right-Low*, five in *Right-High* and nine in *Left-High*). Participants were again randomly assigned to the specific groups.

3.4.3 Number-Transcription Task

The task in terms of transcribing numbers was the same as in Study One and Study Two. The experiment also included a dual task for participants to complete at the same time. The *primary task* set for participants was to enter as many numbers out of a list of 60 in a limited time-frame of six minutes. These numbers were again obtained from logs of syringe pumps (Wiseman, 2011) and presented according to the ISMP standard.

The *secondary task* set for participants was to listen to an audio track while entering the numbers at the same time. The audio track had two levels *Low* and *High*. In the *Low* level participants were asked to listen to letters and count the appearance of the letter 'A'. In the *High* level participants were asked to listen to the same audio track but instead count the number of all appearing vowels (A,E,I,O,U). The letters appeared in an interval of 1.2 seconds. Table 3.3 presents an overview of the total number of appearances of a vowel in both audio streams (*Low* and *High*). Overall there were 650 letters in the audio track (109 vowels and 541 other letters from the alphabet). The order of the letters was randomised so that no pattern of could be identified.

Letter	Number of occurrence
A	20
E	15
I	12
O	30
U	32
Total	109

Table 3.3: Study 3: Frequencies of appearing letters

The secondary task was included in this study solely as means to increase the rate of appearing errors by increasing the mental workload of participants through an auditory stimuli. This workload was manipulated by setting the difficulty of the audio task from *Low* to *High*.

3.4.4 Design

The experiment had a four-condition between-subjects design. Each condition had 10 participants. The conditions were *Left-Low*, *Right-Low*, *Right-High* and *Left-High*.

This experiment had two independent variables. One was the start position of the cursor at initialisation of the interface (either a *left-cursor-start* or a *right-cursor-start* position) and the difficulty of the audio condition (*Low* and *High*).

Left-Low

Participants were told to transcribe all the numbers into the interface as fast as possible using a *left-cursor-start* position. In the secondary task, in which they had to perform at the same time, participants had to listen to an audio track and count the number of appearances of the letter 'A'.

Left-High

Participants were told to transcribe all the numbers into the interface as fast as possible using a *left-cursor-start* position. In the secondary task, in which they had to perform at the same time, participants had to listen to an audio track and count the number of appearances of any vowel.

Right-Low

Participants were told to transcribe all the numbers into the interface as fast as possible using a *right-cursor-start* position. In the secondary task, in which they had to perform at the same time, participants had to listen to an audio track and count the number of appearances of the letter 'A'.

Right-High

Participants were told to transcribe all the numbers into the interface as fast as possible using a *right-cursor-start* position. In the secondary task, in which they had to perform at the same time, participants had to listen to an audio track and count the number of appearances of any vowel.

There were five dependent variables which were measured in this Study: The total number of errors, the total numbers entered, the total error-rate, the number of completed audio task (Low) and the number of the completed audio task (High).

To avoid familiarity effects with the numbers the set of cards was shuffled each time at the start of a run to introduce randomness in the order of numbers.

3.4.5 Materials & Number-Entry Interface

An informed consent document containing instructions were printed out on paper for the participants to read and to sign (see Appendix A.3.1 & A.3.2) before the study. Additionally, there was also a brief demographic on-line questionnaire which was also given to the participants after finishing the task (see Appendix A.3.3).

Each participant was given 60 laminated cards containing a number which they would have to transcribe into the interface. These laminated number cards had a reference number on their back so that they could be identified. The reference numbers of the laminated cards were noted down and the time was recorded by using a smart-phone.

An Apple iPad Mini 2nd generation running iOS 8.2 was used for the study (as shown in figure 3.7). Additionally, a Macbook Pro 2015 (MacOSX 10.10.3) was used to manage the audio stream.

The audio file used for this study was created with MS Word and Apple's natural Samantha voice package. The file listed a single letter per line and could be read by the computer voice via an activation hot-key. A Macbook Pro 2015 (MacOSX 10.10.3) was used to manage the audio stream.

Apart from previous studies this experiment used an implementation of the B.Braun Infusomat Space interface on the iPad Mini 2nd generation (retina display). The implementation was done in Xcode 6 and the source code can be found at the autor's website (`http://cs.york.ac.uk/~feynman`).

Figure 3.7: Study 3: Experimental Setup - (left-cursor-start)

The visual design of the *5-Key* number-entry interface consisted as in the previous experiments of seven components (number display, cursor, four control buttons and the OK button). Pressing the *OK-key* would save the current number into a text log file and store it on the iPad. The files were later transferred for the analysis. The interaction behaviour of the iPad version was identical to the previous keyboard version.

The log files were identically structured as in Study One & Study Two. There were also two versions of the interface used in the study. One where the cursor-start position of the interface was on the far left another one where the cursor-start position was on the far right of the display.

3.4.6 Procedure

The experiment was conducted in a controlled office environment in either the department of computer science or at the Morell Library in York. Every participant was run in an individual session and participants were invited to take part in the study by personal invitation. Every participant was welcomed and given a copy of the informed consent form and the instructions to read and sign. After a participant had signed the form they were asked to become familiar with the use of the iPad and the display of the interface. For this purpose they were given a small set of three training numbers which were separate from the main number set of the experiment. The numbers were presented as laminated cards which have the number to be entered at the front and an index displayed at the back to keep track of the order of the entered numbers. All numbers printed on the cards were again displayed in the ISMP standard. The participants were asked only to use their index finger to enter the numbers. Depending in the audio condition a participant was in, they had to perform the trials while listening to the audio track and count the number of A's or vowels. For this they were asked to wear headphones during the experiment. After the training runs the participant was then asked to proceed to the main study. Participants were asked to enter as many numbers as possible out of a list of 60 numbers either in the *Left-Low*, *Left-High*, *Right-Low* or *Right-High* group within six minutes as quickly and as accurately as possible. The experimenter started the audio track via hot-key on the Macbook. Participants were carefully observed on which number they entered. The index of the cards was then noted down to be able to identify any errors which might occur during the experiment. The experimenter stopped the participant if he or she could not finish entering all the numbers within the given time limit. After finishing the number entry task, participants were asked to fill out a small demographic questionnaire where they had to state the number of correctly remembered A's or vowels. The correct answer would have been either 20 A's or 109 vowels. The participants were then debriefed as to the goals of the study. At the end, each participant was asked if there is anything they want to tell the experimenter about the study, their experience or any general comment.

3.4.7 Study Pilots

The study was piloted which revealed no major issues. The two pilot participants (excluded from the main study) confirmed that the number range of 60 numbers in six minutes was a fitting choice and that the set-up of the audio streams was was working accordingly.

Additionally, a pilot of the secondary task was performed with three additional participants (also excluded from the main study). All three participants were asked to first count the appearance of A's and then the appearance of the vowels. In terms of A's participants performed well (19, 20, 20) and but in terms of the vowels performance varied (100, 80, 126). However, the task set-up reflected the difficulty levels and therefore it was decided to take them forward into the main study.

3.4.8 Results

The data collected in this Study was analysed using a (*2 X 2*) Analysis of Variance (ANOVA), as there is no well established test for non-parametric data in this context. Main effects where they are seen in the ANOVA are confirmed with a non-parametric *Mann-Whitney* test. An alpha level of .05 was used for all statistical tests. Additionally, the Adjusted Rank Transformation (ART) test by Sawilowsky (1990) was used to investigate non-parametric interaction effects. Effect-sizes (partial eta-squared η_p^2 are stated in the relevant sections.

Overall, there were in total 1214 numbers entered by participants. 361 numbers were entered in the *Left-Low* condition, 293 were entered in the *Left-High* condition, 295 in the *Right-Low* condition and 265 numbers were entered in the *Right-High* condition. Table 3.4 shows the relevant descriptives in relation to the total numbers entered in each condition.

		Cursor-Start		
		Left	Right	Row Mean
Audio	Low	36.1 (10.96)	29.5 (6.52)	32.8 (9.41)
	High	29.3 (10.19)	26.5 (8.98)	27.9 (9.47)
	Column Mean	32.7 (10.88)	28.0 (7.81)	

Table 3.4: Study 3: Descriptives: Mean (SD) for the total number entered

In regards to the total numbers entered, there was no main effect of the Cursor-Start, $F(1,36) = 2.54; p = 0.12, \eta_p^2 = 0.07$. There was also no main effect for Audio $F(1,36) = 2.76; p = 0.11, \eta_p^2 = 0.07$ and also no interaction effect (Cursor-Start X Audio) $F(1,36) = 0.42; p = 0.52, \eta_p^2 = 0.01$ (see figure 3.8).

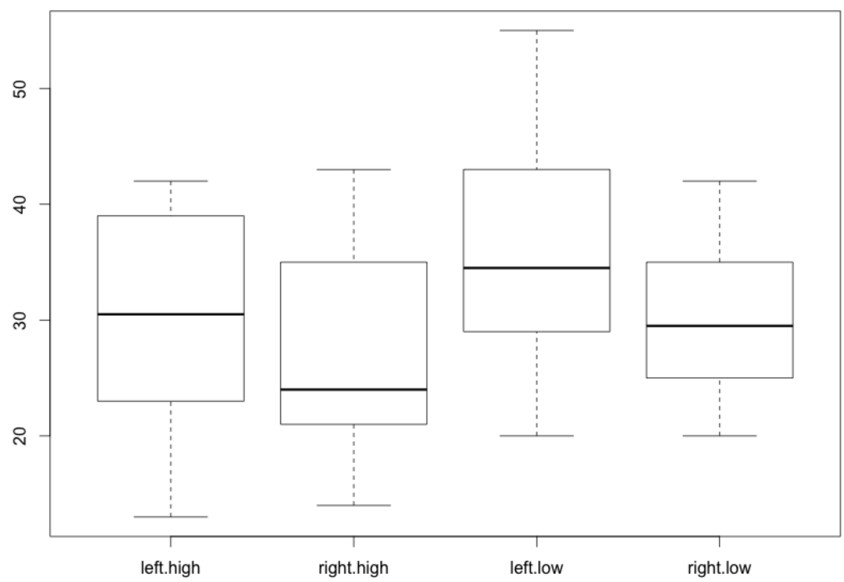


Figure 3.8: Study 3: Boxplot of total numbers entered by each participant in each condition.

There were 27 overall errors made in this study which relates to an error-rate of 2.25%. Overall 12 errors were made in the *High* conditions, 15 in the *Low* conditions, 12 in the *left-cursor-start* positions and 14 in the *right-cursor-start* positions. The detailed error frequencies are: 5 errors in *Right-Low*, 10 errors in *Left-Low*, 3 errors in

Left-High and 9 errors in *Right-High*. Table 3.5 shows the descriptives for the errors made in this study by conditions.

		Cursor-Start		
		Left	Right	Row Mean
Audio	Low	1.0 (1.15)	0.5 (0.53)	0.8 (0.91)
	High	0.3 (0.48)	0.9 (1.19)	0.6 (0.94)
	Column Mean	0.7 (0.93)	0.7 (0.92)	

Table 3.5: Study 3: Descriptives: Mean (SD) for the total errors made

In terms of total errors made, there was no main effect of the Cursor-Start, $F(1,36) = 0.03; p = 0.86, \eta_p^2 = 0.01$. There was also no main effect for Audio $F(1,36) = 0.28; p = 0.60, \eta_p^2 = 0.01$ but a marginal interaction effect (Cursor-Start X Audio) $F(1,36) = 3.69; p = 0.06, \eta_p^2 = 0.10$ (see figure 3.9). The marginal effect was however not confirmed by the ART $p = 0.23$.

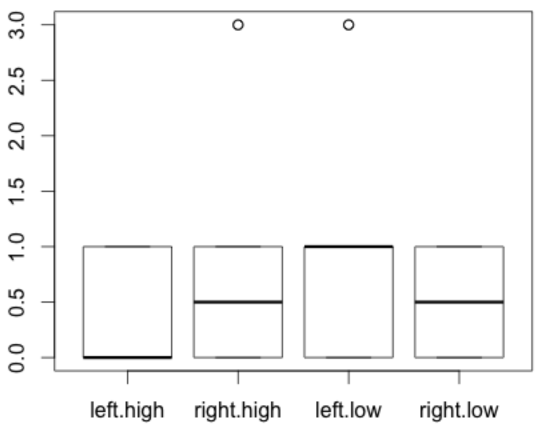


Figure 3.9: Study 3: Boxplot of total errors made by each participant in each condition.

Twenty-one participants made no errors during the entire study. Four participants made no errors in the *Left-Low* condition, five participants made no errors in the *Right-Low* and *Right-High* condition and seven participants made no errors in the *Left-High* condition (49% of all people made at least one error). The highest error-count was recorded in *Left-Low* and *Right-High* (3 errors in both conditions). As in the previous studies out-by-ten errors were of particular interest. Yet, as table 3.6 shows there were very few of these.

Classification	**Number shown**	**Number entered**	**Condition**
Wrong number	0.02	00000.00	left-high
	68	89	left-low
Out-by-ten	27,955	27.95	left-low
	8,333	8.33	left-low
	13,750	13.75	left-low
	27,955	279.00	right-low
	0.49	49.00	right-high
Extra digit	303.03	3303.03	right-high
	250	250.03	right-low
Disposition	77.60	77.06	right-high
Early termination	222.22	222.00	left-low
	806.73	806.00	left-low
	0.49	0.40	right-low
	977.78	977.00	right-high
	1071.67	1071.00	right-high
	608.13	608.00	left-high
Wrong Digit	608.13	606.13	left-low
	2497.50	2498.50	left-low
	4166.67	4166.69	left-low
	30	20	left-low
	20.67	20.66	right-high
	291.12	299.12	right-high
	608.13	608.12	right-high
	79.60	79.90	left-high
	27	28	left-high
	249.4	249.9	left-high

Table 3.6: Study 3: Classification of discovered errors

In regards to the error-rate, there was no main effect of the Cursor-Start, $F(1, 36) = 0.46; p = 0.50, \eta_p^2 = 0.01$. There was also no main effect for Audio $F(1, 36) = 0.02; p = 0.90, \eta_p^2 < 0.01$ but again a (at most) marginal interaction effect (Cursor-Start X Audio) $F(1, 36) = 2.95; p = 0.09, \eta_p^2 = 0.08$ (see figure 3.10). Again this was tested with the non-parametric ART and revealed no significance level $p = 0.36$. Table 3.7 shows the descriptives of the total error-rate.

		Cursor-Start		
		Left	Right	Row Mean
Audio	Low	0.03 (0.05)	0.02 (0.02)	0.03 (0.03)
	High	0.01 (0.02)	0.04 (0.06)	0.03 (0.05)
	Column Mean	0.02 (0.04)	0.03 (0.05)	

Table 3.7: Study 3: Descriptives: Mean (SD) for the total error-rate

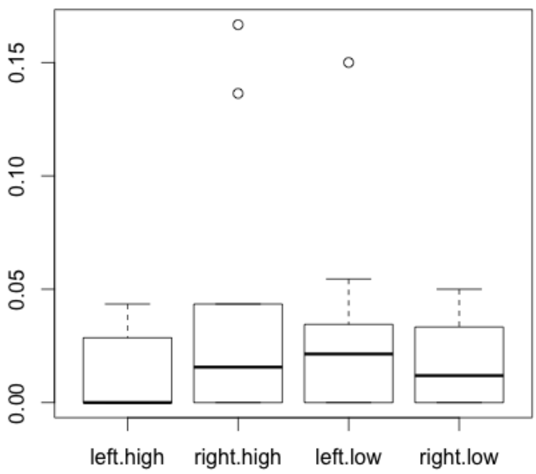


Figure 3.10: Study 3: Boxplot of total error rate by each participant in each condition.

By looking at the dependent variables regarding the audio conditions (Low and High) it was noted that only two participants managed to count the correct number of A's in the low conditions one in *Right-Low* and another one in *Left-Low* but about nine more got very close and stated a range between 16-22. None of the participants in the high conditions was able to tell the correct number of vowels and only one came in a close rage. The worst count was made by a participant in the *Right-High* condition (287 vowels), however this participant did not make any number-entry errors. A table listing the secondary data can be seen in Appendix A.3.4

Additionally, in terms of strategies people used to enter the numbers it was noted that all participants in the *left-cursor-start* conditions entered the numbers from the left to the right but 13 of the 20 people in the *right-cursor-start* positions moved the cursor first to the left and then entered the numbers from left to right (nine participants in *Right-High* and four in the *Right-Low* condition).

3.4.9 Discussion

The results of Study Three showed no significant difference of the total numbers entered, total errors made or the error-rate in terms of the cursor-start manipulation. It was expected that participants would make significantly fewer errors in the *left-cursor-start* positions rather than in the *right-cursor-start* conditions which was not the case. Moreover, there was also no significant difference in terms of the secondary tasks on task-load *High* and *Low*. Therefore the null hypothesis (H_0) could not be rejected.

In terms of total numbers, people entered more numbers in the *left-cursor-start* than in the *right-cursor-start* designs but not significantly more. A possible reason for fewer numbers in the right design could be because a lot of people (13) entered the numbers from the left to the right meaning they first moved the cursor over to the other position and therefore lost valuable time to enter more numbers. This suggests that entering the numbers from left to right was slightly easier for participants. A fact that is reflected in terms of cues for strategy that people use to enter numbers. The data showed that when participants see the *right-cursor-start* position they have a preference to move the cursor first to the left and enter the numbers from the left to the right because it is not very hard to do. This is a really interesting behavioural pattern that is reflected in previous studies One and Two.

Despite a large quantitative amount of numbers being entered the error presentation rate was again well below a systematicity of 5%. The low error presentation rate in relation to studies One and Two speaks for a certain robustness of the *5-Key* design. Additionally, half of the participants made at least one error indicating that the task of entering numbers via touch-screen (iPad) has the possibility of making errors at this low presentation rate. This also suggests that a large number of people who use such devices are susceptible to making errors.

In terms of total errors made, no significance could be detected. However, there was at first a marginal interaction effect but this effect resulted from a parametric test and could not be confirmed by the more accurate non-parametric ART test. Moreover, as there are slight variations across the total number of numbers entered, the error-rate would be a more accurate measure. Yet again in terms of error-rate no main effect for cursor-start or task-load could be detected. However, as with the total errors made there was (at most) a marginal interaction effect though the ART could also not confirm this result. Although not significant, more errors were made in the left conditions than in the right conditions which adds to the argument that previous results proposed by Cauchi et al. (2012) are not coming through. It is interesting that the majority of what is considered to be the most severe errors (out-by-ten) were made in the *left-low* condition (3 out of 5) which is surprising but indicative. Table 3.6 shows the different errors that were made during the experiment classified according to Wiseman et al. (2011). Early termination errors and wrong digit errors occur across conditions indicating possible encoding errors perhaps due to the interference of the counting task. Errors which indicate a possible motor-movement error (transposition errors) only occurred in such little quantity that interpretation is again highly speculative.

No effects were seen In terms of task-load (*Low & High*). People in the *High* conditions enter fewer numbers than people in the *Low* conditions although this was not significant. Maybe the secondary task is interfering with their ability to enter numbers. During the dual-task participants are entering a similar amount of numbers indicating not a huge difference but this is perhaps somewhat interfering with their ability to just get on with the task indicating a small but not significant effect. This is indicated by looking at figure 3.8. In view of the errors, people made more errors in the *Low* conditions than in the *High* conditions which was not expected although again not significant. It may be that the secondary task is interfering in the *Low* condition but was perhaps too strong in the *High* condition. A lot of people (9) in the *Low* condition came very close in the counting task. This supports the argument that people engage with the secondary task and therefore made more errors in the *Low* condition. The *High* condition was perhaps too difficult in combination with the primary task and participants were less engaged in this condition. Although the secondary task can be seen as simple counting task it can be assumed that the students were not heavily trained and therefore experienced higher cognitive load which may have contributed to an even higher load in the *High* condition.

By looking at all the studies presented in this Chapter, people made very few errors (as before in Study One and Two) even with an extra task they made very few errors in the *5-Key* interface. Firstly, that suggests that the *5-Key* design is really robust in terms of errors. Secondly, the rate of occurrence of the errors is far too small. There may be some effect but it is all too small to see. Regardless what Cauchi et al. (2012) said the actual incidence of errors are small and really hard to pick up. There may be some effects and these effects maybe due to different strategies of number-entry with the different device interface but on the whole it is not clear. It looks as if there is a real dominance on one strategy over the other but in many practical situations it is far too small to make a difference. The errors occur in very small numbers which is problematic to have confidence in.

3.5 Overall Discussion & Conclusions

Cauchi et al. (2012), claimed that specific design features, if enabled, can lead to a significantly lower error-rate when using a *5-Key* number-entry interface. In their work, they discovered a constant effect of different design aspects for *5-Key* interfaces by introducing their competitive formal analysis method. For example, they claim that a *5-Key* interface, which enables block error, seems less sensitive to errors than an interface, which does not enable this feature. Moreover, the starting position of the cursor seems to play a very important error-reducing role. According to Cauchi et al. (2012), across all the different combinations of features that they present, whenever a *left-cursor-start* design is contrasted to a *right-cursor-start* a difference in the error-rate can be seen. In other words fewer errors in the *left-cursor-start* would result in significantly fewer errors in a *5-Key* interface. No such difference could be found in the studies presented here.

In detail, Cauchi et al. (2012) discovered some discrepancies in their analysis which lead them to realise that small changes in features can have a substantial effect. They then isolated these features (and combinations of features). As a conclusion Cauchi et al. (2012) recommended to evaluate the discrepancies via empirical experimentation and this is exactly what this Chapter aimed to do. Chapter Three aims to provide empirical evidence for and against their analysis. The experiments presented here use exactly what Cauchi et al. (2012), said in their analysis would have an effect yet no such evidence could be found. A summary of the studies is presented in table 3.8.

Study One was the first attempt and was therefore seen as a trial study to investigate findings by Cauchi et al. (2012) in this thesis. It was a within-participants design and had 12 participants in total. The manipulation was the cursor-start position (either left or right). There was no evidence of a significant difference between the conditions as in this case the study lacked sufficient error data for statistical testing.

Study Two had the same set-up but 40 participants and a between-participants design. However, again no evidence for a significant difference between the cursor-start positions was found. This raised questions about the claims of previous researchers. However, what was seen in both studies was a very low error presentation rate of 2% and 2.9% respectively. It was decided to further investigate the evaluated design features by taking into account measures to increase the errors people make.

Study Three had the same participant size and set-up as Study Two apart from taking into account a secondary task as further measure to increase the error-rate. A procedure widely used in error studies and studies which investigate cognitive load (Paas et al., 2003; Wu & Li, 2013). While the results of this study showed an indication for the secondary task having an effect no significant difference in the amount of errors could be detected in the evaluated design features. This raised doubts of previous claims by Cauchi et al. (2012) on this issue. On the whole, even with an extra task, people made very few errors in these interfaces. A fact that supports findings from other researchers in the field (Oladimeji, Thimbleby, & Cox, 2011) and further speaks for a certain robustness of *5-Key* designs.

Additionally, there were some further differences but also similarities in contrast to their work. For example, Cauchi et al. (2012) suggest entering numbers on the *left-cursor-start* would lead to confusion when entering the numbers but Study One, Two and Three found no evidence for that. On the contrary the results showed that people in the left condition always entered numbers from left to right (only in one or two cases from right to the left). However, Cauchi et al. (2012) mentioned that people would use wraparounds when they are present a claim that could indeed be supported by the findings from Study One. While the studies in this chapter were unable to confirm Cauchi et al.'s proposed effects of specific design features on errors interesting features about people's strategies to enter numbers could be detected. This is also considered a very important finding. People are strategic and tend to change strategy when faced with *5-Key* interfaces for interesting reasons e.g. less workload.

The analysis by Cauchi et al. (2012) seems somewhat appropriate however, it seems not representative enough. Meaning it appears not to be an optimal analytical model as they claim specific design features are beneficial in terms of error reduction and the studies presented in this work found no evidence to support this claim. Cauchi et al. (2012) state some facts about the items that people were about to see when implementing such features in *5-Key* interfaces and some of these items do appear as above mentioned. To some degree Cauchi et al. (2012) seem to be on the right track. The way they are thinking about a device seems to be correct and some behaviours they expect are correct (people would use a wraparound - which in fact they did). They explained that people starting from the left got more confused but they did not. If the analytic work from Cauchi et al. (2012) was really strong then it should be possible to see the effect playing through in actual designs which did not happen in all the studies presented in this chapter. In fact, the studies revealed that other factors of interaction e.g. user strategy may be an important aspect, which the more keystroke level analysis by Cauchi et al. (2012) could not take into account.

In view of future studies based on previous work by Cauchi et al. (2012) it is important to keep in mind that the design features presented in their work are very much ad-hoc i.e. not systematic. This means that the features were taken without particular reason other than trying to find an optimal combination to improve the safety of *5-Key* interfaces and there could be other design features which have not been invented yet. The question about whether their whole analytic technique is correct makes it difficult to find a principled way forward. It is not clear that by further investigating their results would yield in any substantial contribution. It cannot be conclusively said that the analytic approach by Cauchi et al. (2012) is correct or not. It is potentially valuable but at this stage unproven as the studies presented here were unable to confirm their results in actual interfaces. In general, this suggests that these analytic techniques are of value but it is important to really understand the use of the device. A fact that the analysis presented by Cauchi et al. (2012) did not account for at this stage.

In terms of the evaluated interface itself, it is just too hard to see any significant effect of anything evaluated that can safely and confidently be attributed to the difference in the device interface due to the robustness of the *5-Key* design. This is considered an important finding as the presented studies weigh a great number of numbers against a small number of errors. However, this also means that further studies on design features as presented by Cauchi et al. (2012) would be difficult to get empirical access to. Although Study Three included a secondary task to increase the error-rate this was not successful. Because of this reason it was decided to move away from *5-Key* designs for future studies to a *calculator-based* interface which is widely used (cash machine, phone, calculator) but is less restricted in interaction and error potential (Oladimeji et al., 2011).

This Chapter presented three studies which investigated how design features can influence the amount of errors that people make with interfaces used on devices in safety-critical domains. Although the *5-Key* interface studies could not detect a significant difference between the cursor-start designs and their effect on reducing errors, many interesting discoveries have been revealed. Thimbleby and Cairns (2010) already pointed out that number entry seems at first a trivial task, however, people still make errors frequently. The studies presented here have shown that 50% (Study One), 69% (Study Two) and 50% (Study Three) of the participants made at least one error while entering numbers. This showed that although the experiments are designed to investigate an isolated transcription task these simple experiments were an adequate research method to investigate errors. With a minimum of 50% of people in all the studies making at least one mistake in a laboratory study gives a grim indication on real-world transcription errors. Additionally, the error presentation rate in all presented studies is generally low (2%, 2.9% and 2.3% as discovered) when entering numbers in medical devices, but that does not make it less important to mitigate these errors to avoid adverse consequences. The low error presentation rate is further evidence for specific characteristics of these kind of studies. While details such as where the cursor position starts in *5-Key* interfaces seem irrelevant to errors, contextual details however seem more important.

The key findings of Chapter Three are:

- A low but persistent error-presentation rate in all the presented studies could be detected
- No effect of a left-cursor-start compared to a right-cursor-start design on errors in *5-Key* interfaces was seen
- It was seen that people prefer strategies to enter-numbers in a *5-Key* interface
- The analysis by Cauchi et al. (2012) did not account for the strategies people use when entering numbers in a *5-Key* interface
- The *5-Key* interface appears to be really robust in terms of errors

#	Interface	IV	DV	Design	Hypothesis	Outcome	Conclusions
1	5-Key	cursor start (left vs. right)	Total numbers entered Total errors made Total error rate	Within (12)	Significantly fewer errors in the left-cursor start design than in the right-cursor start	No significant difference; Low error rate; Indicative strategies; Wrap feature blocked clear view on strategies; Inability to perform empirical tests	No support for Cauchi et al's results could be seen; There are indications that people prefer specific strategies when entering numbers however due to the Wrap feature this could not be clearly seen; Additionally, the low error-presentation rate did not allow for statistical testing. A follow-up study with an improved experimental design is required;
2	5-Key	cursor start (left vs. right)	Total numbers entered Total errors made Total error rate	Between (20 each)	Significantly fewer errors in the left-cursor start design than in the right-cursor start	No significant difference; Low error rate; Indicative strategies;	Cauchi et al's findings could not be supported, as there was no difference in cursor-start positions; Interestingly strategies became more visible as the Wrap feature was disabled; Perhaps increasing the error-rate would result in a visible effect and provide further evidence for strategies people use to enter numbers.
3	5-Key	cursor start (left vs. right)	Total numbers entered Total errors made Total corrections made Total error rate Completed audio (low) Completed audio (high)	Between (20 each)	Significantly fewer errors in the left-cursor start design than in the right-cursor start Significant increase of errors made in both audio conditions High and Low	No significant difference; Low error rate; Indicative strategies;	Previous findings by Cauchi et al (2012) cannot be supported. What is seen in studies One, Two and Three are strategies that people enter a factor that the analytic analysis by Cauchi et al did not account for. The secondary task did not have the expected effect to increase the error rate. There is a strong indication that the 5-Key interface itself is very robust in terms of errors and therefore effects of specific design features are really hard to see.

Table 3.8: Studies 1 - 3: Summary of findings

Chapter 4

Studies to investigate the DOV effect in Number-Transcription Tasks

"Man is still the most extraordinary computer of all."
John Fitzgerald Kennedy

The previous studies were motivated by Cauchi et al. (2012), Oladimeji et al. (2011) and their analyses. Their work and the three previous studies in Chapter Three can be seen as utilising design features for number-entry interfaces on a functional level. Investigating interface design features and their potential to influence the errors committed in transcription tasks is very likely leading to an endless thread of experiments. Cauchi et al. (2012) themselves essentially proposed almost thirty different variations of design choices for *5-Key* interfaces. All of which present considerable potential to be further researched.

It was therefore decided to step back from the near functional level and to consider what theoretical ideas could be used to reduce number transcription errors. It was at that stage when the results from Diemand-Yauman et al. (2011) (*DOV* effect) in relation to the work of Kahneman (2008) were discovered by further reviewing the literature (Kahneman, 2011). The *DOV* effect presents a theoretical approach with the potential to influence a feature in the context of the task of number transcription. The goal of this Chapter is to discuss the experiments investigating the relevance of the

DOV effect to the reduction of number-entry errors in the context of transcription tasks.

This chapter presents five studies (Study Four - Study Eight) all centred around core aspects (if the effect can be stimulated and detected) of the DOV effect. The first study (i.e. Study Four) used a text-transcription game as a research instrument to investigate the effects of cognitive driven design features on reducing errors. Results showed that introducing DOV style disfluency in the task led to fewer errors being made and therefore higher accuracy. This is partially in line with previous work by Diemand-Yauman et al. (2011). The next study (Study Five) was designed to apply the novel method used in text-transcription tasks to number-transcription task. This first attempt in number-entry produced a non-significant result, which could be attributed to environmental effects. Consequently, Study Six addressed the previous issues and could detect a highly significant effect on reducing errors once the new cognitive driven design features were present. This meant that the confirmed DOV effect is not limited to either one domain: text- or number-transcription but instead there is a strong indication that it has general relevance.

It was assumed but not confirmed that the DOV effect is a source-related effect and not a target-related effect in this interactive context. Study Five and Study Six did not explicitly distinguish between a source or target effect and therefore introduced disfluency in both the source and the target display. Although Study Four did not obscure the target area (i.e. text-input area) the DOV effect was seen, which is indicative for a source related effect as the study did not explicitly consider this independent variable. The promising results of Study Four and Six were then followed by Study Seven, to measure which part, the perception of the information at the source or at the target display, were responsible for the DOV effect to manifest itself. However, Study Seven failed to detect the DOV effect because participants showed measurable fatigue effects in the within-participants design. Study Eight took this into account by repeating the same experiment in a between-participants set-up and was able to detect the DOV effect. Furthermore, the study revealed that indeed it appears that the source is responsible for the effect. These results add to previous work of Diemand-Yauman et al. (2011).

4.1 Experiment 4 - TypoMadness

The following study was designed as a first attempt to look into the *DOV* effect in transcription tasks. Number-transcription experiments, as the previous studies in Chapter Three confirm, generally produce a very low error presentation rate (usually around 5% or lower). If the *DOV* effect has an effect on reducing errors it is necessary to see this effect in a system with lots of errors. This is the reason why a text-transcription study was chosen for this first approach simply to have a greater opportunity for a larger amount of errors to be presented and therefore to be able to detect the *DOV* effect.

4.1.1 Aims & Hypothesis

This study was designed as a test of principle to see if the *DOV* effect would manifest itself in the context of transcription tasks (in this case a text-transcription task). The study used the text-transcription game *TypoMadness* as research instrument where participants had to read short sentences and re-enter them into the system. The to-be-transcribed sentences were represented in two distinct font colours: black (Hex code: #000000) and light-grey (Hex code: #DFDFDF). The condition with the black font presentation is defined as *Clear* and the condition with uses the light-grey font presentation is defined as *Obscured*. The particular colour code for the disfluent i.e. *Obscured* condition was chosen after the similar font in (Diemand-Yauman et al., 2011). Although Diemand-Yauman et al. (2011) did not specify a certain colour code as they only stated a "60% grey-scale" font-colour (p.2) the colour code used in Study Four comes very close to the one used in previous work. The transcribed sentences were then checked for mistakes, which produced the score. The score was measured as how many correct characters were entered by the participant. A higher number of correctly transcribed characters led to a higher score. The hypotheses for this experiment were:

H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.

H_1 : There will be significantly fewer errors in the *Obscured* condition than in the *Clear* condition resulting in the manifestation of the DOV effect.

4.1.2 Participants

Twenty participants (11 women) took part in the study and were randomly recruited by email as well as personal invitation. The participants were between 19 and 35 (mean 27.15; SD 8.52) years old. Participants were mostly students in the department of Computer Science at the University of York except one post-doctoral researcher and a maintenance worker. The following countries were represented by participants: Japan, Portugal, Indonesia, Thailand, Malaysia, China, UK, Canada, India, Netherlands and Russia. Nine participants stated that the task was difficult. However, only four of these nine people were part of the *Obscured* group. Moreover, these four participants further stated other reasons such as being unable to use the backspace-key than the visibility of the given text as the reason why they felt the task was difficult. None of the participants indicated that they had dyslexia or any other disability. Additionally, all of the participants stated that they had normal or corrected to normal vision and only two participants (one from each group) stated that the text presented to them was less readable. All participants were rewarded with chocolate after the study.

4.1.3 Text-Transcription Task

The task set to participants was framed as a game where they had to enter as many sentences as accurately and as quickly as possible within 15 minutes. Each sentence had between 95 and 161 characters and was displayed in either *Clear* or *Obscured* format. Twenty-six sentences were available in total for the study, which were displayed individually on a page while participants progressed through the game. In addition, the sentences were not connected in prose or any other form, so participants could not have predicted them. The sentences were randomly chosen from news and boulevard press websites such as *theguardian.co.uk* or *cnn.com* and can be found in Appendix B.1.3.

4.1.4 Design

The design was a between-subjects design with ten people in each group. The independent variable was the visual representation of the to-be-transcribed text and there were two conditions: *Clear* and *Obscured* (see figure 4.1).

The dependent variable for this study was the score but also the number of sentences and the number of characters entered that participants could achieve when playing the game. The score is the accumulated sum of correctly entered characters in each task reduced by errors. An error in this study is anything that is not identical to the target text. For instance, this could be an additional space, upper or lower case difference, wrong or missing letters or generally any sign different from the displayed text. It should be noted that where an error arose due to inserting an extra character, this did not cause all subsequent characters to be classed as erroneous. To measure character differences (minimal edit distance) a version of the Levenshtein algorithm was implemented followed by manual checks of the datasets (Levenshtein, 1966). The set of 26 text phrases was counterbalanced across the study to avoid ordering effects.

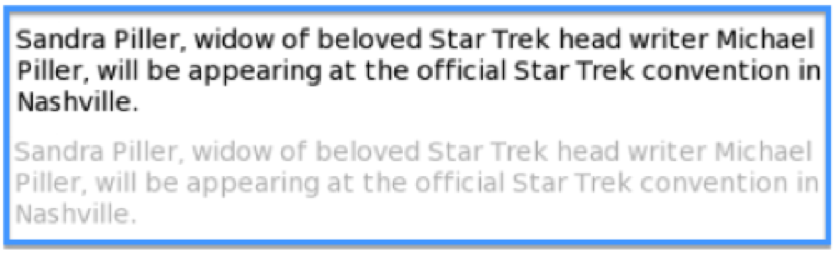


Figure 4.1: Study 4: To-be-transcribed content in the Clear (upper) and Obscured (lower) condition

4.1.5 Materials

The text-transcription game *TypoMadness* was played on a Apple Macbook Pro with a UK English keyboard layout and a 13" size display at a 1280 × 800 resolution. The browser used for the experiment was Google Chrome running under Apple Mac OSX 10.7.5 (Lion). The game website ran in the browser, the browser window was maximised and controls were disabled from being shown so that only the game itself was visible. All other applications were closed. Participants could submit their entered text and proceed to the next text paragraph by clicking on the red button on each page of the game.

In addition to the game website described in section 4.1.6, the instructions as well as an informed consent document were printed out on paper for the participants to read and sign before the beginning of the study (see Appendix B.1.1). Furthermore, there was also a brief demographic questionnaire which was also printed and given to the participants to complete after the trial (see Appendix B.1.2). The demographic questionnaire was used to specify the participants but was not further used in the following analysis. A smart-phone was used as a stopwatch.

4.1.6 TypoMadness - Game Design

The game used for this study was implemented as a website in PHP with JavaScript. The website was connected to a MySQL database allowing the entered text phrases to be stored in the database for each participant. The displayed score was calculated out of the matching characters of each submitted text phrase. Each score for individual text phrases was stored in the database and displayed via cookies on the game website. The game website consisted of a display area in the centre of the page where the text paragraphs were displayed, a text box for typing the text and a red button to submit the entered text and continue with the next task.

The backspace button and the cursor keys were disabled in the text field so that participants could not change their entered content. All transcribed sentences, no matter what they would be, were saved in the database after a participant pressed the red button. The entered sentences were saved as a string in the database and compared to the original text to produce the score.

There were two versions of the game website used in the study. One for the *Clear* group where the to-be-transcribed text was presented as fluent content and another one for the *Obscured* group where disfluency was introduced in the presented content.

Figure 4.2: Study 4: TypoMadness interface in the Obscured condition

4.1.7 Procedure

After reading the instructions and signing the informed consent form, participants completed two training sentences, which were separate from the main sentence set of the experiment. Each of the target text a participant had to transcribe, whether it was for the training to get familiar with the controls or the main study, was displayed in a text box at the centre of the individual website. After the participants finished entering the training session at their own pace a new start page was displayed only containing a start button. On completion of the training period, the participant proceeded to the main study by clicking on the start button. Participants could see their score (in form of a character count as seen in figure 4.2) displayed on the right side of the text field. The target text was presented either *Clear* or *Obscured* depending on the condition. Participants were instructed to transcribe the target text as quickly and as accurately as possible. In order to get the highest possible score, all tasks had to be completed within 15 minutes though participants were not shown how long they had been playing. The experimental design in form of a game made it necessary to put participants under some form of time-pressure. This was also done as another means of increasing the error rate besides choosing a text game for an increased error data pool. Each time the participant pressed the red button, their score was updated though there was no direct indication of any errors. If they completed the task within the 15 minutes, the final pages displayed the overall score. A smart-phone was used to stop participants at the end of the given time limit. Participants were offered a break after finishing the transcription task before they were asked to fill out a demographic questionnaire. The participants were then debriefed as to the goals of the study and rewarded with chocolate.

4.1.8 Pilot

A small pilot study was conducted to measure how many sentences a participant could enter in the given time-frame of three minutes. There were two participants, both of them research students in Computer Science at the University of York. The pilot showed that the initial set-up of 26 given sentences and three minutes to enter these sentences was too narrow. Neither of the participants could manage to transcribe all the text phrases in time. Additionally, the length of the text-phrases (originally 220 characters) was too long as participants struggled with entering lots of phrases in the given time-frame. It was therefore decided to keep the amount of 26 sentences for the participants to enter in both groups (*Clear* and *Obscured*) but to increase the available time-frame from three to fifteen minutes. Moreover, it was also noted that the initial five training sentences were too many and would rather contribute to participants showing fatigue effects. Therefore the number of training phrases was reduced to only two sentences. Additionally, some syntax mistakes were detected and removed during the pilot study as well as the auto-fill option of the browser had to be disabled for the text fields as it had a highly confounding effect on the performance of the participants.

4.1.9 Results

It was hypothesised in this study (H_1) that the DOV - effect would manifest itself in the context of an interactive text-transcription task. The mean of the errors made in the *Clear* group was 830.1 (SD = 516.3) whereas in the *Obscured* group was 360.4 (SD = 127.0) (see figure 4.3).

A *Mann-Whitney-W* (*two-tailed*) test was conducted, as error data does not generally follow a normal distribution, which revealed a significant difference between the two groups ($W = 23$, $p = 0.04$). Though the hypothesis could be seen as a directional

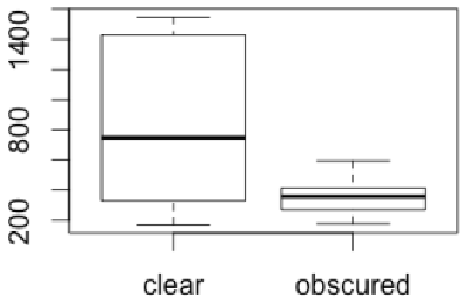


Figure 4.3: Study 4: Boxplot of total errors made by all participants in each condition

hypothesis, a more acceptable *two-tailed* test was chosen to avoid the controversies around *one-tailed* testing (Abelson, 1995). This move resulted in a higher threshold for significance and therefore would provide a higher level of confidence in the data.

The maximum number of sentences each group could transcribe was 260 (33,450 characters). Group *Clear* entered in total 25,269 (77%) characters correctly while as the *Obscured* group entered 30,356 (91%) of the characters correctly. The total amount of text transcribed by the *Clear* group was 230 sentences and 242 sentences in the *Obscured* group. Group *Obscured* made 14% less errors than group *Clear*. A table containing the total achieved scores and transcribed sentences can be seen in Appendix B.1.4.

In terms of speed of completing the task (see figure 4.4), the mean rate for transcribing the sentences for the *Clear* group was 1.69 sentences per minute ($SD = 0.45$) and for the *Obscured* condition 2.02 sentences per minute ($SD = 0.49$). There was no significant difference in the rate of entering sentences ($W = 26$, $p = 0.07$).

Moreover, the participants in the *Clear* group transcribed a mean of 23.1 ($SD =$ 2.99) sentences and the in *Obscured* group 25.0 ($SD = 1.634$). A boxplot illustrates the

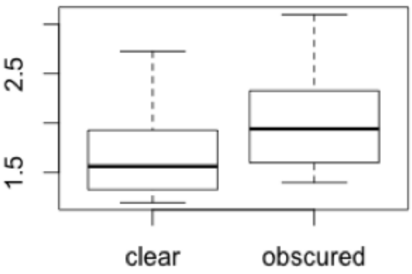


Figure 4.4: Study 4: Boxplot of the rate of typing (sentences per minute) of participants in each condition

sentences entered by group (see figure 4.5). There was no significant difference between the number of sentences participants completed ($W = 32$, $p = 0.15$).

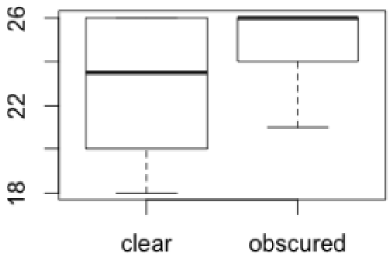


Figure 4.5: Study 4: Boxplot of the number of sentences entered by each participant in each condition

Four participants from the *Clear* group and seven participants from the *Obscured* group managed to transcribe all the twenty-six given text phrases and finished ahead of time. This leaves six participants from the *Clear* and three from the *Obscured* group who could not finish in time and therefore used the full fifteen minutes.

4.1.10 Discussion

This study was designed to be the first attempt to measure if people who are transcribing text in an obscured and disfluent (i.e. harder to read) way apply more deeper cognitive processes to transcribe text and therefore commit fewer errors. The presence of the *DOV* effect could be detected and as predicted, the results showed that participants in the *Obscured* group achieved higher accuracy scores (i.e. committed significantly fewer errors) than those in the *Clear* group. Participants seem to perform better in transcribing text when they were given disfluent, less readable text.

One possible explanation for this reduction in errors could be a speed accuracy trade-off. However there was no significant difference between the two conditions in either the rate of transcribing sentences or the total number of sentences transcribed by each group. Indeed, the participants in the *Obscured* group, if anything, entered more sentences, working at a faster rate suggesting that the introduced disfluency (i.e. reduced readability) of the font was not slowing the participants down. Nonetheless, they made fewer errors than the *Clear* group.

The manipulation applied in the experiment is identical to that of Diemand-Yauman et al. (2011) who suggest that introduced disfluency would stimulate a more active and deeper processing of the information. This can result in an increase of the likelihood of accurately memorising the information (Diemand-Yauman et al., 2011). Moreover, the manipulation used in this study only consisted of representing the to-be-transcribed information in a harder to read font-colour but not also in a different font-type as in Diemand-Yauman et al. (2011), which indicates a more robust outcome.

In view of future studies, an improved experimental design would suggest to use sentences, which are all of the same length to avoid possible control confounds. Additionally, further studies would benefit from only inviting people who have expert typing skills as individual speed might vary and would introduce noise and distortions in the data. However, the *DOV* effect could be detected despite possible variations in typing speed.

Some participants mentioned that they had difficulties and were not quite comfortable with using the Apple Macbook Pro keyboard as they did not own an Apple computer before. The keyboard layout was an UK layout however the feeling is different from a standalone keyboard. A standard Personal Computer (PC) keyboard may be a better approach in terms of the experimental design to avoid any confounding effects for a further text-transcription study. Still, considering this issue an effect in reduced committed errors could be detected.

Grey text can be more legible than black text under specific circumstances, depending on the level of contrast with its background and the overall level of luminance. Some types of dyslexia respond well to contrast reduction for improving readability. However, there were no dyslexic participants (at least not that they were aware of) in the reported study. Additionally, prior experience in data-entry work of participants or even any prior knowledge of the material might have influenced their performance in the given task; therefore a more careful selection of participants for future studies needs to be considered.

In review of this Study being the first attempt to apply the theoretical concepts behind the DOV, the results presented here were overwhelmingly positive which greatly inspired the following number transcription studies.

4.2 Experiment 5 - The first Application in Number-Entry

Having had the very promising results from the initial text-transcription study (Study Four), which suggest that perceiving information presented in a illegible and disfluent state could lead to a reduction in the amount of errors that people make when transcribing such a content, has lead to Study Five with the intention to transfer the results from text- to number-transcription tasks.

The following number-transcription study was being run as a practical session in the undergraduate first year module, Human Aspects of Computer Science (HACS), at the University of York. The idea was to provide the students with the opportunity to get experience of running participants through a study. The study was conducted during practical classes with the students working in pairs with one person acting as the experimenter and the other person as the participant. The study and all its material was designed and prepared by Frank Soboczenski. Students were instructed at the beginning of the module but ran the experiment autonomously on each other. The role of the author in this experiment was:

(i) to produce all the material for the study; this included the informed consent and instructions (see Appendix B.2.1) form, the demographic questionnaire (see Appendix B.2.2) and the experimental tool (programming of the interface)

(ii) to inform the students about the experimental procedure

(iii) to inform the students about the lager purpose of this study i.e. the use of the data in future publications

All the participants gave their consent by signing the form in Appendix B.2.1. Section 5 of the form informed participants of the future use of the data.

The departmental module HACS is designed to teach students how the different areas science, engineering and design are relevant in computer science. Students develop basic skills in experimental methods, rationalise the use of user-centred design on ethical, social and legal grounds and explain aspects of software engineering though the view of user-centred design methods. As part of the module students participate and conduct a series of experiments which they will also learn to statistically analyse and interpret. At the end of the module, students will be able to critique experimental methods used in computer science and human-computer interaction in particular, devise and conduct experiments with users using best practice, including concern for the ethical conduct of experiments, select, perform and interpret basic statistical tests, justify the use of user-centred design on ethical, social and legal grounds and apply user-centred design methods to the design of a system (Cairns & Cox, 2008).

This scheme results in a mutual benefit for students and the lecturer or assisting PhD students. The students are able to engage ethically in designing, running, analysing and interpreting experiments whereas the lecturer or PhD student benefit of using the data generated from the student's experiments for their research. However, all the studies were conducted by following the strict ethics outline presented on the Ethics page at the beginning of this thesis.

4.2.1 Aims & Hypothesis

The aim of this study was to transfer the results from the previous text-transcription study to the domain of number-transcription to investigate if the *DOV* effect can also be measured in this domain. The task and the hypothesis were directly transferred from text- to number-transcription for this experiment. The hypotheses for this study were therefore:

> H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.
>
> H_1 : There will be a significant difference in the number of errors made between the conditions *Clear & Obscured* resulting in the manifestation of the DOV effect.

4.2.2 Participants

The participants for this experiment were 53 first year undergraduates (4 women). The age of the participants was between 17 and 19 (mean 18.43; *SD* 1.18) years. Only one of the participants had previous experience of working in the healthcare sector. Fourteen of the participants stated that they never used a number-pad to enter numbers. However participants were mostly familiar with a serial number-entry interface as used in the study (only two people stated "No" in the questionnaire). All participants voluntarily took part in the experiment as part of the Human Aspects of Computer Science module. As in previous experiments the participants were randomly assigned to the specific groups.

4.2.3 Number-Transcription Task

The task set to participants was also framed as a transcription game to enter as many numbers as quickly and as accurately as possible in a *Serial* i.e. calculator-based interface. Participants had a given time-frame of maximum five minutes. Compared to the previous experiment a five-minute limit was chosen for this study as the pilot study showed that participants could transcribe content that consists of numbers considerably faster than text. The numbers were randomly generated in a range from 1-999.99. About 25% of the random numbers have no decimals the others vary between one and two decimal digits.

4.2.4 Design

A between-subjects design was used for this experiment. Identical to the previous study there were two groups *Clear* and *Obscured*. However, different colour codes were used for the conditions as to the previous one. In light of the previous results it was highly expected that the *DOV* effect would manifest itself in number-entry. Therefore, as Diemand-Yauman et al. (2011) did not use or recommend a specific colour range but rather chose ad-hoc to use grey scale as base colour, this experiment takes this one step further in an attempt to extend the likely generalisability by simplifying original manipulations. Simplification in terms of only having one manipulation to make the information harder to read but also hoping to improve the generalisability. Should this experiment be successful then one interpretation would be that only fading the to-be transcribed information alone might be enough to produce the *DOV* effect. A different background colour for the *Obscured* condition was chosen with the aim to further increase deeper cognitive processes and therefore make the *DOV* effect more prominent (see figure 4.6). The colour used for the *Clear* condition was blue (HEX-code: #5882FA) on white background and for the *Obscured* condition was the same blue font-colour but on slightly brighter blue background (#4E7DFF).

The dependent variables were the total number of uncorrected errors made, particularly the rate of errors and also the number of corrected errors in each group. A single error was counted when any number was entered other than the displayed number. Further, a corrected error is an error where participants noticed that they made an error and entered the number again. These errors were not added to the total number of errors. As previously, error type is not considered in the analysis. *Note: the Obscured condition in the following figure 4.6 may be difficult to see due to the quality of the print. Please consult the digital copy of this thesis.*

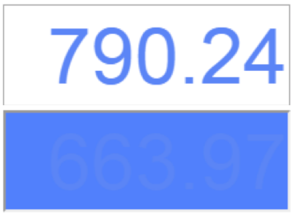


Figure 4.6: Study 5: Number to be entered in both conditions Clear and Obscured

4.2.5 Materials

In addition to the number-transcription game described in section 4.2.6, an instructions and an informed consent document (see Appendix B.2.1) as well as a demographic questionnaire (see Appendix B.2.2) were printed out on paper for the participants to read and to sign. The computers used were Microsoft Windows Lab PCs with a 17" size display at a 1920×1080 resolution. The browser used for the experiment was Mozilla Firefox 10 running under Microsoft Windows 7. The number-transcription game ran just like its predecessor (*TypoMadness*) in the browser. The browser window was also maximised and controls were disabled from being shown so that only the game itself was visible. All other applications on the individual machine were closed. Furthermore, each keyboard was prepared so that participants could only use the number-pad to enter numbers. The following keys were labelled: *0,1,2,3,4,5,6,7,8,9,.,Enter,+*. Stickers with the label *clear* were put in place on the + key to make it easier for the participants to associate the particular function. A smart-phone was also used for this study to measure the time.

4.2.6 Number-Transcription Game - Interface

The visual design of the *Serial* i.e. calculator-based number-transcription interface was created after real world infusion pumps such as the Baxter ©AS40A or the Graseby ®3400 syringe pump (see figure 4.7). The real devices use an identical number-pad layout except for the "Submit" or "Cancel" button. For the purpose of this study only the number pad of the infusion pumps with an *Enter* and a *Clear* key was replicated all other buttons of the original interface design were removed from the interface as they could have a possible unwanted effect on participants.

Figure 4.7: Study 5: Infusion pump models Baxter and Graseby

The number-transcription game was implemented similar to the previous text-transcription game in PHP and JavaScript with a logging feature, which recorded all user interactions with the virtual device and wrote all interactions into a MySQL database. The interface provided thirteen functional buttons usable by the keyboard-only interaction (see figure 4.8). The display itself consisted of 16 components. Three number displays (number to be entered (source display), score and the number input field), *Enter* and *Clear* buttons and 0-9 number buttons including the decimal key. By pressing the *Enter* key the current displayed number was saved and at the same time the input display was cleared and a new number in the source field was randomly generated and presented. The display itself provided participants to enter a 3-digit number with two digits after a separating decimal point (5 digits in total) as previously gathered numbers from infusion pump logs confirmed the validity of this number range

(Wiseman et al., 2011).

There were two versions of the interface used in the study. One where the number to be entered was displayed with a normal font-colour (blue on white background) (group *Clear*) and one where the number to be entered was displayed in a hard to read blue on blue background in group *Obscured*. Each number a participant had to enter, whether it was for the training to get familiar with the controls or the main study, was randomly generated in the game and displayed in the left upper corner of the interface as in figure 4.8.

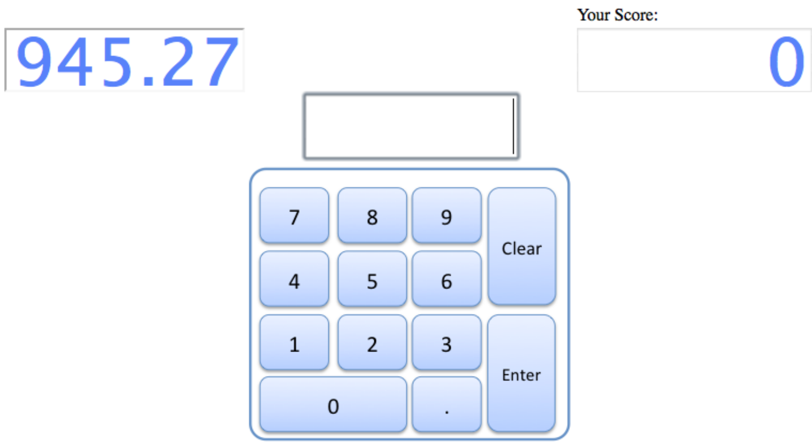


Figure 4.8: Study 5: Number-Transcription interface in the Clear condition

The reason for choosing a *Serial* interface instead of a *5-Key* or *Chevron* interface is again anchored in the low error presentation rate in the previous studies. As previous work, in addition to the studies presented in Chapter Three, showed *5-Key* number-entry interfaces are more robust in comparison with *Serial* interfaces (Oladimeji et al., 2011, 2012). This can be explained in light of the input mechanic that people have to follow when entering numbers. For example, an open keypad does offer more room for errors, in particular slip errors, whereas the interaction with a *5-Key* layout is more restricted to the cursor movements. It is still possible to commit the same kind of errors but at a lower rate therefore choosing a serial layout to investigate the *DOV* in this context was simply the logical decision to make.

4.2.7 Procedure

The experiment was conducted as a Human Aspects of Computer Science (HACS) practical, which took place in one of the computer labs at the University of York. Students were asked to pair up into groups. Each group consisted of one experimenter and one participant. The practical was designed so that students can experience both roles: being a participant and an experimenter during a study. There were two experiments for the practical sessions. One focussed on immersion in games and the other was this experiment. Both students in each group were assigned with running and participating in one experiment. For example one student would be experimenter for the number-transcription study and then after the experiment was finished switch role and become the participant for the immersion study. Students received clear instructions on what to do for both studies. This included a detailed instruction sheet for the experimenter of the number-transcription study as this was designed to mostly run on-line.

Each participant of each group was welcomed by the experimenter and given a copy of the informed consent form and the instructions to read and sign. After a participant had signed the form they were asked to go to the game website (`www.cs.york.ac.uk/numbers`) and to login in with their given user name and password (see figure 4.9). The login details were provided in the instructions. After a participant logged into the system they were asked to become familiar with the use of the keyboard controls and the display of the interface. For this purpose they were asked to run ten training trials, which were separate from the main experiment. The participants were asked only to use their index finger to enter the numbers. After a participant became familiar with how the *Serial* number-entry interface worked he or she was then asked to proceed to the main study. Participants were then asked to enter the given numbers either in the *Clear* group or *Obscured* group within five minutes. The experimenter stopped the participant after the given time-frame. After finishing the number-transcription task, the experimenter offered each participant a small break. Each participant was then asked to fill out a small demographic questionnaire. Participants were then debriefed as a class to the goals of the study. Finally, each participant was asked if there was anything they wanted to tell the experimenter about the study, their experience or any general comment.

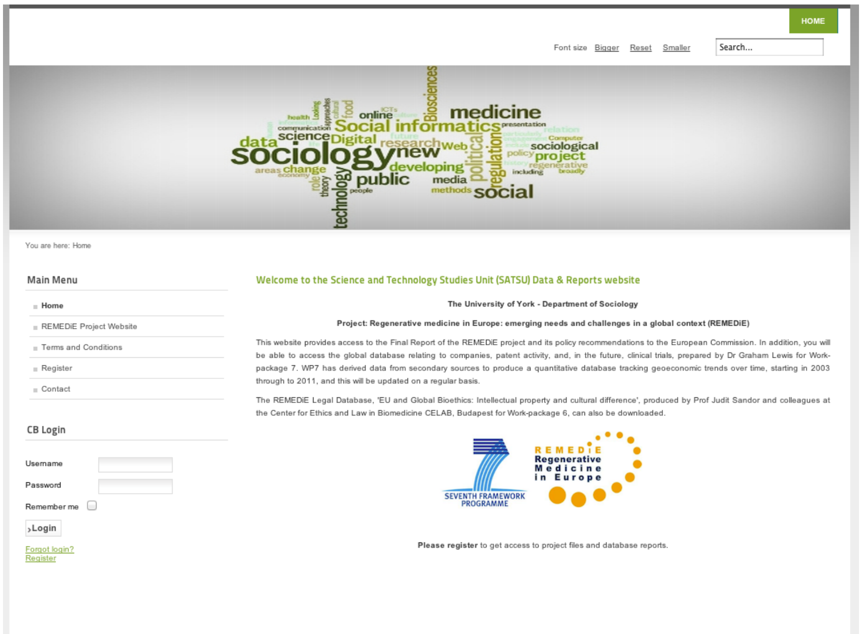


Figure 4.9: Study 5: Experiment website

4.2.8 Pilot

A small pilot study was conducted to test the experimental design and to measure how many numbers a participant could enter in the given time-frame of five minutes. There were two participants who were not students of the University of York. The pilot showed minor design flaws in the developed tool. For example, instructions had to be included for students to disable the browser auto-fill function. This was done to avoid confounding effects when entering numbers. No other changes were made.

4.2.9 Results

In total 4235 numbers were entered for all conditions, which resulted in 202 total errors (4.7%) and 381 corrected errors. 2223 numbers were entered in group *Clear* (117 errors) and 2012 in group *Obscured* (85 errors), respectively. The mean of the numbers entered in the *Clear* group was 85.5 ($SD = 18.1$) whereas in the *Obscured* group 74.52

(SD = 14.03). A *two-tailed Mann-Whitney-W* test revealed that there was a significant difference in the total numbers entered between the groups ($W = 496$, $p = 0.01$) (see figure 4.10).

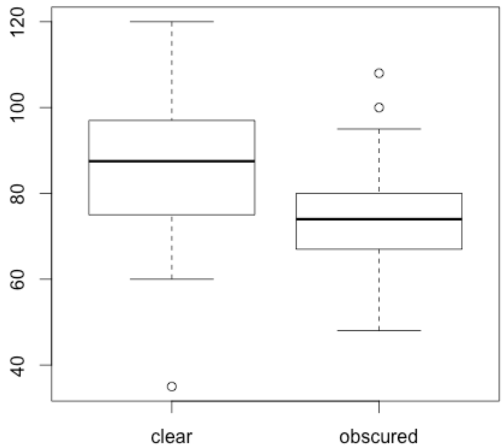


Figure 4.10: Study 5: Boxplot of total numbers entered by all participants in each condition.

There was no significant difference between the errors made in the two groups ($W = 356$, $p = 0.94$) (see figure 4.11). The mean of the errors made in the *Clear* group was 4.5 ($SD = 6.17$) whereas in the *Obscured* group 3.15 ($SD = 2.27$) (Outliers did not have an effect).

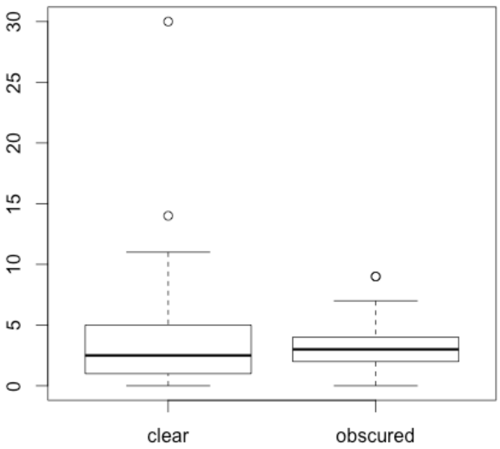


Figure 4.11: Study 5: Boxplot of total errors made by all participants in each condition.

The mean of the error-rate (errors divided by numbers) in the *Clear* group was 0.05 (SD = 0.06) whereas the mean in the *Obscured* group was 0.04 (SD = 0.03). The *Mann-Whitney-W* test revealed no significant difference in terms of the error rate between the groups ($W = 238$, $p = 0.69$) (see figure 4.12).

There was also no significant difference in the number of corrections in the conditions ($W = 301$, $p = 0.37$). Additionally, the mean of the corrections in the *Clear* group was 6.15 ($SD = 5.58$) and in the *Obscured* group 8.17 ($SD = 9.36$).

Forty-eight participants made at least one error when transcribing numbers (93%). Only four people made no errors. Three people in the *Clear* group and one person in the *Obscured* group.

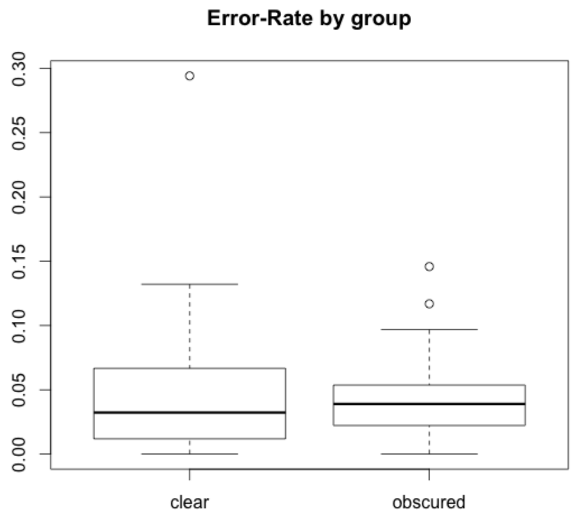


Figure 4.12: Study 5: Boxplot of the error rate made by all participants in each condition.

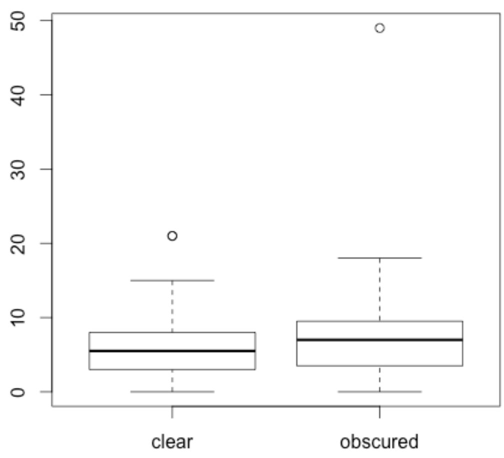


Figure 4.13: Study 5: Boxplot of total corrections made by each participant in each group.

4.2.10 Discussion

The alternative hypothesis (H_1) could not be accepted for this study as there was no significant difference between the groups *Clear* and *Obscured* in the rate of errors people make when transcribing numbers. Moreover, there was also no significant difference in the amount of corrected errors between the groups. This speaks against the possibility of provoking the *DOV* effect based on the construct of *System 2* in transcription tasks (Kahneman, 2008). On the other hand, the reason for a rejection could be that the experimental design was not optimal for this study. In this Study there were 53 people participating at the same time and this subsequently led to a considerable amount of people interacting with each other. Although it could be argued that in real-world transcription tasks such as in hospitals, people are often distracted and influenced by noise, therefore this could be seen as in principle similar to running an experiment in an undergraduate module which does have a certain noise factor of distractions. Effects such as attentional failure (see Chapter Two Section 2.2.4) might come into play. It is also known that distractions lead people to interleave the task which can lead to an increase in errors (Back et al., 2012). Conducting the experiment in a separate environment certainly would address this issue. Additionally, students were running the experiment on each other therefore introducing a lack of strong experimental control.

However, there was a significant difference in the amount of numbers people entered suggesting that people in the *Obscured* condition slowed down and therefore entered significantly less numbers. Moreover, as the study was not run in ideal conditions lighting could have been another confounding factor. For example, when participants looked at the display in the wrong angle they could have perceived the to-be-transcribed information in the *Obscured* condition as too obscured so that it was impossible to read and therefore led to negative results. It could be that the information presented in the *Obscured* condition was simply too faint to provoke the *DOV* effect (i.e. deeper cognitive processing) and rather acted as barrier against it. This was verbally indicated by many participants after the study. An explanation for this could be that initially deeper processes might have been activated by the stimuli but the continuous active processing of the faint and obscured information led to a depletion of their working memory capacity as described in Mayer (2009) therefore inhibited the activation of deeper processes.

Additionally, the use of a different base font-colour and different background colour as means of introducing disfluency could very likely have contributed to inhibit participants engaging in deeper processing. Diemand-Yauman et al. (2011) did not specify if other colours are able to produce the *DOV* effect not did the literature provide any colour specific suggestions. With the initial success of Study Four it was decided to apply the same grade of obscurity on a different colour font. This was done with the hope of extending the generalisability of the manipulations (see section 4.2.4)

Unlike Study Four this experiment could not detect the same effect of reduced errors in the *Obscured* condition. However, as this was the first attempt to transfer the previous findings to the number-transcription domain and as the analysis of the experiment revealed room for an improved experimental design it was unclear at this moment if the stimuli used in this study were actually causing an effect. Additionally, the noise factor and the level of distractions as well as the lighting are indicative for an improved design which were taken into account in the next study.

There was also the argument that the stimuli of this study applied a too high grade of obscurity of the to-be-transcribed content which probably resulted in participants overloading their working memory capacity, getting frustrated or suffer from fatigue and therefore abandon the task. This is regarded as a major issue and influenced the next study to explore performance across a range of levels of obscured information. This range of levels was also aimed to further investigate if it was possible to detect a threshold or in other words a sweet-spot of when the *DOV* effect would become significant.

4.3 Experiment 6 - The Revised Number-Transcription Study

The next study was run as a refined version of the previous number-transcription study. The previous experiment could not find evidence for less visible content contributing to a reduced amount of committed errors, which could suggest a successful activation of the *DOV* effect in number-transcription tasks. However, as Study Five was the first real attempt to apply the techniques used in text-transcription (Study Four) to the number domain, the question arose if the applied stimuli i.e. the faint font, was possibly too faint causing people to perform poorly in the task.

Alternatively, the use of blue coloured fonts and background colour as basis for this experiment did not yield in the expected outcome that colour as a contributing factor could be ignored in future studies. This could be seen as a confound but then again the *DOV* literature did not have any evidence of colours having an impact as this is simply the first time this effect is investigated in an interactive context. At this moment it is unclear if a different colour could impact the *DOV* effect or if the negative result resulted out of the obscured background colour alone. Of course one option was to investigate this finding in further studies and to explore if any colours or combinations of colour can be of influence. It was however decided not to follow this avenue and to rather focus on further exploring the effect itself by turning the attention towards a possible sweet spot - a threshold of obscurity of where the effect could be seen.

This study is identical in its set-up to Study Five but in order to avoid introducing a less visible font colour which could act as a confound and in order to see if a possible sweet spot could be identified this study used three conditions: *Clear*, *Medium* and *Obscured*. The reason for having three conditions is to have another level of measurement to detect if the *DOV* effect manifests itself. Additionally, the *Obscured* condition is using a less faded font colour compared to the previous study meaning that the stimuli applied in this study are not as strong as in Study Five.

4.3.1 Aims & Hypothesis

The aim of this study is identical to the previous number-transcription experiment. Albeit the hypotheses are also similar and only differ in the number of conditions (the conditions are explained in the design section 4.4.4):

H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.

H_1 : The number of errors made in groups *Obscured* or *Medium* will be significantly lower than the number of error made in the *Clear* group.

4.3.2 Participants

Participants were randomly recruited by personal invitation or by email. There were in total 30 participants (14 women) for this study. The mean age of all the participants was 26.56 with a *SD* of 8.24 making them between 18 and 34 years old. Participants were mostly familiar with a *Serial* based number-entry interface as used in the study. None of the participants had previous experience of working in the healthcare sector. However, three participants stated that they were familiar with infusion pumps. Two people stated that they found the task difficult (one in the *Clear* and one in the *Obscured* group) and two different participants stated that they did not consider themselves to have normal vision (also one in the *Clear* and one in the *Obscured* group). Five participants had never used a number pad before, seven rarely used it, fourteen sometimes used it and only four participants (one each in each of the *Clear* and *Obscured* groups and two in group *Medium*) often used a number pad to enter numbers. All participants who took part in the experiment were rewarded with chocolate. Participants were again randomly assigned to the specific groups.

4.3.3 Number-Transcription Task

The tasks set to participants was identical to the previous number-transcription study. Participants had to enter as many numbers as possible in the given time-frame of five minutes. The same number-transcription game was used therefore the mechanisms were responsible for randomly generated numbers in the range from 1-999.99 adhering to the same number presentation standard and decimal to integer distribution.

4.3.4 Design

A between-subjects design was chosen for this experiment (ten participants in each group). The difference to the previous study was that unlike using two groups *Clear* and *Obscured* this study had three distinct groups. Condition *Clear* which is identical to Study Four, *Medium* which uses a light-grey coloured font (RGB Hex-Code: $\#BDBDBD$) and *Obscured* which uses an even fainter-grey faded presentation of the numbers (RGB Hex-Code: $\#DFDFDF$). Notably, the *Obscured* condition is not as strongly obscured as in the previous number-transcription study but the same as in Study Four. The *Medium* condition is in fading gradient located between the other two conditions. The three conditions were chosen to provide a better range to measure a possible effect and to avoid a possible overpowered stimuli as indicated in previous results. The independent variable was the grade of visibility of the number presentation. The dependent variables were identical to the previous study. This included the total number of errors made, the rate of errors and also again the number of corrected errors in each group. Similar as before corrected errors were not added to the total amount of errors.

Figure 4.14: Study 6: Source display illustrated in all three conditions

4.3.5 Materials & Number-Transcription Game

Apart from the technology used for this study the only other materials were printed instructions and an informed consent sheet for participants to read and to sign before running the trials (see Appendix B.3.1).

As for the technical part the study was identical to the previous number-entry experiment except that only a single computer was used to collect data not multiple ones like in a teaching lab setting. An Apple MacBook Pro 13" connected to an external keyboard and monitor with a 17" size display at 1920 × 1080 resolution was used. Additionally, the browser used for the experiment was Mozilla Firefox 10 running under

MacOSX 10.7.5 (Lion). The external keyboard was again modified with stickers as before. Unlike the previous study a Google on-line form was used to collect demographic data (a structure of the questionnaire can be seen in Appendix B.3.2). In order to measure the time a smart-phone was used throughout the study. The number-transcription game was the same as described in section 4.2.6. The only difference was that three different set-ups existed in this study in respect to the three conditions.

Each time a participant pressed the *Enter* key the number he or she just entered, which was shown in the centre display, was saved to the database and a new number was instantly and randomly generated at the upper left display. Additionally, at the same time the numbers in both displays were compared and if they matched the score was increased by 100 points. Any other entry than the correct number resulted in a decreasing of the score by 100 points.

4.3.6 Procedure

In contrast to Study Five, this experiment was conducted in an empty office in the Computer Science building. All participants were run in individual sessions which were managed by email and personal invitation. Participants were given the instructions to read and to sign. After they had signed the documents, they were asked to become familiar with the use of the keyboard controls and the display of the interface. For this purpose, participants were then asked to run ten training trials, which were separate from the main experiment. Participants were asked to sit straight in front of the monitor to avoid a possible effect on the perception of the numbers as the *Medium* and *Obscured* information could be perceived differently from a different angle. The lighting in the room was also controlled. After a participant became familiar with how the *Serial* interface worked he or she was then asked to proceed to the main study. Participants were asked to enter the given numbers either in the *Clear*, *Medium* or *Obscured* group within five minutes as quickly and as accurately as possible. Participant were stopped at the end of the given time-frame. During the experiment, participants themselves could not see how much time had passed. After finishing the number-entry task, the experimenter offered each participant a small break. Each participant was then asked to fill out a small on-line demographic questionnaire and the participants were debriefed as to the goals of the study.

4.3.7 Pilot

A pilot study revealed some technical issues with the database connection. There were two participants who were students in computer science. Additionally, the pilot revealed that the angle of how participants looked at the monitor had to be controlled which was then implemented in the main study. Pilot participants were excluded from the main study.

4.3.8 Results

The hypothesis for this experiment was that there is a significant difference in the number of errors between the three conditions *Clear*, *Medium* and *Obscured*. Participants entered in total 2391 numbers for all three groups, which resulted in 44 total errors (1.8%) and 149 corrected errors. 849 numbers were entered in group *Clear* (28 errors), 796 in group *Medium* (12 errors) and 746 in group *Obscured* (4 errors), respectively.

The mean of the numbers entered in the *Clear* group was 84.9 ($SD = 5.45$) whereas in the *Medium* group was 79.6 ($SD = 13.95$) and in the *Obscured* group 74.6 ($SD = 17.61$). Additionally, the mean of the corrected errors in the *Clear* group was 5.2 ($SD = 5.18$), whereas in the *Medium* group was 5.3 ($SD = 6.80$) and in the *Obscured* group 4.4 ($SD = 3.59$). A *Kruskal-Wallis* test was conducted, as error data does not generally follow a normal distribution. There was no significant difference in the total numbers entered between the groups ($H = 2.08; df = 2; p = 0.35$) (see figure 4.15) and the number of corrections ($H = 0.23$; $df = 2$; $p = 0.89$) (see figure 4.16).

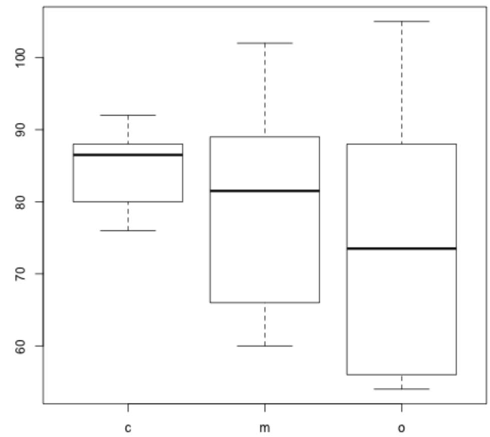


Figure 4.15: Study 6: Boxplot of total numbers entered by all participants in each condition.

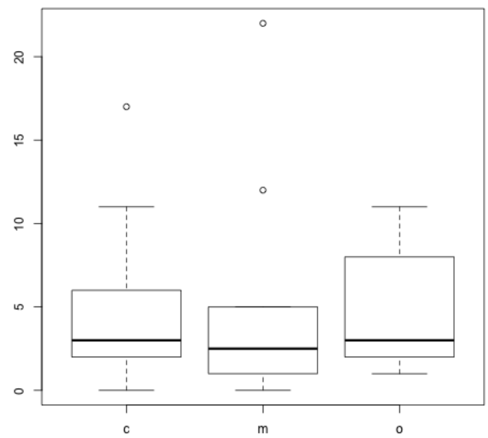


Figure 4.16: Study 6: Boxplot of total corrections made by all participants in each condition.

There was a highly significant difference between the total errors made in the three groups ($H = 9.78$;df= 2; $p = 0.007$). The mean of the errors made in the *Clear* group was 2.8 ($SD = 2.20$) whereas in the *Medium* group was 1.2 ($SD = 1.39$) and in the *Obscured* group 0.4 ($SD = 0.52$) (see figure 4.17).

Additionally, taking into account the number that each individual entered in relation to the errors, the rate of errors was also significantly different ($H = 8.89$; $df = 2$; $p = 0.01$) (see figure 4.18).

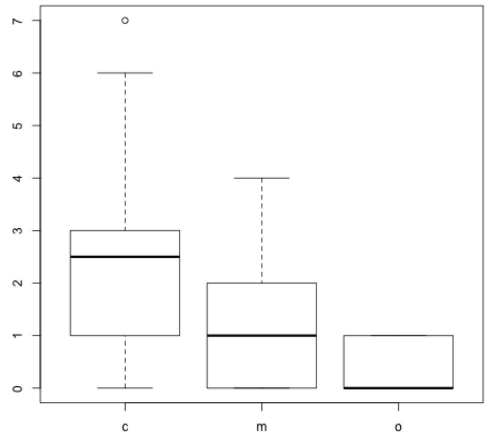


Figure 4.17: Study 6: Boxplot of the total errors made by all participants in each condition

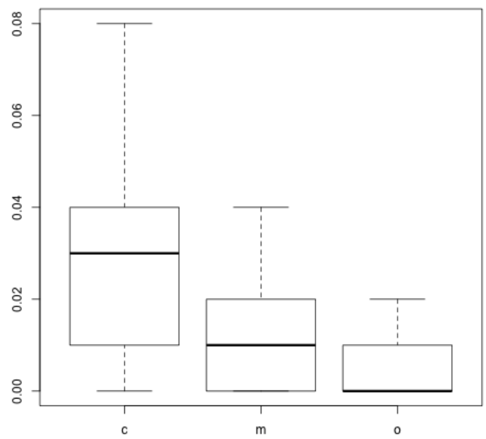


Figure 4.18: Study 6: Boxplot of the error-rate of each participant in each group.

A follow-up multiple comparisons test was used to determine which means among a list of means differ from the rest i.e. where the exact difference was between the tested groups. This test compares significance levels of single and multiple comparisons to be directly compared in contrast to comparing the set as a whole. This specific follow-up test for the *Kruskal-Wallis* has a critical threshold value and the pairs of groups which have observed differences higher than this threshold are considered statistically different (Siegel & Castellan, 1988).

The follow-up multiple comparisons test of the total errors made by each participant in each group revealed, that only the difference between group *Clear* and *Obscured* is significant (observed difference of 11.75 > critical threshold of 9.43). A second follow-up test of the rate entered by each participant in each group revealed also that only the group *Clear* is significantly different from group *Obscured* (observed difference of 11.30 > critical threshold of 9.43).

Eleven participants (36%) made at least one error with only one person making no error in the *Clear* group, four people making no error in the *Medium* group and six people making no error in the *Obscured* group.

4.3.9 Discussion

Like in the text-transcription study (Study Four) the results of this study showed that participants in the *Medium* and *Obscured* group achieved higher accuracy i.e. less errors than those in the *Clear* group. However, only the difference between *Clear* and *Obscured* was significant, which again suggests that the experimental manipulation similar to those suggested by Study Four has worked and there is a trend indicating that a more obscured representation of information can lead to fewer errors being made in data-transcription tasks.

As observed in Study Five, the presentation of information that is too obscure might have the opposite effect of making numbers simply too hard to read and therefore resulting in people abandoning the task. Yet at this moment it is not clear to what extent the different font colour or background colour have contributed to this.

While the results of Study Five suggest that the manipulation used was too strong, the contrast with previous findings suggest that there could be a threshold or sweet-spot of when the *DOV* effect comes into play during a transcription task. It was therefore decided to have a closer look at different levels of obscurity (grades of the faded font) to further examine if a specific level could be detected.

Results in Study Six could indeed detect a *DOV* effect by taking this into account, which speaks in favour of the revised experimental method. Participants entered fewer numbers in the *Obscured* group than in the other groups but this was not significantly different. Furthermore, taking into account that the rate of errors was still significantly different would mean that even if participants are possibly going slower they are also making errors at a lower rate. This could be due to the fact that they are taking slightly more time over each entry but is also very strongly indicative that this is the type of results showing the *DOV* effect coming into play.

In terms of levels of obscurity, the manipulation in Study Six was not as strong as in Study Five meaning that even the *Obscured* group in Study Six (highest fading grade) was not as strong as the one in the previous study. This elevated the *Medium* condition in this study even higher (lower fading grade). This also means, as Study Five was not significant and as the level of obscurity was too high, a potential threshold of when the *DOV* manifests itself must be above the fading grade of Study Five (lower fading grade) but lower than the one from the *Medium* condition (higher fading grade) as only *Obscured & Clear* had a difference in the results of Study Six. It might be that the grade of fading can be higher than the one in the *Obscured* condition of Study Six but an exact measure can not be given at this time as it is unclear which other factors contributed to the non-significance in Study Five (font colour and background).

Overall, there are few corrections, which suggest different mechanisms behind corrected and uncorrected errors. For example, that people use different strategies to correct errors or people are self-detecting their slip errors. While people who commit the errors think their action was correct, those who correct errors are aware of their mistakes suggesting that they are self-evaluating their actions on the fly. Interestingly, whilst the number of committed errors was fewer there was no difference in the number of the corrected errors.

An indication that the *DOV* effect could be a source related effect. If people use a deeper more effortful cognitive processing i.e. a strong encoding in memory then perhaps the corrected errors are made by comparison with memory rather than comparison with the original. According to Salthouse (1986), when people transcribe text they perceive the information at the source and encode it in memory (a view that may be directly applicable to number transcription tasks as well). This is where, in theory, the *DOV* effect interferes. Others describe this as *System 2* engaging in the task (Kahneman, 2008). The *DOV or System 2* engage a more effortful deeper processing of the information and therefore move away from a shallow error prone *System 1* to a better encoding in memory. Participants in this study therefore could have simply made fewer errors due to the better encoding and subsequent retrieval of the information. This subsequently also led to fewer corrections. If they would have made more corrections than that would be in indication that people would notice the error when they see the wrong number on the display and compare it again with the one in the source field. However, as there were not many corrections, not even a significant difference, this was not the case.

Similar to the studies presented in Chapter Three and other comparable studies by Oladimeji et al. (2012) and Thimbleby and Cairns (2010) the rate of number-entry errors is still low (1.8% in this case) but this does not mean it is less important to investigate why these errors happen.

In order to get a better understanding of the manipulations leading to the *DOV* effect the attention in the following study was shifted towards the interesting fact that the manipulation at the source is responsible for the *DOV* effect rather than the display. Given the fact that a manipulation which can successfully provoke the *DOV* effect was already discovered, further studies which focus on a threshold or sweet-spot are planned for the future.

4.4 Experiment 7

Looking back upon the previous studies, it can be seen that it is possible to let people make significantly fewer errors when transcribing information, which is presented to them in a less visible form. Even though there is much room to investigate the application of additional stimuli such as different font-types as in Diemand-Yauman et al. (2011) the applied font colour grade seems to be enough to stimulate and subsequently to detect the hypothesised effect. This seems to be the case for both text- and number-transcription tasks.

Additionally, from the previous work on the *DOV* effect it would be expected that it is the source information which is important to stimulate the effect. This seems to be the case in Study Four but was not explicitly considered and in the subsequent experiments both the to-be-transcribed source and the numbers in the target display were presented in the same font colour. This was done partly by a way of replication. This means, for example, that in the *Obscured* condition the source and the target area would have been displayed in the same grey faded font colour. Participants could have been stimulated by the stimuli present at the target field. As this has never been done before in an interactive context, an interesting question is, which of the two design elements is responsible for stimulating the *DOV* effect? The corrected errors might be an important factor if the target is obscured or not to discern if the *DOV* is a source or target related effect.

Additionally, there is the theory that people tend to use a memory-intensive, less accurate strategy in transcription tasks and this can lead to higher error rates (Gray et al., 2006). This would speak against increasing the effort to access information. Moreover this would also rule-out that the information which is harder to read at the source would have an effect on reducing errors as this would simply just increase the error rate. However, the results gained from the text-entry study and the previous number-entry study suggest otherwise as there was a decrease in errors when people perceived the information in an obscured way.

In conclusion, increasing the effort for people to access the information by obscuring the to-be-transcribed information would align with the theory surrounding the *DOV* effect. People would have to increase their effort and think harder to access the presented information and therefore encode it better in their memory.

4.4.1 Aims & Hypothesis

The aim of this experiment was to investigate which obscured interface design feature (target or source) can stimulate the *DOV* effect and let people make fewer errors in transcription tasks. Previous studies matched the visibility grade in the obscured conditions on both the source and the target display. Therefore the hypotheses for this experiment was:

H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.

H_1 : There will be a significant difference in the number of errors made between the conditions *Clear & Obscured*.

H_2 : There will be interaction effects on errors or corrected errors due to the source and the target being greatly different.

4.4.2 Participants

Participants were all students and randomly recruited by personal invitation or by email. There were ten participants (three were women). The age of the participants was between 19 and 25 (mean 22.7; *SD* 2.53) years. None of the participants had previous experience of working in the healthcare sector and only three people stated that they were familiar with an infusion pump. Participants were mostly familiar with a calculator-based number-entry interface as used in the study (except two). All participants were again rewarded with chocolate and as usual all of the participants were randomly assigned to the specific groups.

4.4.3 Number-Transcription Task

The four tasks set to participants were to transcribe as many numbers as possible in the given time-frame of five minutes for each condition. Other than that, the tasks itself where identical to that of the previous study (Study Six) in terms of number display (structure) and distribution (decimal to integer).

4.4.4 Design

In order to facilitate data collection a within-subjects design was chosen for this study resulting in four experimental conditions. The two independent variables were the grade of the visibility of the presented numbers in the source display and in the target input area resulting in four experimental conditions. The dependent variables were identical to the previous study. As in the previous study the numbers are randomly generated at the start of each trial. To avoid ordering and rehearsal effects, the order in which participants saw the four conditions was fully counterbalanced. This study used the same font-colour coding for the *Clear* and *Obscured* condition as in Study Four and Study Six.

4.4.5 Materials

The same computational platform as in the previous number-transcription study was chosen. The computer used was again an Apple Macbook Pro 13" connected to an external monitor with a 17" size display at 1920 × 1080 resolution. Also, the browser used for the experiment was Mozilla Firefox 10 running under Mac OSX 10.7 (Lion). Hardware preparations were again the same as before i.e the keyboard was labelled with stickers accordingly.

Instructions and an informed consent form were printed and given to participants to read and to sign before they started the trials (see Appendix B.4.1). Participants were instructed to fill out a small demographic questionnaire after each condition (see Appendix B.4.2) followed by a final questionnaire at the end of the study (see Appendix B.4.3). The questionnaires between the tasks contain questions about the difficulty of the task, task engagement as well as questions about the effort a participant put into the task. As the design of the experiment was a within-participants design meaning that one participant would experience all the conditions sequentially it was required to place these questions after each task to be able to identify any changes between the conditions in terms of demographics. Some of the questions focus on task difficulty therefore only one questionnaire at the end might hide this information for the individual tasks. All these questionnaires were provided on-line via Google forms. As previously a smart-phone was used to manage the time.

4.4.6 Number-Transcription Game

There were four versions of the interface used in the study according to the experimental conditions. The only modification of the game were the unmatched number presentation conditions. Other than that the game was identical in its design and mechanics (see figure 4.19).

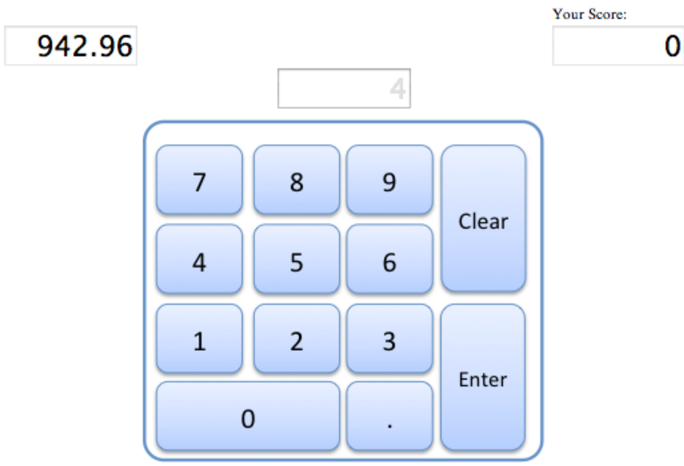


Figure 4.19: Study 7: Number-transcription game with unmatched source and target display

4.4.7 Procedure

This study was also conducted in an empty office in the Computer Science building. All participants were again run in individual sessions which were managed by email and personal invitation. The monitor and keyboard were prepared accordingly so that participants looked straight at the display and only interacted with the number-pad of the keyboard.

A participant was welcomed and given a copy of the informed consent form and the instructions to read and sign. After that, the training started, where participants were asked to enter ten training trials which were separate from the main experiment and also randomised between participants. After a participant became familiar with the controls he or she was then asked to proceed to the first main task. Participants were then asked to enter the given numbers in one of the four conditions (depending on the randomisation of the groups) within five minutes. The experimenter stopped the participant after the time limit was up. After each main task participants were asked to fill out a small on-line questionnaire and then asked to take a short break. The study then continued with the second task by letting participants enter again the given numbers within five minutes in the second condition. The same procedure was followed for all the four conditions. After finishing all four number-transcription tasks and the following small questionnaires between the main tasks, participants were asked some demographic questions in form of an additional on-line questionnaire.

4.4.8 Pilot

A small pilot study with two participants (two research students) was conducted to test the experimental design. Only minor data storage errors were detected and later corrected. The participants of the pilot were excluded from the main study.

4.4.9 Results

The data which is analysed in this study is non-parametric data (error-data). As this analysis is looking for two-way effects, a (2 x 2) *Analysis of Variance* (ANOVA) is the suitable test, as there is no well established test for non-parametric data. However, main effects where they are seen in the ANOVA are confirmed with a non-parametric *Wilcoxon* test.

The hypothesis for this experiment was that there would be a significant difference between the four experimental conditions. Participants entered in total 3494 numbers for all four groups, which resulted in 102 total errors (2.92%). 862 numbers were entered in condition *Clear & Clear* (20 errors), 859 in condition *Clear & Obscured* (30 errors), 883 in condition *Obscured & Clear* (18 errors) and 890 in condition *Obscured & Obscured* (34 errors), respectively. Table 4.1 shows the descriptives for the total numbers entered in each condition.

		Display		
		Clear	Obscured	Row Mean
Source	Clear	86.2 (14.00)	88.3 (12.60)	86.1 (13.51)
	Obscured	85.9 (13.76)	89.0 (16.00)	88.7 (14.01)
	Column Mean	87.6 (14.72)	87.1 (12.87)	

Table 4.1: Study 7: Descriptives: Mean (SD) for the total numbers entered

An alpha level of .05 was used for all statistical tests. The data was analysed using a two-way (2 x 2) ANOVA for a within-participants design, with the grade of the visibility of the source and the target display as related samples variables.

In regards to the numbers entered in total, there was no main effect of the Source, $F(1, 32) = 0.08; p = 0.78; \eta_p^2 = 0.001$, with participants transcribing about the same amount of numbers in both conditions *Clear* and *Obscured*. There was also no main effect for the visibility of the Display $F(1, 32) = 0.001; p = 0.98; \eta_p^2 = 0.001$. There was also no interaction effect (Source x Display) $F(1, 32) = 0.353; p = 0.56; \eta_p^2 = 0.001$ (see figure 4.20).

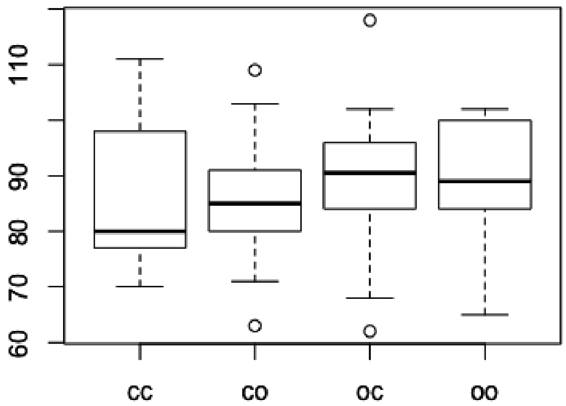


Figure 4.20: Study 7: Boxplot of total numbers entered by each participant in each condition.

In regards to the total errors, there was no main effect of the Source, $F(1,32) = 0.008; p = 0.93; \eta_p^2 = 0.001$. There was also no main effect for Display $F(1,32) = 0.027; p = 0.87; \eta_p^2 = 0.003$ and also no interaction effect (Source x Display) $F(1,32) = 1.62; p = 0.21; \eta_p^2 = 0.07$ (see figure 4.21). Table 4.2 presents the descriptives in relation to the uncorrected errors.

		Display		
		Clear	Obscured	Row Mean
Source	Clear	2.0 (2.00)	3.0 (3.10)	2.5 (2.60)
	Obscured	1.8 (1.48)	3.4 (3.06)	2.6 (2.48)
	Column Mean	2.7 (2.62)	2.4 (2.46)	

Table 4.2: Study 7: Descriptives: Mean (SD) for the total errors made

Only one participant made no errors during the whole study. Three participants made no errors in both the *Clear & Clear* and the *Clear & Obscured* condition whereas two participants made no errors in the *Obscured & Clear* condition (80% made at least one error) and *Obscured & Obscured* condition. The highest uncorrected error-count and number of corrected errors was recorded in a *Obscured & Clear* condition (ten errors in one task, 63 corrections).

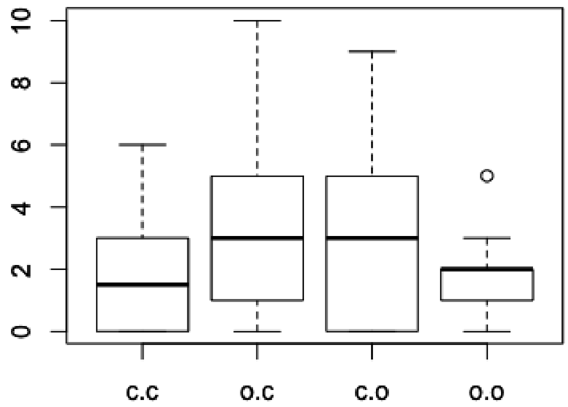


Figure 4.21: Study 7: Boxplot of total errors made by each participant in each condition.

In regards to the error rate, there was no main effect of the Source, $F(1, 32) = 0.04; p = 0.84; \eta_p^2 = 0.001$. There was also no main effect for Display $F(1, 32) = 0.01; p = 0.94; \eta_p^2 = 0.003$ and also no interaction effect (Source x Display) $F(1, 32) = 1.67; p = 0.21; \eta_p^2 = 0.08$ (see figure 4.22). Table 4.3 presents the descriptives in relation to the total rate of errors.

		Display		
		Clear	Obscured	Row Mean
Source	Clear	0.02 (0.02)	0.03 (0.03)	0.03 (0.02)
	Obscured	0.02 (0.02)	0.04 (0.03)	0.03 (0.02)
	Column Mean	0.03 (0.03)	0.03 (0.03)	

Table 4.3: Study 7: Descriptives: Mean (SD) for the total error rate

A closer look at the order of the consecutive number-transcription tasks revealed that the number of errors increased after participants competed the first task i.e. transcribed the first set of numbers. Figure 4.23 shows the total errors made by participants ordered by the task.

Participants made in total 701 corrections during the study. 163 corrections were made in condition *Clear & Clear*, 191 in condition *Clear & Obscured*, 124 in condition *Obscured & Clear* and 223 in condition *Obscured & Obscured*, respectively. However, there was no main effect for Source, $F(1, 32) = 0.018; p = 0.89; \eta_p^2 = 0.001$. There

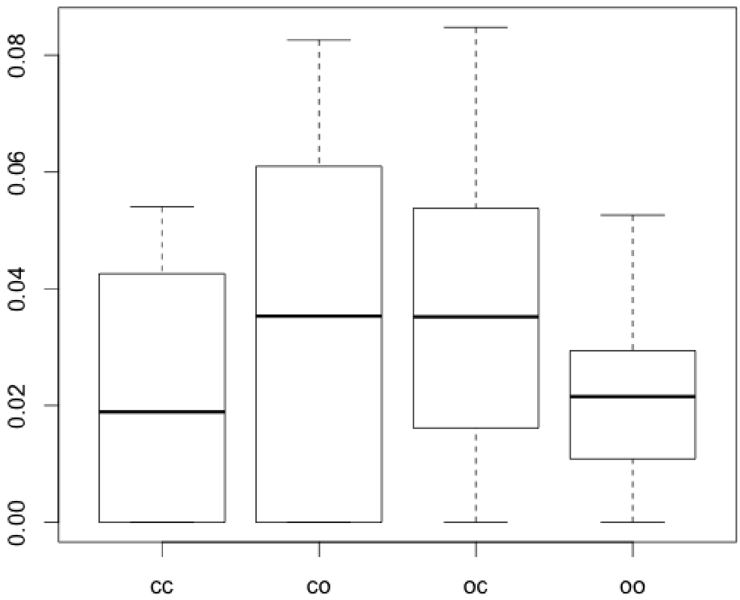


Figure 4.22: Study 7: Boxplot of total error rate by each participant in each condition

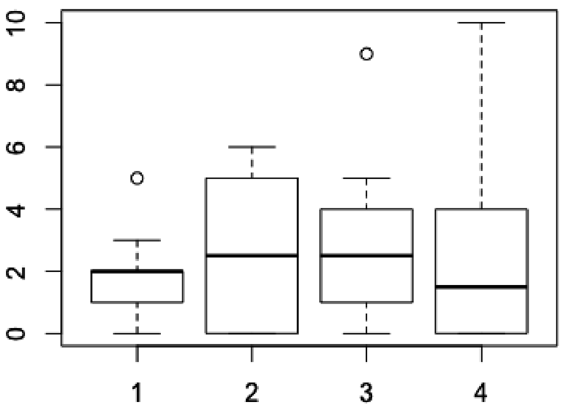


Figure 4.23: Study 7: Boxplot of the total errors presented in the order of tasks

was also no main effect for Display $F(1, 32) = 0.123; p = 0.73; \eta_p^2 = 0.008$ and also no interaction effect (Source x Display) $F(1, 32) = 0.024; p = 0.88; \eta_p^2 = 0.03$ (see figure 4.24). Table 4.4 presents the descriptives of the corrections.

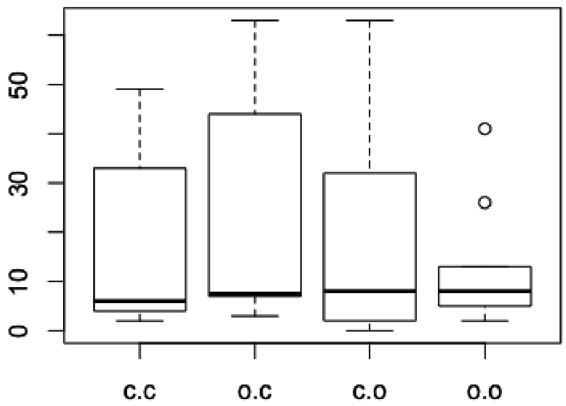


Figure 4.24: Study 7: Boxplot of the corrections made by participants in each condition

		Display		
		Clear	Obscured	Row Mean
Source	Clear	16.3 (18.55)	19.1 (23.85)	17.7 (20.85)
	Obscured	12.4 (12.11)	22.3 (23.50)	17.4 (18.91)
	Column Mean	19.3 (20.85)	15.6 (18.91)	

Table 4.4: Study 7: Descriptives: Mean (SD) for the total corrections made

4.4.10 Discussion

No significance of the total errors, error rate, numbers entered or corrected errors could be measured between the four conditions. This study has, like previous experiments, shown that although a large quantity of numbers have been transcribed, the error presentation rate was well below 5% (2.9% in this Study).

The results have shown that a within-participants design seems to be a suboptimal choice for number-transcription experiments as there are indications in the data, which suggest that they were less engaged in the task after they completed the first number-transcription condition (see figure 4.23). Participants seem to make more errors as the tasks progressed (regardless of condition). The task itself is simple yet it seems too monotonous if repeated four times over twenty minutes. Results of the demographics between the tasks also indicate a downwards trend in engagement. While all participants initially stated that they were highly focused on the task (4-5 on the 5-point likert scale) they were already less focused after the second task (1-3 on the likert scale). An independent sample i.e. between-participants design would perhaps allow the DOV effect to present itself.

Although it was not possible to detect a significant difference in the amount of errors between the four conditions, this study has shown that, for example, only one participant made no errors during the entire study. All others made one or more errors. This can be seen as further indication that in even small tasks errors happen and transferred to transcription tasks in safety-critical domains such as the healthcare domain can cause serious adverse events.

4.5 Experiment 8 - The Revised Design

Two of the previous studies (Study Four and Study Six) indicated that there is potential for improvement of accurately transcribing information once the to-be-transcribed information is presented in a harder to read font (faded colour). However, only conditions where the disfluency of the source was matched to the disfluency of the display were measured. Study Seven was the first attempt to measure all possible conditions i.e. combinations but was unable to detect the DOV effect possibly due to its disadvantageous experimental design (within-participants design). Study Eight is therefore identical to Study Seven but addressed the previous issue by utilising a between-participants design.

4.5.1 Aims & Hypothesis

This study had the identical aim as Study Seven namely to measure whether the introduced disfluency at the source or the at the input display is responsible for the DOV

effect to present itself. This also led to the identical hypotheses:

> H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.
>
> H_1 : There will be a significant difference in the number of errors made between the conditions *Clear & Obscured*.
>
> H_2 : There will be interaction effects on errors or corrections due to the source and the target being greatly different.

4.5.2 Participants

Participants were again randomly recruited and were almost all students (one resource developer). There were 40 participants (25 women). The mean age of all the participants was 22.57 with a *SD* of 2.78 making them between 19 and 25 years old. Participants came from a variety of countries such as UK, Germany, Latvia, Thailand, Saudi Arabia with a majority coming from China (22 people). Three of the participants had previous experience of working in the healthcare sector and four stated that they were familiar with an infusion pump. Twenty-one participants were not familiar with a calculator based number-entry interface as used in the study. Only one participant found the task difficult but made no more that one error and four corrections in total. The person from Saudi Arabia was the only one who did not enjoy the task and stated that the study was tiring and exhausting. All the participants were randomly assigned to the groups. Additionally, all participants who voluntarily took part in the experiment were rewarded with chocolate.

4.5.3 Task, Materials, Design and Interface

The number-transcription task was identical as to the one in Study Seven. In order to address previous unwanted effects a between-participants design was chosen for this study. There were again the same four versions of the interface used in the study as in Study Seven. The independent variables were again the grade of visibility of the source and input display. All the dependent variables were also identical to the previous study. The same font-colour codes were used again for both conditions identical to Study Four, Six and Seven.

The electronic set-up, the number-transcription game itself and the paper-based instructions as well as the informed consent were also identical (see Appendix B.5.1). There was a difference in the demographic questionnaires. Unlike Study Seven, this study only used one demographic questionnaire at the end of the experiment (see Appendix B.5.2).

4.5.4 Procedure

The experiment was conducted in an empty office in the Computer Science building and at the Morell Library (see figure 4.25) at the University of York. All participants were run in individual sessions which were managed by email and personal invitation. The hardware and monitor set-ups were performed as before and no other program was running on the computer. The browser used for the experiment (Mozilla Firefox) was therefore set to full screen mode.

Figure 4.25: Study 8: Experimental set-up - Morell Library

Each participant was first welcomed and then given a copy of the informed consent form and the instructions to read and to sign. The training trials were again ten numbers and after participants completed the training they were asked to proceed to the main study. Participants were then asked to enter the given numbers in one of the four conditions within the usual five minutes. Participants were stopped after the given time-frame was up. After finishing the main number-transcription task, participants were asked to complete a few demographic questions. This was followed by the debriefing of a participant.

4.5.5 Pilot

A small pilot study with one participant (research student at the department of computer science) was conducted to test the experimental design. One data link error to the database was detected and fixed other than that no errors or design flaws were detected.

4.5.6 Results

The data in this experiment was also analysed using a $(2X2)$ *Analysis of Variance* (ANOVA), as there is no well established test for non-parametric data. Main effects where they are seen in the ANOVA are confirmed with a non-parametric *Mann-Whitney* test.

Participants entered in total 3232 numbers combined in all the groups, which resulted in 65 total errors (**2%**). 801 numbers were entered in condition *Clear & Clear* (21 errors), 822 in condition *Clear & Obscured* (23 errors), 827 in condition Obscured & Clear (17 errors) and 782 in condition *Obscured & Obscured* (4 errors), respectively. Table 4.5 illustrates the relevant descriptives in relation to the total numbers entered in each condition.

There was no main effect of the Source, $F(1, 36) = 0.031; p = 0.86; \eta_p^2 = 0.01$ in regards to the total numbers entered. There was also no main effect for Display $F(1, 36) = 0.090; p = 0.77; \eta_p^2 = 0.01$ and also no interaction effect (Source x Display)

		Display		
		Clear	Obscured	Row Mean
Source	Clear	80.1 (11.45)	82.2 (13.80)	81.15 (12.39)
	Obscured	82.7 (11.16)	78.2 (13.87)	80.45 (12.47)
	Column Mean	81.4 (11.09)	80.2 (13.62)	

Table 4.5: Study 8: Descriptives: Mean (SD) for the total numbers entered

$F(1, 36) = 0.68; p = 0.21; \eta_p^2 = 0.02$ (see figure 4.26).

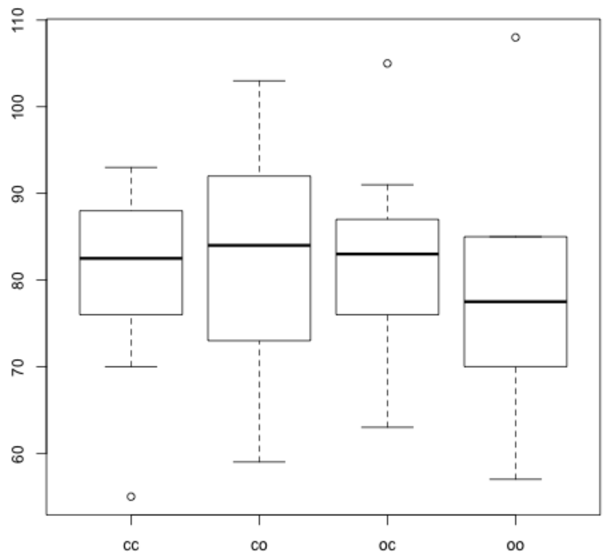


Figure 4.26: Study 8: Boxplot of total numbers entered by each participant in each condition.

In view of the total errors made, there was a main effect for Source $F(1,\ 36) = 6.31;\ p = 0.016; \eta_p^2 = 0.15$) but no main effect for Display $F(1,\ 36) = 1.442;\ p = 0.24; \eta_p^2 = 0.04$) nor an interaction effect: $F(1,\ 36) = 2.682\ p = 0.11; \eta_p^2 = 0.07$). The main effect for source was confirmed by a *Mann-Whitney* test $p = 0.007; W = 296.5$ (see figure 4.27). The following table 4.6 illustrates the descriptives regarding the total errors made by condition:

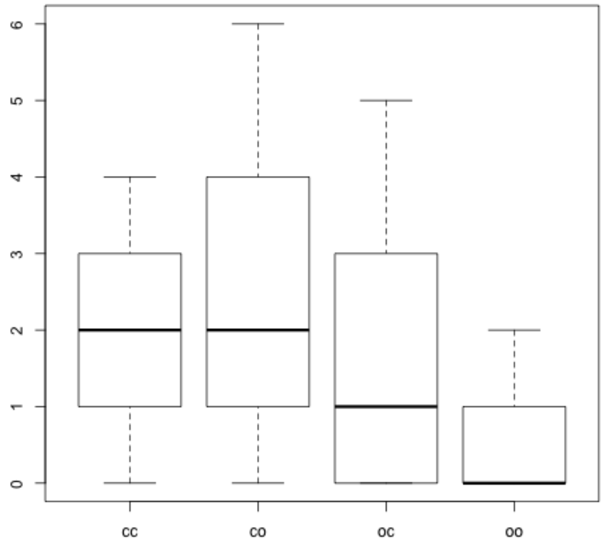


Figure 4.27: Study 8: Boxplot of total errors made by each participant in each condition.

		Display		
		Clear	Obscured	Row Mean
Source	Clear	2.1 (1.20)	2.30 (1.83)	2.20 (1.50)
	Obscured	1.7 (1.77)	0.40 (0.70)	1.05 (1.47)
	Column Mean	1.9 (1.48)	1.35 (1.66)	

Table 4.6: Study 8: Descriptives: Mean (SD) for the total errors made

In view of the total error rate, there was also a main effect for Source $F(1, 36) = 9.01$; $p = 0.004; \eta_p^2 = 0.20$ but no main effect for Display $F(1, 36) = 2.20$; $p = 0.15; \eta_p^2 = 0.06$ nor an interaction effect: $F(1,\ 36) = 1.53\ p = 0.23; \eta_p^2 = 0.04$. The main effect for source was again confirmed by a *Mann-Whitney* test $p = 0.007; W = 296.5$ (see figure 4.28). Notably, the *Mann-Whitney* did not reveal a significance for the display manipulation in this regard $p = 0.16; W = 251.6$. Table 4.7 presents the descriptives in relation to the error rate.

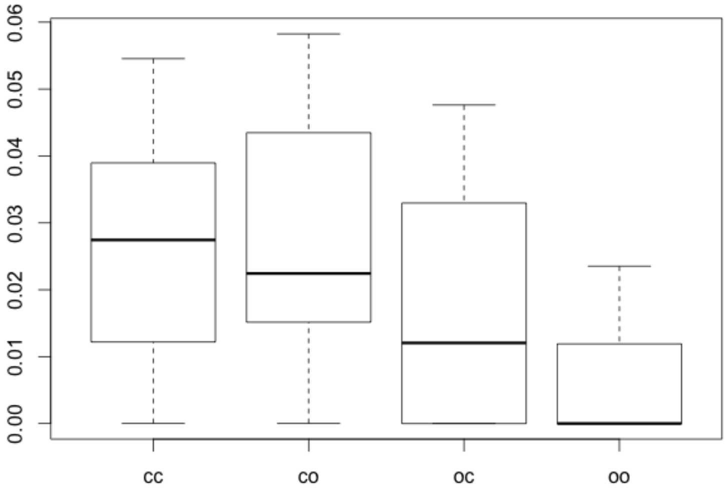


Figure 4.28: Study 8: Boxplot of total error rate by each participant in each condition.

		Display		
		Clear	Obscured	Row Mean
Source	Clear	0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
	Obscured	0.02 (0.02)	0.01 (0.01)	0.01 (0.02)
	Column Mean	0.03 (0.02)	0.02 (0.02)	

Table 4.7: Study 8: Descriptives: Mean (SD) for the total error rate

Twelve participants made no errors during the whole study. One participant made no error in the *Clear & Clear* condition, one participant made no errors in the *Clear & Obscured* condition, three participants made no errors in both the *Obscured & Clear* condition whereas seven participants made no errors in the *Obscured & Obscured* condition (70% of all people made at least one error). The highest error-count and number of corrected errors was recorded in a *Clear & Obscured* condition (6 errors in one task, 18 corrections).

Participants made in total 274 corrections during the study (see figure 4.29). 52 corrections were made in condition *Clear & Clear*, 55 in condition *Clear & Obscured*, 119 in condition Obscured & Clear and 48 in condition *Obscured & Obscured*, respectively. Table 4.8 shows the descriptives for the corrections in this study.

		Display		
		Clear	Obscured	Row Mean
Source	Clear	5.2 (4.83)	5.5 (5.12)	5.35 (4.85)
	Obscured	11.9 (10.0)	4.8 (5.55)	8.35 (8.69)
	Column Mean	8.55 (8.39)	5.15(5.21)	

Table 4.8: Study 8: Descriptives: Mean (SD) for the total corrections made

Regarding the total corrections made, there was no main effect for Source $F(1, 36) = 1.989$; $p = 0.16; \eta_p^2 = 0.05$ and no main effect for Display $F(1, 36) = 2.555$; $p = 0.12; \eta_p^2 = 0.07$ nor an interaction effect: $F(1, 36) = 3.026$ $p = 0.09; \eta_p^2 = 0.08$ (see figure 4.29).

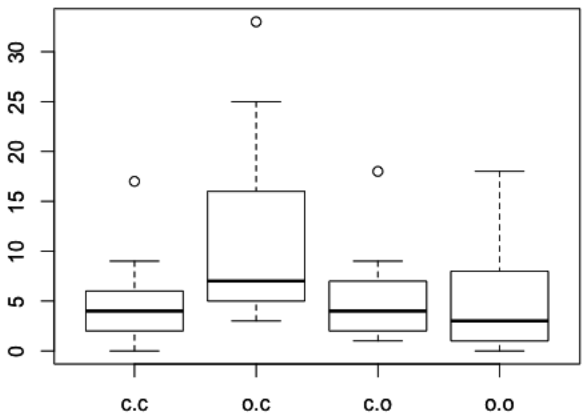


Figure 4.29: Study 8: Boxplot of the total corrections by condition

4.5.7 Discussion

The hypothesis (H_1) for this experiment was that there would be a significant difference in the number of errors between the four conditions and therefore a DOV effect could be detected. Indeed, there was a significant difference in the total errors and the total error rate committed between the conditions, forming a stronger case for supporting the argument in the literature that by introducing disfluency or in other words by making information harder to access or in this case harder to read can be beneficial in terms of accuracy in transcription tasks (Diemand-Yauman et al., 2011).

The results showed that was a significant difference in terms of source, where people perceive disfluent information is significantly different from the other conditions in terms of errors being made but no significant effect of display. This would confirm that the *DOV* is a source related effect and possibly can not be influenced by disfluent information presented at the target area. The effect size is modest meaning that 15% of the variance in the total error data are due to the manipulation at the source. In terms of the error-rate this is even higher (20%). Though a modest effect, in safety critical domains even modest effects can have a considerable impact.

Additionally, no significance of the total numbers entered or corrections could be measured between the four conditions indicating that no speed-accuracy trade-off was present. It appears that no effect in the total corrections contributed to the fact that the *DOV* effect is a purely source related effect and does not influence error detection (i.e. participants checking the display). By looking at the boxplot (figure 4.29) it appears that the highest variation in the data occurred in the *Obscured Source & Clear Display* condition indicating that people have made more corrections in this condition but not significantly more.

An explanation for this could also be that once the Source is obscured people encode the numbers better in their memory therefore compare the numbers in their memory not via the display and therefore make fewer errors in the first place. In case they do make a mistake then as a consequence they notice this because of the better encoding. Further to this, additional motor-movement errors might also come into play which means that once a number is encoded in memory and then retrieved it is translated into motor-movements (Salthouse, 1986) it could be that at this stage an interference causes people to type the wrong number although they had the correct number in mind and then notice this mistake because of the better encoding therefore make a higher number of corrections in this case. This is however at this stage a possible speculative explanation but not confirmed with enough evidence.

These results further strengthen the argument that within-participants designs are not advantageous in laboratory experiments investigating these particular number-transcription tasks. In addition, this study has, like previous experiments, shown that although a large quantity of numbers has been entered the presentation rate of the errors was again well below what is typically considered systematic, namely of 5% (Abelson, 1995).

4.6 Overall Discussion & Conclusions

This chapter presented a series of studies motivated by the healthcare context outlined earlier to examine the *DOV* effect for text- and number-transcription tasks. The first study (Study Four) used the text-transcription game *TypoMadness* as research instrument to increase the error-rate and produced a very promising effect. People made significantly fewer errors without getting slower when transcribing text phrases in a grey faded font as previously proposed by Diemand-Yauman et al. (2011). This was seen as key-stone experiment and greatly inspired further experiments in the domain of number transcription tasks. Study Five was the first attempt to do this in a class room setting but resulted in a non-significant inconclusive outcome due to the environment and the manipulation used in this study. The classroom setting contributed noise factors and the colour coding of the manipulation (grade of obscurity) was disadvantageous in terms of readability. While it was initially expected that none of the colour coding base is effecting the *DOV* effect, it suggested that this manipulation was too strong in this case making it too hard to read and therefore undermining the *DOV* effect bringing into question a possible threshold of obscurity of when the effect might be seen. It was inconclusive at that stage if the inability to produce the effect was due to the base colour of the manipulation or due to the grade of fading and background or the environmental effects. It is possible that all three of them together somehow contributed to prevent the *DOV* effect to be seen. Study Six was a much more controlled laboratory study and took the issues of Study Five into account. The same successful manipulation as in Study Four was used (although not as strong) and resulted in people making significantly fewer errors (without slowing down during the task) in the *Obscured* conditions. Moreover, this study included an additional *Medium* condition to identify a possible threshold of then the *DOV* effect could come into play. An indication of such a sweet-spot could be identified but not conclusively fixed. Moreover, due to the number of corrections and to previous results in the literature (Diemand-Yauman et al., 2011; Sungkhasettee et al., 2011), it was assumed that the *DOV* effect could be a Source related not a Display related effect. This would mean that once people read the number (or text) at the source then a better encoding (triggered by the deeper processing mechanism behind the effect (Kahneman, 2008) would be responsible for them to retrieve the information better from memory and therefore make fewer errors. Additionally, comparisons before and after an error was made is due to comparisons in memory rather than due to comparisons with the number presented at the display (where people re-enter the numbers or text). Study Seven aimed to identify whether the Source or Display is responsible for the *DOV* effect but resulted in a non-significant outcome. The number of errors between the conditions indicate that participants were already less engaged after the first task (within-participants design) due to the nature of the task being possibly too monotonous. Study Eight used a between-participants design, which resulted in the detection of the *DOV* effect. Both the errors and error rate was lower in the *Obscured* condition. Moreover, it became also apparent that the effect is related to the Source which is consistent with the theory surrounding the *DOV* effect (Alter et al., 2007; Diemand-Yauman et al., 2011) and strongly suggests that the *DOV* effect is widely applicable in the context of number- and text-transcription tasks. The effect size of the study suggests a modest effect however, even modest effects can have a critical impact in safety-critical domains.

The key findings of Chapter Four are:

- The *DOV* effect does work in both text- and number-transcription tasks
- The *DOV* seems to be a source related effect
- A within-participants design is disadvantageous for studies as presented

The studies presented here are essentially a first step into unknown territory and while the thesis is mapping this territory much is still to be discovered of how the *DOV* effect works. The key findings of this chapter lead to several interesting directions of future studies.

When focusing on design, the discovered results represent a potentially important contribution for the design of future medical devices and transcription tasks in general. However, there is also the possibility that this effect will wear off over time. The literature described the *DOV* effect being constructed upon concepts of DPT such as two distinct cognitive processes (Kahneman, 2008; Kuehl et al., 2014). According to DPT the part in people's mind that thinks harder and more effortful is also a slow learner. So in theory the effect might endure for a while but may disappear over time. Future work will therefore need to focus on a long-term study, to investigate how long the discovered effect lasts in transcription tasks.

Additionally, fatigue is another issue that needs to be examined in this context. There is evidence in the literature that, for example, people who are on extended work duty possess a significantly decreased level of vigilance. For example, workers in the healthcare domain (Scott, Rogers, Hwang, & Zhang, 2006). This is something future studies need to take into account. Furthermore, the grade of frustration is another interesting aspect to consider. For example, if the *Obscured* text is too strong i.e. if the text or the number is too hard to read then there is the potential that people will get frustrated in entering numbers (or text) and simply lower their efforts towards a successful completion of the task or even abandon the task completely. This is something that needs to be explored - especially in light of possible longitudinal effects. Although, in the laboratory Study Five presented here it appears that people were less engaged in the task whereas in contrast to other studies (Study Four, Six, possibly Seven and Eight) the grade of obscurity resulted in the effect being present. Yet the studies presented here only have a time-frame of a few minutes and such effects might occur after a longer period of time.

Moreover, as healthcare personnel do not usually transcribe information from an electronic source but also from paper-based sources there is further room to explore the application of the findings in this medium.

These are several avenues of directions future *DOV* studies could take. All of them are interesting and give further insight into how the *DOV* effect works. However, what was striking in this thesis is the low error presentation rate of all the studies presented here and in this field - even with an interface that is more susceptible to errors (Oladimeji et al., 2011). Further to that all the studies presented here (except Study Five) are ideal laboratory studies and give little insight how the *DOV* effect might perform in a real-world trial i.e. how robust the effect is once distractions come into play. The following Chapter Five will present studies to investigate the behaviour of the *DOV* effect in relation to distractions.

The results in this chapter revealed important findings based on a new approach to understanding the errors in context of transcription tasks. This may have significant implications for the design of tasks in safety-critical systems.

#	Interface	IV	DV	Design	Hypothesis	Outcome	Conclusions
4	Keyboard	Text presentation (Clear vs Obscured)	Number of sentences Number of characters Score (correct-errors)	Between (10 each)	Fewer errors in Obscured than Clear	Significant difference in score (errors) between Clear and Obscured; No significance in total numbers entered; Fewer errors in Obscured; DOV presented;	DOV effect present – a novel way to reduce errors in text transcription tasks; No speed accuracy trade-off; Urge to try in number-entry studies;
5	Serial	Number presentation (Clear vs Obscured)	Total numbers entered Total errors made Total corrections made Total error rate	Between (26/27)	Fewer errors in Obscured than Clear	Significant difference in numbers entered; No significant difference in terms of errors corrections nor rate; Noise factor high; Manipulation too strong;	The DOV effect could not be seen; Participants slowed down but also made more errors; Working memory probably depleted; Noise factors & lack of strong experimental control; Chosen colour scheme was probably disadvantageous and too ambitious; Attempt to increase the generalisability of the manipulation failed; Re-run in a more controlled environment and fall back to default colour scheme which was successful in the first place; Instead of exploring further colour schemes go back to original design but explore if a possible sweet spot can be found;
6	Serial	Number presentation (Clear/ Medium and Obscured)	Total numbers entered Total errors made Total corrections made Total error rate	Between (10 each)	Fewer errors in either Obscured and Medium than Clear	Significant difference in errors and error rate; No significant difference in total numbers entered nor corrections; Follow up test revealed difference between Clear and Obscured	DOV presented; No speed accuracy trade-off; Indications for a possible sweet spot but only indicative; Number of corrections and literature indicate source effect; No difference in corrections supports theories surrounding the DOV effect that it is a source related effect; This needs further investigation; A possible range for a sweet-spot could be identified;
7	Serial	Number presentation at the Source (Clear vs Obscured); Number presentation at the Display (Clear vs Obscured)	Total numbers entered Total errors made Total corrections made Total error rate	Within (10)	Fewer errors in Obscured than Clear Fewer errors (interaction) in Source condition	No significant difference in any of the conditions in terms of total numbers, corrections, errors or rate;	DOV not present; Number of errors between the conditions indicate that participants were less engaged as task progressed; task too monotonous for a within-participants design; This was also reflected in the demographic questionnaires; Recommendation: do not use within-participants design for future number-transcription experiments; Theory surrounding source or display inconclusive; Follow-up study required;
8	Serial	Number presentation at the Source (Clear vs Obscured); Number presentation at the Display (Clear vs Obscured)	Total numbers entered Total errors made Total corrections made Total error rate	Between (20 each)	Fewer errors in Obscured than Clear Fewer errors (interaction) in Source condition	Significant difference in errors and error rate; No significant difference in total numbers entered nor corrections; Tests revealed difference in Source but not Display;	DOV present; Significantly fewer errors (and error rate) in Obscured than Clear; Tests revealed Source is significantly different in terms of errors and rate leading to the conclusion that the DOV can be provoked by obscuring source information in transcription tasks; no speed accuracy trade-off; no difference in corrections which might further indicate that the DOV is source related; effect size is modest (15% for errors and 20% for error-rate)

Table 4.9: Studies 4 - 8: Summary of findings

Chapter 5

Extended studies on design features with underlying cognitive processes

"The brain is like a muscle. When it is in use we feel very good. Understanding is joyous."
Carl Sagan

5.1 Introduction

The previous chapter presented a series of studies, which investigate the relevance of the *DOV* effect to the reduction of number-entry errors in the context of transcription tasks. These studies are regarded as core experiments of this thesis as they primarily analyse if the *DOV* effect itself can be stimulated, detected and what design feature (Source or Target) is responsible for the effect to be activated.

The results surrounding the *DOV* effect in an interactive context as presented in this thesis are a novel contribution to knowledge. That also means that much is still unknown about this effect and how it ultimately behaves in the real-world despite the growing body of theories in the literature surrounding it. While Section 4.6 in Chapter Four presented several avenues for further research it was decided to explore how robust the *DOV* effect is once distractions come into play. In terms of real-world distractions and existing safety protocols, studies, which provide analogous distractions, would be useful to establish the interplay between these concepts. This was done by introducing distractors into the experimental design to divert attention away from the primary transcription task to increase the errors people make and to get a better understanding how the effect may behave outside of an ideal laboratory environment.

This chapter presents two studies, which add further knowledge to the core research questions covered in Chapter Four. The first study (Study Nine) investigated how robust the *DOV* effect is related to attention, effort and distractions by introducing two auditory distractors (*numerals* and *letters*). Interesting features in relation to the audio distractors have been discovered, which were not expected. Study Ten then took the promising results of Study Nine into account and refined the auditory distractors based on suggestions in the literature. Study Ten evaluated how the *DOV* reacts once *narrative speech* and *white-noise* auditory distractors are present. This study showed clearer results in terms of the *DOV* and its performance in an environment with auditory distractors.

5.2 Experiment 9

In order to examine the behaviour and the potential of the *DOV* effect as a way to reduce transcription errors in safety-critical domains, it was decided to consider the robustness of the effect in a distracting environment, such an environment can be found in busy hospital wards or on a flight deck. The *Air Inter Flight 148* incident (see Chapter Two) described a situation where all safety layers failed which led to disaster. Among these layers was a distraction i.e. a moment where the Captain's attention was refocused which caused him that led him not to recognise the speed of the aircraft after a wrong number was entered into the autopilot. Although, a simple task, number entry still requires focus and attention. Distractors could move attention away from this task leading to such errors. The *DOV* has been seen to be effective in what might be considered ideal laboratory conditions, it may be that there is no benefit from it in environments when there are competing demands on attention.

Although there is a considerable body of research to understand the dynamics of attention and how attention can be captured by distractors, arguments have been focused on the concepts of early or late selection. One such theory mentions that whether early or late selection occurs depends on how much perceptual load someone experiences during a task (Lavie, 2005) (see Section 2.2.4 in Chapter Two). In the context of studying the *DOV*, it could be the case that presenting the numbers to be transcribed in an obscured i.e. hard to read format could itself be a source of perceptual load and therefore lead to less susceptibility to distraction. Nevertheless, it seems that this is not the case. Lavie and De Fockert (2003) made their stimuli harder to read by reducing contrast between the text and the black background (similar to Study Five) and also by reducing the size of the text. They found that, by examining the effect of this on the processing of visual distractors during a visual search task, there was no effect of the text format on susceptibility to distractors. Their study has similarities to this one in that a reduced contrast (faded number) is used to provoke the *DOV*. In contrast Study Nine uses a more interactive task, number entry, rather than visual search for a target as well as auditory distractors rather than visual distractors. Nonetheless, the results from Lavie and De Fockert (2003) suggest that the stimuli presented here is not simply influencing perceptual load and therefore influencing awareness of distractors in that approach.

While some researchers suggest that unlike in visual search tasks where perceptual load can result in filtering out the distractors in an auditory search task, the auditory distractors are still perceived regardless of the audio perceptual load (Tellinghuisen & Nowak, 2003; Murphy et al., 2013). There is also some evidence in the literature that increased engagement with a task can reduce the effect of auditory distractors (Halin, Marsh, Haga, Holmgren, & Sörqvist, 2014a; Halin, Marsh, Hellman, Hellström, & Sörqvist, 2014b). They use the *DOV* to examine how background speech could influence people's abilities to engage in proofreading tasks (Halin et al., 2014a) and recall tasks (Halin et al., 2014b). In the easy-to-read conditions, participants perform less well with the auditory distraction. Whereas, in the hard-to-read condition people perform equally well. This is maybe because speech engages complex processes and so the *DOV* effect interferes with some of those processes to reduce the effectiveness of speech as a distraction. Also, though this work shows promise for the robustness of the *DOV*, it should be noted that they contrasted the situation of auditory distraction with silence as opposed relevant versus irrelevant distractors.

Nevertheless, similar effects are seen in more "pure" tasks (Soerqvist, 2014) inattentional deafness (Macdonald & Lavie, 2011): people remain unaware of changes to the auditory environment depending on the visual perceptual load they are experiencing. The stimuli used in these experiments are much simpler and comparable to the visual search tasks discussed above and moreover compare irrelevant to relevant distractors. The role of distractors though, as a changing soundscape, is very different from the effect of general background speech as a distraction. A pure task as mentioned by Soerqvist (2014) is a task used in a laboratory study to study basic cognitive functions such as pressing a response key only if a stimuli is active.

Therefore, auditory distractions have an analogue to the noisy, distracting environments that motivate this work in that area. The influence of auditory distractions in relation to the *DOV* as applied in the number entry task is unknown. Although, Study Five might hint that a certain degree of noise might inhibit the effect the results of Study Five were inconclusive as if a single item or a combination of items (manipulation etc.) was responsible for the failure to stimulate the *DOV*. The *DOV* may enable better focusing on the number transcription task and hence reduce errors or it may be that, for this task, it is unable to block the effect of distractors. The existing psychology literature surrounding the *DOV* suggests the potential of the effect to overcome auditory distractions but the specific effect in the context of number entry is unknown.

5.2.1 Aims & Hypothesis

The aim of this study was to reproduce the *DOV* in a number transcription task in the presence of auditory distractors. The distraction in this study consists of an audio track, which was played to participants as they took part in the number-entry task. This study introduces two distinct audio tracks (numerals and letters) as distractions in two visibility conditions (*Clear & Obscured*) to investigate the robustness of the *DOV* in terms of memory encoding and retrieval. An audio track of numerals is intended to present a greater distraction in the number entry task than letters as the entry of numbers into a device requires the participants to encode and then enter numerals. Hearing letters ought not to interfere with this, following Tellinghuisen and Nowak (2003).

The expectation is therefore that numeral distractors will provoke more errors than letter distractors. However, if the DOV is robust, it should still be seen in both the numeral and letter conditions, which means that errors should be reduced under both conditions. This study can be seen as the first step towards investigating the efficacy of the DOV in real-world settings. The hypotheses for this study were therefore:

H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.

H_1 : The DOV effect can be reproduced in the number-transcription task. Hence there would be a significant difference in the number of errors between the *Obscured* and *Clear* conditions.

H_2 : Numeral distractors lead to more errors the letter distractors.

5.2.2 Participants

Forty participants (17 women) were randomly recruited by personal invitation in the Morrell library of the University of York. There were 10 participants in each group. The sample size was considered adequate as a previously studies showed a highly significant effect in a similar experimental set-up (Study Six). All of the participants were students at the University of York in different departments, except one sales administrator. Participants were between 19 and 29 years old (mean 24.25; SD 5.12). Five participants had previous experience in the healthcare sector. Among those were two nursing students. Four participants stated that they never used a tablet before, one in the *Clear-Letters* condition, one in the *Obscured-Numerals* condition and two in the *Obscured-Letters* condition. None of the participants stated that they never used a touch-based device. All of the people stated that they had normal vision and normal hearing. One participant in each condition stated that they did not enjoy the task and only one in the *Obscured-Letters* condition stated that the task was difficult. All participants were randomly assigned in to the different groups.

5.2.3 Number-Transcription Task and Interface

The tasks for this study were framed as a number-entry game on the iPad to enter as many numbers as quickly and as accurately as possible in a serial interface. The tool used for this study was identical to that in Study Eight. The number-entry task was again designed as a game with a score where participants could evaluate themselves on how well they perform while entering numbers. The score would increase by 100 for each correct number a person entered and decrease by 100 for a wrong number or for accidentally pressing the enter key identical to Study Eight. Participants had to enter randomly generated numbers in a range of 0-999.99 (5% without decimal places). The font-type used to represent the numbers was System Bold as in all the previous studies. The number-entry interface did not provide any cue as to a potential error (e.g. double taps on decimals) other than the visible transcribed number in the target display. However, the log file provided detailed insight into timestamps and all button presses. The application always used a white background. Participants had a time limit of five minutes (see figure 5.1).

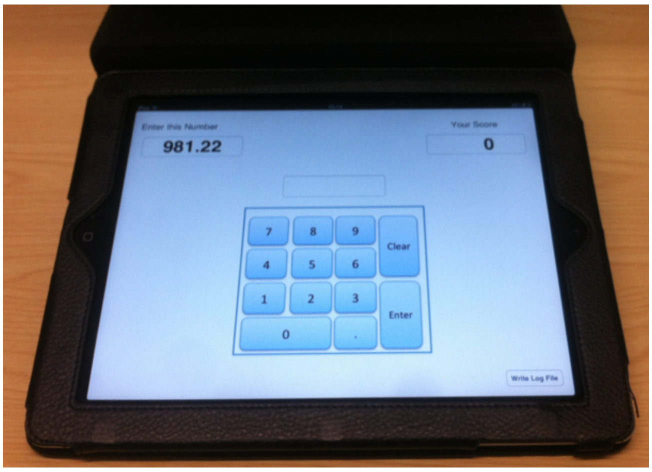


Figure 5.1: Study 9: Experimental Setup - iPad Interface

5.2.4 Design

A between-subjects design was used for this experiment. The study, had the following four conditions (as a result of two independent variables): the grade of visibility in which the to be transcribed information was presented in either black or grey (same colour code as Study Eight), and the type of audio track participants had to listen to while transcribing numbers (Letters or Numerals):

Clear-Letters (CL): In this group participants saw the numbers to be transcribed in a clear black font while listening to an audio track speaking single letters from the alphabet.

Clear-Numbers (CN): In this condition people saw the to be transcribed information in a clear black font while listening to an audio track speaking single numerals.

Obscured-Letters (OL): Participants in this condition saw the to be transcribed numbers in a grey colour font while listening to the single letters audio track.

And in the ***Obscured-Numbers (ON)*** condition, participants listened to the numeral audio track while transcribing information presented in a grey coloured font.

The dependent variables were again the total numbers entered, the total number of errors made, including the rate of errors and also the number of corrections made in each condition.

Provided by the results of Study Eight, it should be noted that the manipulation in Study Nine was only present at the source of the presented numbers not at the target display!

5.2.5 Materials

Participants were given paper based instructions and an informed consent form (see Appendix C.1.1) to read and to sign before proceeding to the training and the main task. A Google demographic questionnaire (see Appendix C.1.2) was given to participants after the study. The research instrument used for this study was an iPad2 running iOS 6.3.5. Additionally, this experiment required participants to listen to one of two audio tracks: one with random numerals and another one with random letters.

The audio track consisted of a string of numerals or letters. The audio track lasted for as long as the experiment ran (five minutes) and with each letter or numeral being read out with a one-second gap. Both tracks were set-up as a MS Word document and then read by the inbuilt voice of the used Macbook Pro. For this purpose Apple's Samantha voice package was downloaded and installed prior to the study as this voice showed the closest resemblance to a human voice. The numerals read to participants were random single numerals from 0 to 9. The letters were random single letters from the whole alphabet.

Unlike the experiments tailored to the study of distraction in either visual or auditory search, it was not possible to tightly synchronise the timing of the auditory distractors to the number-entry task. This was because of the difference between a visual search task and a number transcription task. In a visual search task, the participant simply has to respond to having found the target. In a number transcription task, the participant needs to make a series of key presses and that may depend on the participants typing speed as well as any of the intended manipulations. The headset used was a Somic Stereo ST-908 and a smart-phone was used to manage the time.

5.2.6 Procedure

The experiment was conducted in a controlled study room in the library. All participants were run in individual sessions, which were managed by personal invitation. Each participant was first welcomed and then given a copy of the informed consent form and the instructions to read and to sign. This was then followed by participants entering six training numbers while listening to the specific audio track (either letters or numerals).

Participants were then asked to enter the given numbers in one of the four conditions within five minutes as quickly as possible. The experimenter stopped the participant after the five minutes were up. After finishing the main number-entry task, participants were asked to complete the demographic questions followed by the debriefing of the study.

5.2.7 Pilot

A small pilot with two students revealed minor procedural errors. The data retrieval process from the iPad was sensitive to the applications memory. It was necessary to check if the data was written to a log file (otherwise the data in the memory would have been overwritten) and double check the data after it was copied off the device. Alternatively, it was also possible to copy the complete data-set at once after the complete study. As usual people who participated in the pilot were excluded from the main study.

5.2.8 Results

Again the data which is analysed in this study is non-parametric data (error-data). As this analysis is looking for two-way effects, a *Analysis of Variance (ANOVA)* is the suitable test, as there is no well established test for non-parametric data. However, main effects where they are seen in the ANOVA are confirmed with a non-parametric *Mann-Whitney* test.

3680 numbers were entered in total during the study, which resulted in 66 total errors and an error-presentation rate of 1.8%. 957 numbers were entered in condition *Clear-Letters* (19 errors), 844 in condition *Clear-Numbers* (23 errors), 914 in condition *Obscured-Letters* (14 errors) and 925 in condition *Obscured-Numbers* (10 errors), respectively. Table 5.1 illustrates the relevant descriptives in relation to the total numbers entered in each condition.

		Visibility		
		Clear	Obscured	True Mean
Audio	Letters	95.70 (21.81)	91.40 (16.49)	93.55 (18.95)
	Numbers	88.40 (17.90)	92.50 (13.20)	90.45 (15.45)
	True Mean	92.05 (19.77)	91.95 (14.54)	

Table 5.1: Study 9: Descriptives: Mean (SD) for the total numbers entered

An alpha level of 0.05 was used for all statistical tests. The data was analysed using the ANOVA, with the grade of the visibility and the type of audio distraction as unrelated samples variables. In regards to the numbers entered in total, there was no main effect of visibility, $F(1,36) = 0.0; p = 0.99; \eta_p^2 < 0.01)$, with those in the *Obscured* conditions transcribing slightly more numbers than in the *Clear* conditions but not significantly. There was also no main effect for audio $F(1,36) = 0.309; p = 0.58; \eta_p^2 < 0.01)$, with a few more numbers entered in the letters conditions but again not significantly more than in the numbers conditions. There was also no interaction effect (Visibility X Audio) $F(1,36) = 0.568; p = 0.46; \eta_p^2 = 0.016$ (see figure 5.2).

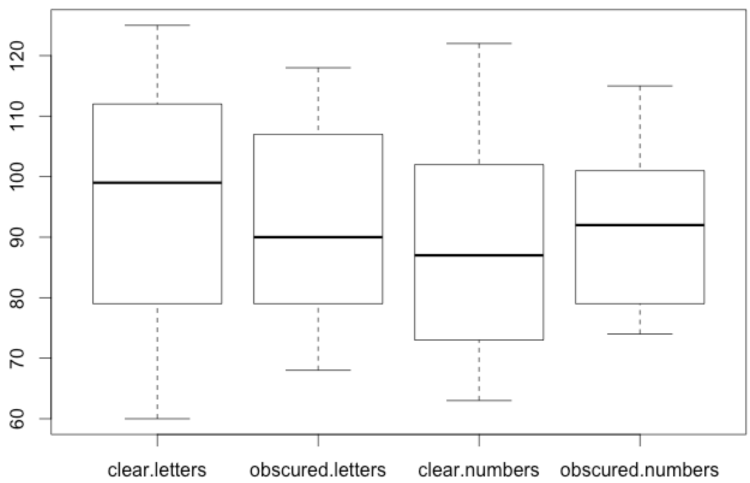


Figure 5.2: Study 9: Boxplot of total numbers entered by each participant in each condition.

In regards to the total errors made, there was a marginal main effect of visibility, $F(1,36) = 4.084; p = 0.051; \eta_p^2 = 0.10)$, with those in the *Obscured* conditions making marginally fewer errors than participants in the *Clear* conditions. This main effect was confirmed by the *Mann-Whitney* test ($p = 0.04; W = 272.5$). There was however no main effect for audio $F(1,36) = 0.0; p = 1.0; \eta_p^2 < 0.001)$, with both audio tracks causing participants to commit about the same amount of errors. There was no interaction effect (Visibility X Audio) $F(1,36) = 0.807; p = 0.38; \eta_p^2 = 0.022)$ (see figure 5.3). The following table 5.2 illustrates the descriptives regarding the total errors made by condition:

		Visibility		
		Clear	Obscured	True Mean
Audio	Letters	1.90 (1.10)	1.40 (1.43)	1.65 (1.27)
	Numbers	2.30 (1.70)	1.00 (1.34)	1.65 (1.63)
	True Mean	2.10 (1.41)	1.20 (1.36)	

Table 5.2: Study 9: Descriptives: Mean (SD) for the total errors made

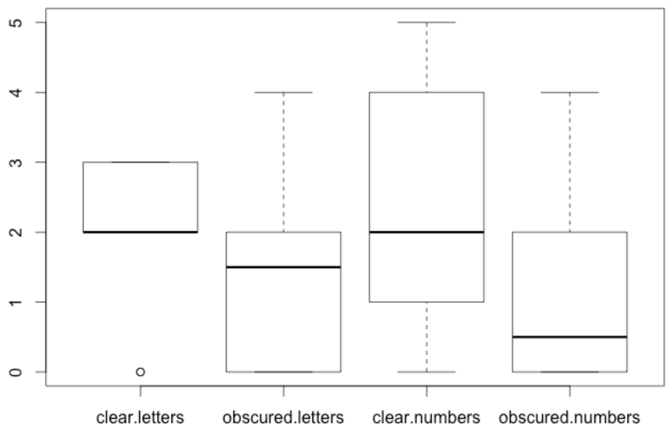


Figure 5.3: Study 9: Boxplot of total errors made by each participant in each condition.

Thirteen participants made no errors during the entire study. Two participants made no errors in the *Clear-Letters* condition, two participants made no errors in the *Clear-Numbers* condition, four participants made no errors in both the *Obscured-Letters* condition whereas five participants made no errors in the *Obscured-Numbers* condition (68% of the participants made at least one error). The highest error-count was recorded in the *Clear-Numbers* condition.

In regards to the error-rate, there was a marginal main effect of visibility, $F(1, 36) = 2.31$; $p = 0.06; \eta_p^2 = 0.09$). There was no main effect for audio $F(1,36) = 0.005; p = 0.95; \eta_p^2 < 0.001$) and no interaction effect $F(1,36) = 1.18; p = 0.285; \eta_p^2 = 0.032$ (see figure 5.4). A *Mann-Whitney* test confirmed the marginal effect of visibility on error rate ($W = 264.5; p = 0.078$).

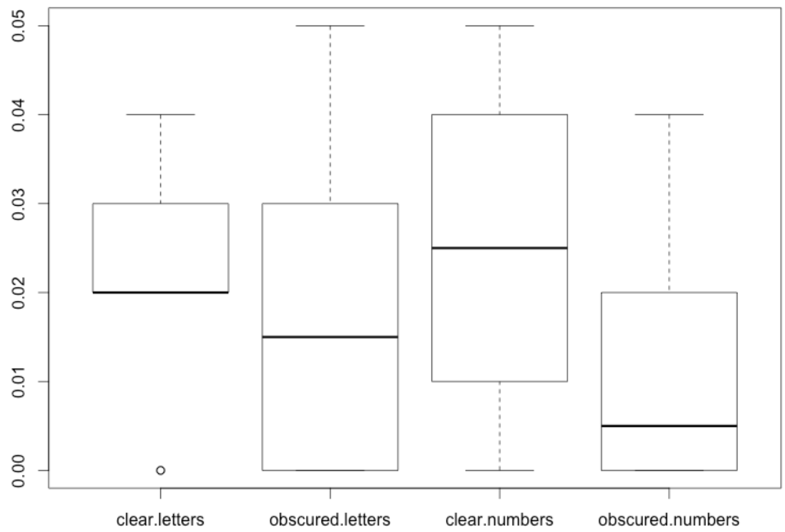


Figure 5.4: Study 9: Boxplot of the error rate by each participant in each condition

Participants made in total 268 corrections during the study. 75 corrections were made in condition *Clear-Letters*, 42 in condition *Clear-Numbers*, 90 in condition *Obscured-Letters* and 61 in condition *Obscured-Numbers*, respectively. The highest amount of corrections made by a participant was 23 (resulting in 0 errors) in the *Obscured-Letters* condition.

In regards to the total corrections made, there was no main effect of visibility, $F(1, 36) = 1.565; p = 0.22; \eta_p^2 = 0.042$), with those in the *Obscured* conditions making more corrections than participants in the *Clear* conditions but not significantly more. Table 5.3 shows the descriptives in relation to the total corrections. There was however a main effect for audio $F(1, 36) = 5.202; p = 0.03; \eta_p^2 = 0.13$), with participants making significantly more corrections in the letters conditions than the numbers conditions, which the *Mann-Whitney* test confirmed ($W = 275; p = 0.04;$). There was also no interaction effect (Visibility X Audio) $F(1, 36) = 0.022; p = 0.88; \eta_p^2 < 0.001$) (see figure 5.5).

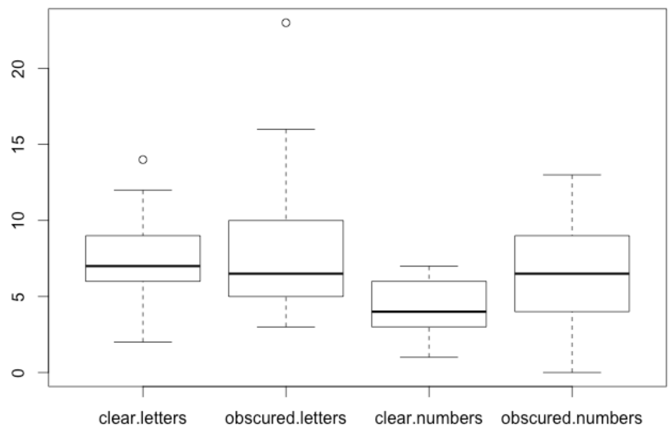


Figure 5.5: Study 19: Boxplot of the total corrections made by each participant in each condition

		Visibility		
		Clear	Obscured	True Mean
Audio	Letters	7.50 (3.72)	9.00 (6.20)	8.25 (5.04)
	Numbers	4.20 (2.04)	6.10 (4.18)	5.15 (3.35)
	True Mean	5.85 (3.38)	7.55 (5.36)	

Table 5.3: Study 9: Descriptives: Mean (SD) for the total corrections made

The following table 5.4 presents a summary of the errors participants made classified according to Wiseman et al. (2011)) and structured in the four conditions.

Type	Occurrence in condition				Example (to-be-entered -> entered)
	ClearLetters	ClearNumerals	ObscuredLetters	ObscuredNumerals	
Out-by-factor-of-10	4	4	2	3	658.71 -> 65.87
Transposition Error	2	3	6	3	289.32 -> 298.32
Wrong Digit	10	10	5	4	43.02 -> 43.03
Wrong Number	1	1	1	0	186.47 -> 0.81
Extra Digit	1	0	0	0	548.81 -> 5483.81
Missing Digit	0	5	0	0	81.04 -> 8.04
Early Termination	1	0	0	0	79.91 -> 79.9

Table 5.4: Study 9: Table of number-transcription errors

5.2.9 Discussion

This study aimed to replicate the *DOV* in a number-transcription task and additionally investigate its robustness in the presence of audio distractors. The results show that the total number of errors made was reduced by the *DOV* and the error-rate was also marginally reduced as well.

Moreover, this is not a speed-accuracy trade-off with participants slowing down in order to be more accurate which aligns with identical effects in Study Four, Six and Eight. Although these results reflect the ones from previous studies they are not so unequivocal. Of course, the main difference is that the current task was done in the presence of an audio distraction. Though the intention was that the letters would be less distracting than the numerals, this does not seem to have come through in the measures used. There are no main effects of audio distractors on either the total errors made or the error rate. The overall error-presentation rate was 1.8% and is again equally low compared to previous studies.

A possible explanation could be that some way the numeric distractors also contributed to the *DOV* by increasing focus and concentration on the task. This would mean that the numeric distractors were not sufficiently distracting. There are indications in the literature that people process distractors depending on the type of processing demands. Specifically, Lavie and De Fockert (2003) found that higher perceptual load stimuli in a “relevant” task reduced the processing of “irrelevant” distractors.

In general this could be transferred to this study. It appears that numeric distractors are somewhat task relevant to numerical entry whereas letters are not which is reflected in the hypothesis. Therefore the implication would be that the task is not sufficiently demanding so that all distractors are being processed equally. However, Lavie and De Fockert (2003) were limited to the visual domain only.

In the mixed modality domain, Tellinghuisen and Nowak (2003) found no reduction in the effect of distractors depending on visual load but even so the effect of distractors was dependent on the relevance of the distractors to the task in hand. This is therefore an interesting feature contrasting with the existing literature that encourages further exploration but cannot be examined in any more detail from the data here as it was not the primary focus of this study.

If it is therefore the case that letters and numerals are equally distracting for the number-entry task, then the spoken numerals are not more relevant than the spoken letters. This may be because numbers can be multiply represented by people as strings of digits, spoken words or spatially (Dehaene, 1997) and the spoken representation of numerals is not used in the number-entry task.

Alternatively, it might be that all auditory distractors are equally intrusive for this task but this might perhaps have led to a higher error-rate overall than previous studies, which was not the case. In fact, given the generally lower error-rate, it may be that there was some increased engagement as a consequence of the distracting audio and the DOV could only marginally improve upon this.

The only significant difference due to distractors was in corrections with participants making significantly more corrections (both totally and as a rate of corrections) when listening to the letters distractors. If anything, this suggests that letters were more distracting than numerals but in such a way that participants were able to catch the errors for themselves.

This might be a result of interference between the motor memory to type a letter with the intention to type a particular digit (Salthouse & Saults, 1987) or perhaps an indication that letters and numerals are processed differently (Dehaene, 1997). However, this is essentially a surprising result that would need substantially more investigation before a more insightful explanation can be provided and therefore remains speculation at this point.

Table 5.4 showed a summary of the different types of errors that participants made structured according to previous work in the field (Wiseman et al., 2011). Transposition errors seem to happen across the conditions indicating possible motor skill dependence (Salthouse & Saults, 1987). Whereas missing digit errors only occurred in one of the clear conditions indicating what could be seen as possible encoding errors. Similarly, wrong digit errors also seemed to occur more in both clear conditions.

However, due to the few number of occurrences of errors these differences can also only be seen as indicative but not conclusive. These could however be used as working hypotheses to consider in future studies.

The results presented here support a potential effect of the DOV in improving number transcription accuracy, even in the presence of auditory distractors. The effect size is modest and only marginally significant but bearing in mind the safety-critical context, which motivated this thesis, where number entry is a widespread and accuracy dependent task, even small effects of this kind can be of importance.

Notably, Study Nine only used the manipulation to provoke the DOV at the source not the target. Further strengthening the previous argument that the DOV is in fact source related.

A previously mentioned, the DOV is a novel interactive way to reduce errors and therefore much is yet to be explored of how the effect behaves. This thesis is the first attempt to map this unchartered territory but there is still some way to go from definitively establishing the robustness of the effect. In particular, that the numeral distractors were not more distracting than the letter distractors is surprising and it may even be the case that any auditory distraction for this task can be effectively screened out. It may indeed be the case that when it comes to the number transcription task, it brings its own focusing effect that is therefore naturally resistant to distraction. It would be worth comparing this with other interactive tasks to see if it is possible to perceive the effectiveness of distractors in this context.

Study Nine successfully replicated the *DOV* effect and further investigated its robustness once levels of distractions are present. However, the effect was only beginning to manifest itself as results were marginal. Surprising facts of the used distractors were revealed yet require further investigation as current explanations have to be viewed with care. The absence of control conditions which only examine the *DOV* effect on its own in contrast to the auditory distractors is considered a limitation of Study Nine.

The following Study Ten will further examine the *DOV* effect and was aimed to provide further knowledge as to how the effect behaves in an distracting environment. Different distractors are chosen based on previous work in the literature with the intention to have more effective distractors present. The following experimental design also provides control groups to be able to examine the effects more clearly.

5.3 Experiment 10

The previous Study Nine has shown that the *DOV* effect may still be seen in a distracting environment. This finding presents a first indication on how the effect would behave in real-world busy environments were distractions are constant such as hospital wards. However, there were issues with the chosen auditory distractors. The numeric distractors were not, as expected, more distracting than the letters and even the letters distractors did not have a significant effect on the errors or on the error rate but only on the corrections raising questions about the effectiveness of the chosen distractors. It was therefore decided to run a follow-up study with alternative, what was believed more effective, auditory distractors namely *white-noise* and *narrative speech*.

The auditory distractors in Study Nine seemed sensible by comparison with the literature on perceptual load and distraction. However, the stimuli in Study Nine were not very realistic and the task itself is interactive which could explain why the stimuli were not powerful enough. Because of this reason, Study Ten moved towards more effective and known to be more reliable effects namely *white-noise* and *narrative-speech* which can be compared to a general hub of noise. Study Ten moved towards possibly more effective stimuli for a wider range of tasks.

The literature presents evidence that non-task related auditory stimuli i.e. irrelevant sound has the capacity to break selective attention and even impair cognitive processes (Banbury, Macken, Tremblay, & Jones, 2001). Studies Nine and Ten can be seen as falling under what Banbury et al. (2001) described as the irrelevant sound paradigm. This paradigm describes studies which have a short-term memory task where participants have to recall visually presented items for example numbers or words but in an environment where auditory distractions such as *white-noise* or *narrative speech* are introduced. The studies in Chapter Five have identical characteristics but lack the active recall task. Nevertheless, according to the surrounding literature and the established findings in this thesis on how the *DOV* works it is assumed that participants use the same memory encoding processes only strengthened by the *DOV*.

White-noise as irrelevant background noise is one of the common distractors found in studies related to attention and task performance (Salame & Baddeley, 1989). It consist of a continuous sound of equal levels in every frequency. However, the effectiveness of *white-noise* as a distractor on cognitive processes are not clear (Jones & Broadbent, 1991). While some researchers present significant interference effects of *white-noise* (Salame & Baddeley, 1989; Jones, Madden, & Miles, 1992) others report minor effects or no effects at all (Beaman, 2005). For example, Dalton and Behm (2007) report that *white-noise* can result in both distractions and increased stress. Smith (1988) presented results that *white-noise* reduced performance during a number detection tasks and further interfered with an estimation task. Whereas Harrison and Kelly (1989) report that *white-noise* improved performance compared to *no audio* conditions. Additionally, Soederlund, Sikstroem, Loftesnes, and Sonuga-Barke (2010) found that *white-noise* improved task performance in inattentive children, but significantly impaired those who were normally attentive. Despite the variances of the effect of *white-noise* as a distractor it was chosen for this study as one level of distraction.

Narrative speech in contrast to *white noise* has a much more consistent record on its effect on task performances (Salame & Baddeley, 1982; Szalma & Hancock, 2011). For example, if *narrative speech* is included as a distractor during irrelevant sound studies i.e. studies with a recall task, accuracy can decline to 50% or even to 30% (Ellermeier & Zimmer, 1997). According to Banbury et al. (2001) this result is reliable in multiple experiments. Jones, Marsh, and Hughes (2012) showed that memory retrieval was significantly impaired once semantically meaningful *narrative speech* was present in contrast to *no audio*. This was however not the case once semantically meaningless speech such as reversed *narrative speech* was present. Such a semantic effect was also detected in multiple other experiments (Jones, Miles, & Page, 1990; Soerqvist, Noestl, & Halin, 2012). This could be a further explanation why the letters distractor did not have the expected effect on errors in Study Nine as it lacked semantic meaning. Macken, Phelps, and Jones (2009) discussed *narrative speech* in contrast to working memory performance and state that an interference in such performance by *narrative speech* is due to the competition of task relevant information and task irrelevant information during recall (Szalma & Hancock, 2011). Perhaps this is in line with the perceptual load theory as previous studies have found that tasks with low perceptual load leave spare capacity which is then used for the processing of task irrelevant information, whereas tasks with high perceptual load consume all the capacity which results in no available space left for processing task irrelevant information (Lavie, 1995). Mack and Rock (1998) found that reduced perceptual processing of task irrelevant information in high perceptual load leads to “inattentional blindness” i.e. failure to recognise distractors partially or completely during tasks ((Lavie, Lin, Zokaei, & Thoma, 2009; Macdonald & Lavie, 2011). In general, speech can be distracting and to some degree even task irrelevant speech is processed as seen in the “Cocktail Party Effect” (Norman, 1968). This effect shows the ability of humans to tune into conversations and to filter out other conversations in a noisy environment by grasping specific auditory stimuli for example the name of a person. For the above mentioned reasons and its well known effects in the literature *narrative-speech* was therefore chosen as a second level of distraction for this study.

5.3.1 Aims & Hypotheses

The aim of this study was to reproduce the *DOV* effect in a number transcription task identical to Study Nine. This study also aims to investigate how robust the detected *DOV* effect is in two different levels of distraction. In theory, the *DOV* effect is built upon deeper cognitive processes which result in better memory encoding and retrieval of the to-be-transcribed disfluent information. Study Nine indicated that even in the presence of distractions the DOV may still be seen.

It is expected that the two levels of auditory effects have significant interference effects on the number of errors and the error-rate in this experiment. Based on previous findings in the literature it is expected that narrative speech has a higher distracting impact on the task performance. The hypotheses for this study are therefore:

H_0 : There is no difference in any of the evaluated conditions. The manipulations do not have any effect.

H_1 : The DOV effect can be reproduced in the number-transcription task. Hence there is a significant difference in the error-rate between the *Obscured* and *Clear* conditions.

H_2 : The audio distractors influence the errors and error-rate.

5.3.2 Design

A between-subjects design was used for this experiment. This study had the following six conditions resulting out of two independent variables the grade of visibility in which the to-be-transcribed information was presented (either black or grey) and the type of audio track participants had to listen to while transcribing numbers (*no audio, white-noise* or *narrative-speech*). The two groups with no audio acted as control groups in this experiment. Identical to Study Nine this study also only used the manipulation at the source.

Clear-NoAudio: In this group participants saw the numbers to be transcribed in a clear black font but no audio track was played. Participants were instead instructed to wear the headphones to shield from outside noise.

Clear-WhiteNoise: In this condition people saw the to be transcribed information in a clear black font while listening to *white-noise* (more specific details on the audio file used can be seen in Section 5.3.4).

Clear-NarrativeSpeech: In this condition people saw the to be transcribed information in a clear black font while listening to *narrative-speech* (again details can be seen in Section 5.3.4).

Obscured-NoAudio: Participants in this condition saw the to be transcribed numbers in a grey faded colour font but no audio track was played. Participants were instead instructed to wear the headphones to shield from outside noise.

Obscured-WhiteNoise: Participants in this condition saw the to be transcribed numbers in a grey faded colour font while listening to *white-noise*.

Obscured-NarrativeSpech: In this condition people saw the to be transcribed information in a grey faded colour font while listening to *narrative-speech*.

The dependent variables in this study were identical to the previous one: the total numbers entered; the total errors made; the total error rate; the total corrections made

5.3.3 Participants

Ninety participants (65 women), fifteen in each condition, were randomly recruited by personal invitation in the library of the University of York. All of them were students in different departments except for one accountant. Participants were between 19 and 25 years old (mean 22.89; *SD* 3.4). Six people stated that they were familiar with infusion pumps and twenty-eight had previous experience in the healthcare sector. All of the participants had previously used a touch-based device and mentioned that they had normal vision. Only two participants mentioned that they did not consider themselves as having normal hearing (these two participants did not perform differently in the experiment). Two participants in the *ObscuredNoAudio* condition stated that they did not enjoy the task and none stated that the task was difficult. All participants were randomly assigned to a group.

5.3.4 Number-Transcription Task, Materials and the Interface

The tasks set for this experiment were identical to the previous Study Nine. Participants had to enter as many numbers as possible using a *Serial* interface. The given time frame for the task was six minutes. This time-scale was chosen in relation to Study Nine to increase the amount of error error-data available for analysis.

Consistent to the previous studies participants were given the paper-based instructions and an informed consent form to read and to sign before proceeding to the training and the main task. The two different forms can be seen in the Appendix (C.2.1 and C.2.2). The difference in the forms was only the in the headset instructions. One group was told the purpose of the headset was to shield against outside noise and the other was told that they had to perform the transcription task while listening to an audio track. An on-line Google demographic questionnaire was given to people after the study was completed. The questionnaire can be seen in Appendix C.2.3.

The research instrument used for this study was an iPad Mini 2 running iOS 8.4. Additionally, this experiment required participants to listen to two audio tracks: *white-noise* and *narrative-speech*. The *white-noise* file was a ten hours constant track taken from Youtube (`https://www.youtube.com/watch?v=wzjWIxXBs_s`). Additionally, by searching for a piece of narrative that would be freely available and of general interest of somebody speaking clearly as well as articulately the commencement speech of Admiral William H. McRaven was chosen as it had some attributes of value. The speech of Admiral McRaven to University of Texas graduates in 2014 is engaging and slightly humorous as it presents his career experience as guidance for the graduates. The file was always played from 4:49 for the duration of the task (six minutes) and can also be accessed on Youtube (`https://www.youtube.com/watch?v=pxBQLFLei70`). Both tracks were set-up via a Macbook Pro and a headset (Somic Stereo ST-908). A smartphone was used as a stopwatch.

5.3.5 Procedure

Identical to previous experiments, each participant was first welcomed and then given a copy of the informed consent form and the instructions to read and to sign. This was then followed by participants entering three training numbers while listening to the specific audio track (depending in which condition either *white-noise, narrative-speech or no audio* at all). Participants were then asked to transcribe the given numbers in one of the six conditions within the given time-frame. The experimenter stopped the participant after the six minutes were up. After finishing the main number-entry task, participants were asked to complete the demographic questions followed by the debriefing of the study.

5.3.6 Pilot

The study was piloted by two graduate students at the University of York. No flaws or technical problems were detected. Both pilot participants did not take part in the main study.

5.3.7 Results

As in previous studies, the data which is analysed in this study is also non-parametric data (error-data). As this analysis is looking for two-way effects, a *(2 x 2) Analysis of Variance (ANOVA)* is again the suitable test, as there is no well established test for non-parametric data. If main effects are seen in the ANOVA they are confirmed with the appropriate non-parametric tests (*Mann-Whitney* & *Kruskal-Wallis*).

In total, 10287 numbers were entered during the study, which resulted in 185 total errors and a total error-presentation rate of 1.8%. Table 5.5 shows the total numbers entered and the corresponding errors per condition and table 5.6 illustrates the relevant descriptives in relation to the total numbers entered in each condition.

		Visibility	
		Clear	Obscured
Audio	NoAudio	1679 (36)	1760 (20)
	WhiteNoise	1676 (24)	1755 (20)
	NarrativeSpeech	1659 (46)	1558 (39)

Table 5.5: Study 10: Total numbers entered (total errors made) per condition

		Visibility		
		Clear	Obscured	True Mean
Audio	NoAudio	111.94 (18.32)	117.34 (13.68)	114.64 (16.12)
	WhiteNoise	111.74 (11.20)	117.00 (18.30)	114.37 (15.15)
	NarrativeSpeech	110.60 (18.48)	117.20 (17.84)	113.90 (18.16)
	True Mean	111.43 (15.99)	117.18 (16.35)	

Table 5.6: Study 10: Descriptives: Mean (SD) for the total numbers entered

Again an alpha level of .05 was used for all the following statistical tests. In regards to the numbers entered in total, there was no main effect of visibility, $F(1, 84) = 2.722; p = 0.10; \eta_p^2 = 0.031$), with those in the *Obscured* conditions transcribing slightly more numbers than in the *Clear* conditions but not significantly. There was also no main effect for audio $F(1, 84) = 0.015; p = 0.98; \eta_p^2 < 0.001$), with a few more numbers entered in the *NoAudio & WhiteNoise* conditions than in the *NarrativeSpeech* condition but again not significantly more. There was also no interaction effect (Visibility X Audio) $F(1, 84) = 0.015; p = 0.98; \eta_p^2 < 0.001$ (see figure 5.6).

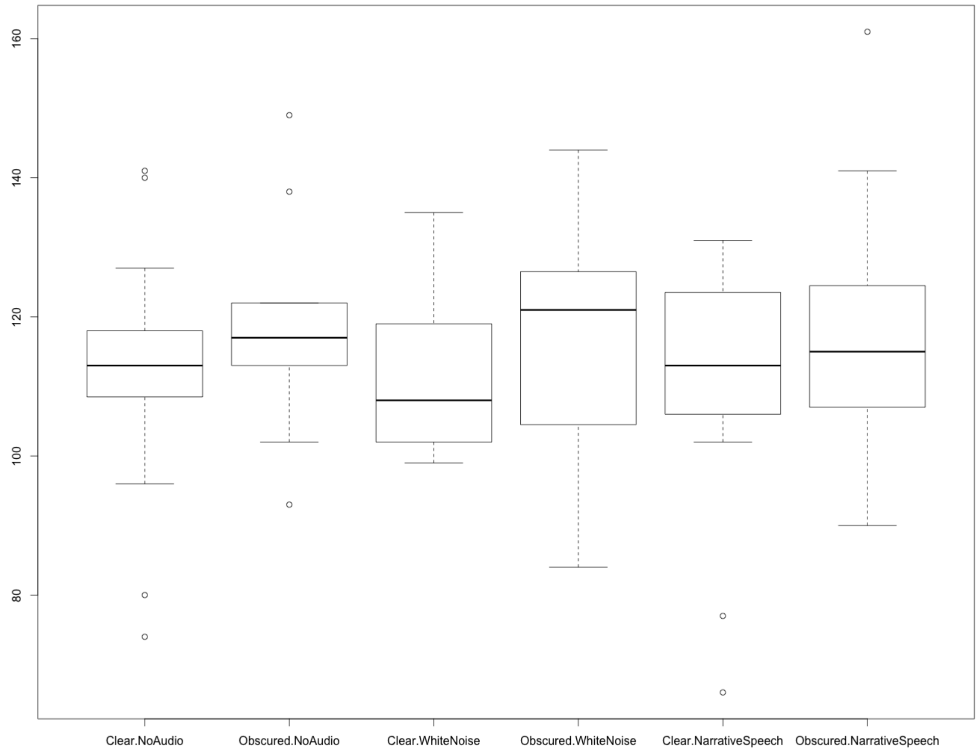


Figure 5.6: Study 10: Boxplot of total numbers entered by each participant in each condition.

In regards to the total errors made, there was a marginal main effect of visibility, $F(1,84) = 3.772; p = 0.055; \eta_p^2 = 0.04$), with those in the *Obscured* conditions making marginally fewer errors than participants in the *Clear* conditions. This main effect was not confirmed by the *Mann-Whitney* test ($W = 1187.5; p = 0.15$). There was however also a main effect for audio $F(1,84) = 6.897; p = 0.002; \eta_p^2 = 0.14$), with the *NarrativeSpeech* audio distractor causing participants to commit more errors than *WhiteNoise* or *NoAudio*. This main effect could be confirmed by the *Kruskal-Wallis* test ($H = 17.35; p < 0.001$). There was no interaction effect (Visibility X Audio) $F(1,84) = 0.605; p = 0.54; \eta_p^2 = 0.01$) (see figure 5.7). The following table 5.7 illustrates the descriptives regarding the total errors made by condition:

		Visibility		
		Clear	Obscured	True Mean
Audio	NoAudio	2.40 (1.84)	1.34 (1.05)	1.87 (1.57)
	WhiteNoise	1.60 (1.76)	1.30 (0.82)	1.46 (1.36)
	NarrativeSpeech	3.07 (1.75)	2.60 (1.24)	2.83 (1.51)
	True Mean	2.36 (1.85)	1.76 (1.19)	

Table 5.7: Study 10: Descriptives: Mean (SD) for the total errors made

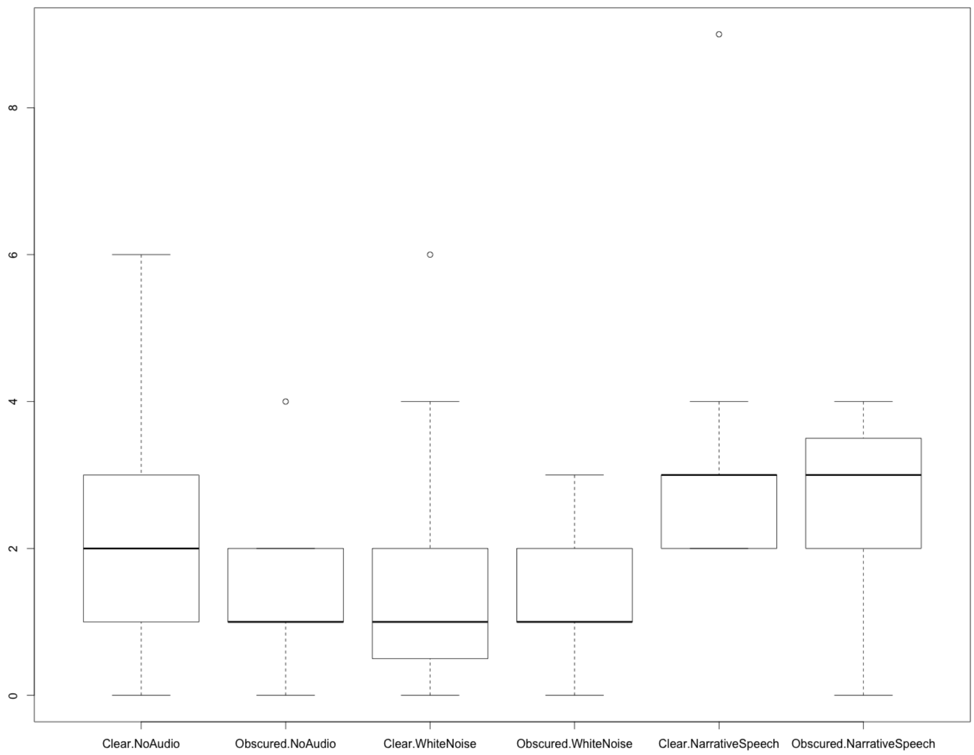


Figure 5.7: Study 10: Boxplot of total errors made by each participant in each condition.

Twelve participants (13%) made no errors during the entire study that means that 87% were susceptible to errors. The highest error-count (9) was recorded in the *Clear-NarrativeSpeech* condition.

In regards to the error-rate, there was a main effect of visibility, $F(1, 84) = 4.755; p = 0.03; \eta_p^2 = 0.05$). This main effect could be confirmed by the non-parametric *Mann-Whitney* test ($W = 1260.5; p = 0.04$). There was also a main effect for audio $F(1, 84) = 7.796; p < 0.001; \eta_p^2 = 0.16$) which could also be confirmed by the *Kruskal-Wallis* test ($H = 18.48; p < 0.001$). Again no interaction effect (Visibility X Audio) $F(1, 84) = 0.643; p = 0.53; \eta_p^2 = 0.02$ could be detected (see figure 5.8). Table 5.8 shows the descriptives regarding the total error-rate made by condition.

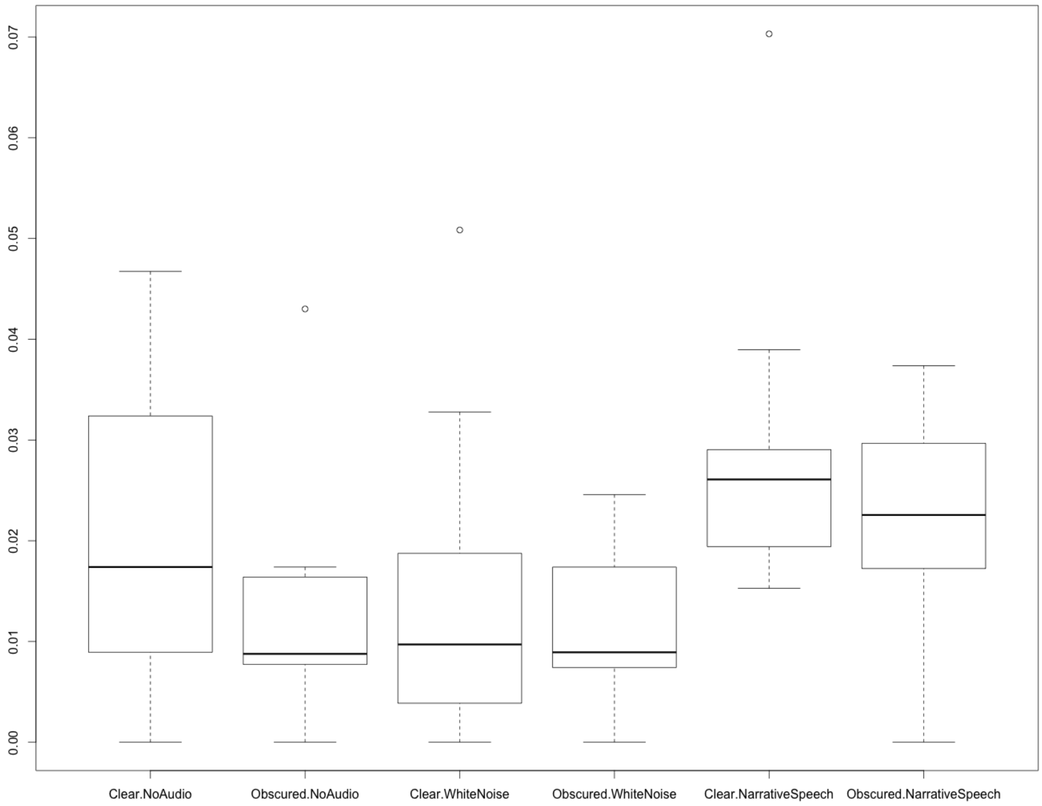


Figure 5.8: Study 10: Boxplot of the error rate by each participant in each condition

		Visibility		
		Clear	Obscured	True Mean
Audio	NoAudio	0.021 (0.016)	0.012 (0.011)	0.016 (0.014)
	WhiteNoise	0.014 (0.014)	0.012 (0.007)	0.013 (0.011)
	NarrativeSpeech	0.028 (0.013)	0.022 (0.011)	0.025 (0.012)
	True Mean	0.021 (0.015)	0.015 (0.011)	

Table 5.8: Study 10: Descriptives: Mean (SD) for the total error rate

Participants made in total 907 corrections during this experiment. The highest amount of corrections made by a participant was 25 (resulting in 1 error) in the *Obscured-NoAudio* condition. Table 5.9 shows the corrections per condition and table 5.10 shows the descriptives in relation to the total corrections made in each condition.

		Visibility	
		Clear	Obscured
Audio	NoAudio	195	146
	WhiteNoise	153	162
	NarrativeSpeech	157	94

Table 5.9: Study 10: Corrections made per condition

		Visibility		
		Clear	Obscured	True Mean
Audio	NoAudio	13.0 (6.12)	9.7 (6.98)	11.4 (6.66)
	WhiteNoise	10.2 (4.39)	10.8 (5.89)	10.5 (5.11)
	NarrativeSpeech	10.5 (6.89)	6.3 (4.32)	8.4 (6.04)
	True Mean	11.2 (5.89)	8.9 (6.03)	

Table 5.10: Study 10: Descriptives: Mean (SD) for the total corrections made

In regards to the total corrections made, there was a marginal main effect of visibility, $F(1, 84) = 3.430; p = 0.06; \eta_p^2 = 0.04)$, with those in the *Obscured* conditions making fewer corrections (103) than participants in the *Clear* conditions but only marginally significantly fewer. This effect was confirmed by a *Mann-Whitney* test ($W = 1246.5; p = 0.05$). There was no main effect for audio $F(1, 84) = 2.081; p = 0.13; \eta_p^2 = 0.04$), with participants making fewer corrections in the *Obscured-NarrativeSpeech* conditions than the other conditions. There was also no interaction effect (Visibility X

Audio) $F(1, 84) = 1.414; p = 0.25; \eta_p^2 = 0.03$) (see figure 5.9).

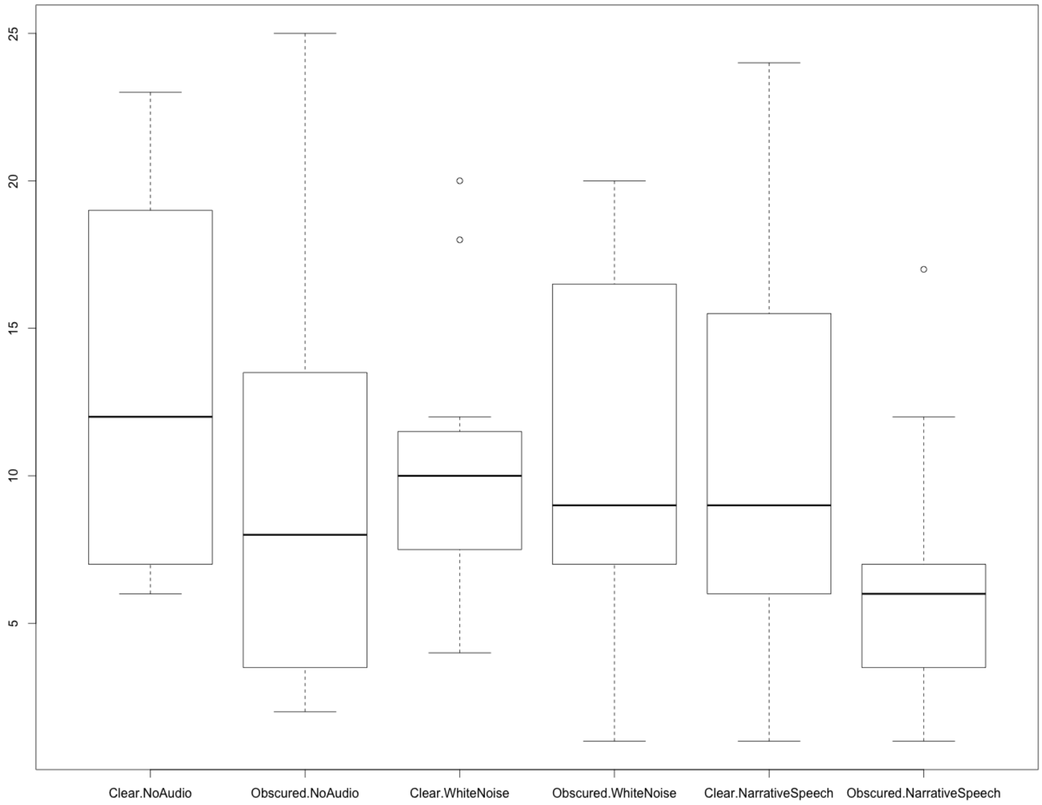


Figure 5.9: Study 10: Boxplot of the total corrections made by each participant in each condition

5.3.8 Discussion

This study was aimed to further investigate how the *DOV* effect would behave once auditory distractions are introduced into the environment. It had the alternative hypothesis (H_1) that the *DOV* effect could be reproduced in a number-transcription task as in an earlier studies. This would be seen in a significant difference in the error-rate between the *Clear* and the *Obscured* conditions. Moreover, it was expected that the distractors have a significant impact on the performance of the participants in the task i.e. the total number of errors and the error-rate (H_2). In detail, it was expected that the *narrative-speech* distractor would have a stronger distracting effect than *white-noise*. How the *DOV* effect would react to the two levels of distractions (*white-noise* and *narrative-speech*) was unknown. A further set of *no audio* groups acted as control condition.

Results showed that participants did not enter significantly different amounts of numbers in the different conditions. This is in line with previous studies (Four, Six, Eight and Nine) where the *DOV* was present indicating no speed-accuracy trade-off. By examining figure 5.6 there seems to be a downwards trend of numbers entered in the *Clear-WhiteNoise* condition which is however not reflected in the descriptives presented in table 5.6.

Overall, the study presents a low overall error presentation rate of 1.8% adding to previous studies in this field despite the presence of distractors. In terms of total errors made the *DOV* was only marginally present but the non-parametric *Mann-Whitney* test did not confirm an effect. However, the audio distractors worked in this study as there was a main effect for audio. Participants made significantly more errors in the *narrative-speech* condition than in the others as expected. However, the *white-noise* distractor did not seem to have an effect as by examining table 5.6 the means of both *white-noise* conditions are the lowest of the audio conditions in terms of errors. A result that can be seen in previous work (Harrison & Kelly, 1989). The fewest errors were seen in the *Obscured-NoAudio* and in the *Obscured-WhiteNoise* conditions further indicating an effect of the *DOV* coming through and perhaps an inability of white-noise to distract.

The *DOV* was clearly seen by taking the error-rate into account which could also be confirmed by the *Mann-Whitney* test. There was a clear significant difference between the *Clear* and *Obscured* conditions in terms of error-rate. The main effect for audio could also be seen and confirmed by the non-parametric *Kruskal-Wallis* test. People made more errors in the *narrative-speech* conditions. The unequal effect of the auditory distractors could again be based on how the distractors are processed or how much resources they need to be processed (Lavie & De Fockert, 2003). It could be that the task irrelevant *white-noise* simply was not demanding enough to be processed at the same level as *narrative-speech*. A much more prevailing explanation is simply the presence of the *DOV* effect and the possible relation of the effect to the perceptual load theory (Lavie, 2005). The deeper cognitive processes and therefore the better encoding of the visually perceived information could have resulted on more working memory capacity being available. Therefore more capacity was available to process irrelevant information faster or perhaps the irrelevant information was screened out early (early selection) (Lavie, 2005). Another explanation would be that *white-noise* distractor was equally distracting in the first place but the *DOV* reduced its effect to the same level as no distractor was present. However, if that was the case than the *Clear-WhiteNoise* condition should show an increase in errors which it did not.

Perhaps *white-noise* induces the *DOV*. It could be that *white-noise* is just making it harder to concentrate for participants and therefore invokes *System 2* processes resulting in the decrease of errors. Or perhaps *white-noise* has an effect of helping to focus which has nothing to do with the *DOV*. Therefore participants made fewer errors which subsequently left fewer opportunities for the *DOV* to come into play.

By focusing on the distractors it could be seen that in each audio condition there is a reduction in errors where the to-be-transcribed numbers were presented in an obscured form. All the error-rates are low at a level of the control group *Obscured-NoAudio*. Except the one in *narrative-speech*. In the *narrative-speech* conditions all error-rates are up at the same level as *Clear-NoAudio*. In case the *DOV* effect would only be present in the *no audio* condition then an interaction effect and not a main effect would be seen which was not the case. As a consequence the *DOV* is working across the conditions. In conclusion, there is an overall effect of the *DOV* across all conditions which is most pronounced in the control groups (*Clear-NoAudio* & *Obscured-NoAudio*) and having some effect in the other conditions but this can only be seen in aggregate. Yet the question arises, which audio distractor can be found more prominently in the real-world?

Participants in all *Obscured* conditions made fewer errors then in the *Clear* conditions which adds to the key finding that the *DOV* could be seen, although in aggregate, in number transcription tasks regardless of *narrative-speech* or *white-noise* distractors. Perhaps, the *DOV* can even be a focusing instrument that can essentially screen out certain distractors of having any effect in the first place such as *white-noise* or specific distractors are actually part of the effect itself.

Participants made significantly more errors in the *narrative-speech* condition. It might be that the *narrative-speech* was semantically meaningful to participants and therefore had such a distracting effect (Jones et al., 2012). Given the fact that the narrative was a slightly humorous story told by Admiral McRaven (see Section 5.3.4), although irrelevant to the task, may have captured the attention of participants to such an extent that they made more errors. This possibly resulted in influencing the *DOV* effect.

Hughes, Vachon, Hurlstone, Marsh, Macken, and Jones (2011) report that making the primary task more attentionally-demanding by making the to-be remembered items (digits or letters) harder to read would result in participants slowing down i.e. a speed-accuracy trade-off but the effects of the audio distractor would diminish. The *DOV* could require a certain level of attention which would account for such an effect. In relation to Hughes et al. (2011) a speed-accuracy trade-off might indicate this but was not seen in the study presented here. It should be mentioned that the study by Hughes et al. (2011) was however a serial recall task.

In terms of corrections there was a main effect for visibility although marginal the effect was confirmed by the *Mann-Whitney* test. This means that obscuring the to-be transcribed information had a direct impact on now many corrections participants made in this study. Overall, fewer corrections were made in the *Obscured* conditions except in the *Obscured-WhiteNoise* conditions. Earlier in Study Nine it was speculated that the *DOV* effect, being a source related effect, with having only the source manipulated in Study Nine and Ten this can now be seen as established. People make fewer errors in the first place therefore make fewer corrections due to deeper cognitive processes. The higher mean in the *Obscured-WhiteNoise* can be explained by the better encoding of the to-be-transcribed information and therefore people would detect that they made an error and correct themselves. No main effect of audio not an interaction effect was detected. That means that the distractors did not have any influence on how many corrections people made. Another reason why the mean in the *Obscured-WhiteNoise* was higher could be a simple motor movement effect that causes participants to make more errors and therefore more corrections (Salthouse & Saults, 1987).

5.4 Summary

This chapter has presented two studies that focused on further extending the knowledge of the *DOV* effect in an interactive number-transcription task. Study Nine successfully replicated the *DOV* effect and further investigated its robustness once levels of distractions are present. However, no effect of the auditory distractors on the errors and error-rate made by participants could be detected. The only effect that the distractors had was the letters distractor on the total corrections made in the study. This raised questions about the effectiveness of the chosen distractors. While there were explanations for the results in the literature it was decided that more effective auditory distractors are needed to further examine the robustness of the *DOV* effect.

The second study (Study Ten) employed two levels of auditory distractors *white-noise* and *narrative-speech*. Both distractors were well established in the literature of being effective on the task performance. Yet, the effects of *white-noise* are not clear and still debated in the literature. Nevertheless, the *DOV* could be seen across all the conditions. However, in detail only the control conditions showed a clear presence of the *DOV*. Although *narrative-speech* increased the errors and the error-rate more than other distractors people made still fewer errors in the individual *Obscured* condition (although not significantly except in the control conditions). As hinted in the literature but not directly expected the *white-noise* distractor did not have a greatly distracting effect. Surprisingly, it may be that *white-noise* might itself be a part of the *DOV* and therefore did not have a distracting effect. Table 5.9 shows a summary of the two studies.

The results of both studies indicate that the *DOV* has the potential to provide real-world benefits in safety critical domains, for example by obscuring drug prescriptions to promote more careful number entry into syringe pumps. Although studies in Chapter Four and Five have presented evidence that there might be a novel way to make transcription tasks more accurate by counter-actively presenting the information in a harder to read font colour, much is still unknown about the dynamics of this effect. More work is required to understand the full application (i.e. benefits and limitations) of this novel approach.

A real-world application of disfluency relies on the results of further investigation into the long-term use of the technique, and the potential interplay between disfluency, real-world distractions, and existing protocols. Further investigation is needed to establish if the effects of disfluency change over time, i.e. do people become immune to its effects? Additionally, further studies need to focus on identifying clear and effective distractors for the task and demonstrate the *DOV* in this context as well as including a multi-tasking environment rather than straight distractions to move closer to a real-world ward environment.

At this moment there is no real-world implementation in safety-critical systems of the findings presented here or the *DOV* in general. Additionally, current *DOV* literature only focuses on reproducing the effect (Alter & Oppenheimer, 2009; Diemand-Yauman et al., 2011). This thesis on the contrary showed over several studies that the *DOV* can be reproduced or applied in transcription tasks. Moreover, it went one step further in presenting studies with the aim to move the *DOV* into real-world applications in safety-critical environments. Yet there are still many questions about how the *DOV* reacts in such an environment, a concern that is reflected in the current literature (Yue et al., 2013; Kuehl et al., 2014). Future studies need to address this gap to gain a more complete understanding of the *DOV* and its impact on error reduction in transcription tasks. The *DOV* represents a small and cheap manipulation i.e. a small change in the presentation of the to-be transcribed content with a potentially huge impact on safety in terms of accuracy. Key findings of this Chapter are:

The key findings of Chapter Five are:

- The *DOV* is present in distracting environments
- Narrative-speech significantly increase the error-rate in transcription tasks
- The *DOV* is a source related effect
- No speed-accuracy trade-off once DOV is present
- Persistent low error presentation rate in number-transcription tasks

#	Interface	IV	DV	Design	Hypothesis	Outcome	Conclusions
9	Serial	Number presentation at the Source (Clear vs. Obscured); Auditory distractors (letters vs. numerals)	Total numbers entered Total errors made Total corrections made Total error rate	between (10 each)	Fewer errors in Obscured than Clear Numerals more distracting than letters	No significant difference in the total numbers entered; Significant difference in the total errors and error rate (marginal); Main effect for audio in terms of total corrections;	No speed-accuracy trade-off; DOV present (although marginal) despite distractors; surprisingly numerals are not more distracting as letters; perhaps task not sufficiently demanding enough; possible motor-memory interference; modest effect size; maybe the task has its own focusing interference – perhaps naturally resistant; closer examination of the effect of the distractors not possible due to the lack of control groups; further study required to give more conclusive results; next study should focus on alternative audio distractors based on the literature and provide control groups;
10	Serial	Number presentation at the Source (Clear vs. Obscured); Auditory distractors (none vs. white noise vs. narrative speech)	Total numbers entered Total errors made Total corrections made Total error rate	between (15 each)	Fewer errors in Obscured than Clear Distractors have impact on errors and error rate (Narrative speech more distracting than white noise)	No significant difference in the total numbers entered; marginal significant difference of visibility in the total errors (obscured fewer errors than in the clear conditions); significant difference in the audio distractors (narrative speech causing significantly more errors); same seen in error rate (but DOV can be seen more clearly here); marginal effect of visibility on corrections but no effect of audio;	No speed-accuracy trade-off; DOV is present despite distractors; Narrative Speech had the most distracting impact on the performance of participants whereas white-noise seem to have had no distracting effect at all; perhaps white-noise is part of the DOV; no significant difference of corrections in terms of audio distractors however marginal significant difference of visibility meaning the DOV has an impact on how many corrections people make; in this case fewer corrections are made in obscured conditions despite distractors supporting earlier views that better encoding support people to compare numbers in memory therefore make fewer corrections in the first place or perhaps do not need to compare them as the better encoding prevents errors; Narrative speech distractor worked as expected in increasing the errors; DOV somewhat counteracted the distraction effects; Perhaps the better encoding resulted in more capacity to process task irrelevant information;

Table 5.11: Studies 9 - 10: Summary of findings

Chapter 6

Thesis Conclusion

"Blast medicine anyway! We've learned to tie into every organ in the human body but one. The brain! The brain is what life is all about."

Dr Leonard McCoy

Star Trek - The Original Series, "The Menagerie"

6.1 Overview & Synthesis

This thesis focused on the effect of design features for transcription tasks to reduce errors. More specifically, the thesis was situated in the context of healthcare systems, focusing on number transcription tasks, such as programming infusion pump devices or entering patient data in forms. At first, Chapter Three presented studies to reduce number entry errors in a transcription task in a more feature driven level by empirically evaluating the impact of different design features for *5-Key* interfaces on errors. It was found that previous results presented by Cauchi et al. (2012) could not be supported as no evidence for their proposed design features of significantly reducing errors could be found. While the analysis presented by Cauchi et al. (2012) might be somewhat appropriate it is not representative enough. Chapter Three, however, presents a rich body of knowledge surrounding number-transcription studies with valuable insights on how people specifically interact with *5-Key* interfaces. Of particular importance was the discovery of the specific strategies that people use when entering numbers into *5-Key* devices. Critically, the analysis of Cauchi et al. (2012) did not take user strategies into account. Moreover, another important finding was that the *5-Key* interface seems to be a very robust design itself in terms of errors. It is not trivial to see any significant effects and for this reason it was decided to move away from this design to another one that is both widely used in real interfaces and the literature: the *Serial* interface (Oladimeji et al., 2011). After further reviewing the literature it was then decided to step back from the feature driven level of Chapter Three to a theoretical driven level in Chapter Four to reduce number transcription errors based on the concepts surrounding the DPT (Stanovich & West, 2002; Kahneman, 2008, 2011; Diemand-Yauman et al., 2011).

Dual-Processing Theory suggests that people possess two distinct ways of thinking *System 1 & System 2*. One which is the quick thinking and another one which is the slow thinking or in other words this part provokes a more effortful thinking. Inspired by these theories it was further investigated how specific design features could somehow engage *System 2* i.e. more effortful, deeper cognitive processes when transcribing information. It was discovered that the literature provides another theory which builds upon DPT and presents a way to stimulate *System 2*: disfluency theory. Disfluency theory states that by interrupting the normal flow of reading information or in other words by making content harder to read can lead to higher learning outcomes (Yue et al., 2013). Further evidence from previous experiments by Diemand-Yauman et al. (2011) showed that indeed there are indications that people could be influenced by design elements to memorise and retrieve information more accurately.

As a result the thesis moved to identify and to collect evidence for what the author called the *DOV* effect in text transcription tasks in Chapter Four. The first attempt was made by using a text-entry game as research instrument and the study indeed revealed evidence for the *DOV* effect. The next step was to investigate if this effect could be successfully transferred to a number transcription task, which if so would indicate that the effect is possibly applicable for multiple domains and not limited to text transcription. Study Five aimed to detect this effect but failed due to possible environmental effects. Study Six addressed these issues and was able to successfully measure, for the first time, a significant effect of the *DOV* effect in number transcription. Subsequently, the question arose what design feature would be responsible for the detected effect. Was obscuring the source of the to-be-transcribed information responsible or was it due to the obscured display on the interface? Study Seven failed to detect any significant difference but was influenced by possible fatigue effects. However, Study

Seven could establish that for number-transcription tasks it is not recommended to use a within-participants design as in this particular study participants had to enter a large amount of numbers in four consecutive tasks, which lead to participants being less engaged. Study Eight, which used a between-participants design, was able to measure the *DOV* effect. Moreover, Study Eight could establish that the source was responsible for this effect.

In Chapter Five two additional studies were presented, which focus not only on the existence of the *DOV* effect in transcription tasks but also look further into how this effect would behave in a distracting environment moving towards more ecologically validated domains, for example, hospital wards. Study Nine looked at how robust the *DOV* effect is related to levels of audio distractions and revealed a marginally significant effect when people transcribe numbers. However, the introduced auditory distractors did not have the expected effect on errors. It could be the case that the obscured numbers could itself be a source of perceptual load and therefore lead to less susceptibility to distraction.

There is also a discussion in the literature that people process distractors to different extents depending on the type of processing demands. For example, a higher cognitive load i.e. deeper cognitive processes in a relevant task reduced processing of irrelevant distractors (Lavie & De Fockert, 2003). Study Ten introduced more more effective auditory distractors based on previous findings in the literature (Banbury et al., 2001; Salame & Baddeley, 1982) namely *white-noise* and *narrative-speech*. This had the result that indeed the auditory distractors had an effect on the errors people made. Although only *narrative-speech* was responsible for that. Interestingly, *white-noise* might be a part of the *DOV* effect itself. Nevertheless, the *DOV* was present and to some degree counteracted the effects of the distractors. People made fewer errors in each obscured condition. In summary the key findings of this thesis are:

- The error presentation rate in all experiments is low but persistent
- No effect of a *left-cursor-start* design in *5-Key* interfaces on errors could be seen
- It was seen that people prefer strategies to enter-numbers in a 5-Key interface
- The analysis by (Cauchi et al., 2012) is not sophisticated enough to account for strategies
- The *5-Key* interface appears to be really robust in terms of errors
- DOV does work in interactive text- as well as number transcription tasks
- DOV is a source related effect
- Within-participants designs are disadvantageous in this particular number transcription experiments
- No speed-accuracy trade-off was seen once the DOV is present
- The DOV is present in environments with auditory distractors

Section 1.4.5 in this thesis discussed the systematicity of the error rates in this thesis (typically between 1-2%) in contrast to the established view of systematicity in cognitive science (5%). Although, the systematic error rates in the experiments presented here are well below 5% they are persistent at the same level and suggest that this is simply the systematicity in number-transcription tasks. Yet there is a discrepancy between what can be viewed as “good science” (5%) and “good safety”. In terms of “good safety”, aviation standards classify low error rates less than 1.0×10^{-5} , but greater than 1.0×10^{-7} as extremely remote and errors less than 1.0×10^{-7} but greater than 1.0×10^{-9} as extremely improbable (FAA, 2000). It could be argued that the presented error rate influenced by the DOV can still not be seen as “safe” in the industry, however the DOV depict an appreciable reduction in errors. On the other hand more naturalistic experiments were chosen as despite their lower error rates than, for example, approaches presented in (Cauchi et al., 2012) to empirically investigate how people interact with interfaces influenced by the DOV. Such studies give valuable insight into behavioural patterns that would otherwise not have been possible to see.

6.2 Research Contributions

6.2.1 What design features can be used to ameliorate the effects of human error in entering numbers into digital healthcare devices?

There are a vast number of medical devices in use everywhere. A great number of them are infusion pump devices used in hospitals. Many of these devices use different interfaces and follow different operating procedures. Cauchi et al. (2012) have proposed that some design features for infusion-pumps have a greater impact on how many errors people would commit should they be present or not based on their analysis. This thesis evaluated the effect of one particular feature: the start position of the cursor in a *5-Key* interface used on, for example, B.Braun infusion pumps. The reason for choosing this particular feature is because, according to Cauchi et al. (2012), this feature seem to analytically have the highest impact on people's performance in relation to errors. Study One, Two and Study Three focused on investigating if this was indeed the case. The first initial experiment was not able to collect sufficient data to be able to run a proper empirical analysis and can be considered as a pilot but Study Two identified no significant effect of where the cursor starts on the errors committed in a *5-Key* interface. This indicated that designing *5-Key* interfaces with a permanent *left-cursor-start* position would be no different from having permanent *right-cursor-start* positions. Study Three introduced a secondary task to increase the detected low error rate but again was unable to detect a difference between the cursor-start position in terms of errors. However, the strategies that people used to transcribe the numbers seem to indicate that people prefer to enter the numbers from left-to-right. Study One revealed that people who started in the *right-cursor-start* position changed their strategy once they spotted the *Wrap* functionality and later in Study Two, which had the *Wrap* feature disabled are indications that people change their strategy from right-to-left to enter the number from left-to-right when they use a *right-cursor-start* design. This is a very interesting behavioural pattern that was also indicated in Study Three. The findings about the strategy that people use to enter numbers is only indicative but not conclusive but it does support the theory that people who use a *right-cursor-start* design are uncomfortable with entering the data from right-to-left and prefer the other way around which might introduce an interference effect and influence the amount of errors. The literature clearly provides many more features which could be of importance in this context (Cauchi et al., 2012). While no impact on the errors could be detected important behavioural patterns could be unveiled. Other features as the one presented in (Cauchi et al., 2012) might reveal similar behavioural insights. The general low error-rate in Study One, Tow and Three (even with a secondary task present) has shown that the *5-Key* design itself is robust in terms of errors.

6.2.2 Can the *DOV* be used to ameliorate the effects of human error by reducing the error-rate in transcription tasks?

Earlier work by Alter et al. (2007), Diemand-Yauman et al. (2011) showed that there is a strong indication that harder to read i.e. obscured information can be beneficial in terms of memory and accuracy. Yet no evidence for such an effect in an interactive context existed.

This thesis has presented a series of studies to answer the above mentioned research question. The results show that the *DOV* does indeed have an appreciable effect on ameliorating errors in transcription tasks. While initial experiments showed that the *DOV* significantly reduced the error-rate in text-transcription tasks, later studies also confirmed the *DOV* in number transcription. This also means that the *DOV* is not limited to either domain and potentially widely applicable to text and number content. Moreover, the effect seems to show some robustness towards auditory distractors providing a first view of how the *DOV* would work in a real-world environment. Surprisingly, findings in Chapter Five suggest that some auditory stimuli may be beneficial in terms of reducing errors or may even contribute or stimulate the *DOV*.

Previous work in the context of studying human error has suggested that a way to ameliorate errors is to focus on cognitive features - especially for errors where training and experience is not much of help (Back et al., 2007). The studies in Chapter Four and Five have shown that the cognitive driven design features which stimulate the *DOV* effect are in fact a novel approach to reduce transcription errors in an interactive context.

6.2.3 How does the *DOV* work with a view of moving to real-world applications?

Humans perceive information through their sensory inputs such as via visual and auditory perception. This information is then processed and encoded in memory. Once that information is needed it is then recovered from memory in retrieval processes. In case of writing or typing, the retrieved information is translated into motor-movements (Salthouse, 1986; Inhoff et al., 1986). Figure 6.1 illustrates this process. For example, a person who transcribes a number would collect the number from the source via various eye-movements and encodes and stores the number in memory. The person would later retrieve the number, plan motor actions and execute these motor actions to enter the number in, for example, an infusion pump interface.

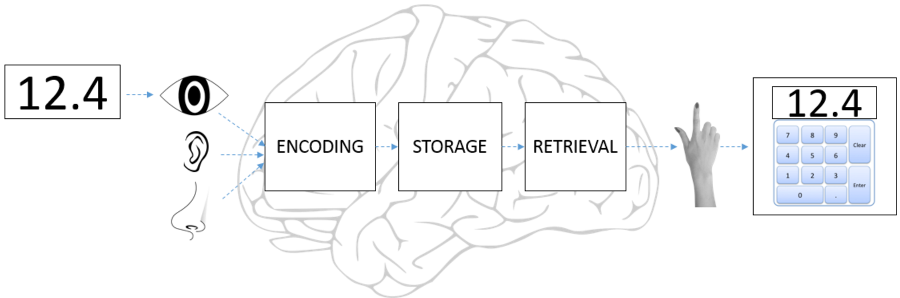


Figure 6.1: Number processing diagram

Study Eight has shown that the *DOV* effect is a source related effect which aligns with the theory surrounding it in the literature (Salthouse, 1986; Kahneman, 2008). The work by Diemand-Yauman et al. (2011) used a visual perceptual manipulation to activate the *DOV* effect in the same way as Study Eight (although the manipulation of Study Eight was only limited to the font colour). Other researcher support this view and propose similar visual manipulations to trigger deeper cognitive processing of the perceived content (Alter et al., 2007; Kahneman, 2011). Study Six and Eight give further indications that the *DOV* effect is source related in terms of the corrections participants made during the studies. It is assumed that people make more comparisons in memory due to the stronger encoding not via the display and therefore make fewer errors in the first place.

Based on the supporting arguments in the literature and the evidence presented in this thesis. It seems the way the *DOV* works is by obscuring information (grey faded font colour) at the source. Deeper cognitive processes are then activated and are responsible for an enhanced encoding, storage and retrieval process. Figure 6.2 illustrates how the *DOV* operates once activated. The *DOV* effect does suggest a stronger encoding after collecting the harder to read visual information leading to a subsequently more accurate motor action once the information is about to be entered in the interface.

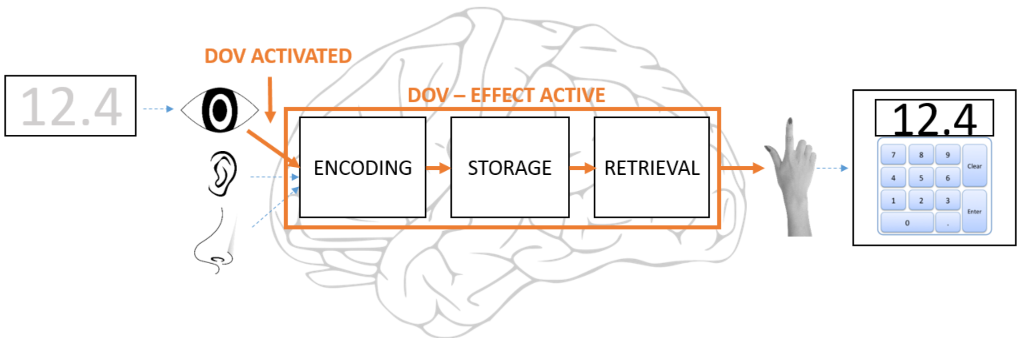


Figure 6.2: Number processing once DOV is present

Slip errors and lapses occur due to memory or attentional failure ((Reason, 1990)) and are prominent in automatic *System 1* processes (Sternberg, 1996). Sternberg (1996) argues that slips also occur when people depart from a routine task, due to task interruptions caused by distractions. Firstly, the *DOV* effect presents an appreciable effect that can override *System 1* processes by invoking a deeper controlled (*System 2*) thinking and therefore reduce the error-rate in transcription tasks. Chapter Four presents studies which support this argument. Secondly, Chapter Five presents studies which show that the *DOV* seems to have some robustness towards auditory distractions. The studies in Chapter Five show how the *DOV* effect works in view of moving towards real-world applications.

In general, it is assumed that auditory distractors are interfering or disrupting the encoding process of the perceived information leading to a higher error-rate (see figure 6.3). Yet some auditory distractor may have been more intrusive than others. For example, the letters and numerals distractors did not have an effect on the error-rate in Study Nine. Perhaps, the numeric distractor was somewhat task relevant to numerical entry whereas letters was not therefore the irrelevant distractor was not processed equally (Lavie & De Fockert, 2003). The *DOV* seem to show some robustness towards numerical and auditory distractors as the effect could be seen (although marginal). It is assumed that the enhanced encoding process that is triggered by the *DOV* increased engagement in the task which reduced the effect of auditory distractions or shielded the cognitive processes of being interrupted by the distractors (Halin et al., 2014a, 2014b). However, Narrative-speech has increased the error-rate and possibly has a weakening or disrupting influence on the *DOV* effect. It may be that narrative-speech was semantically meaningful to participants and therefore had such a distracting effect (Jones et al., 2012).

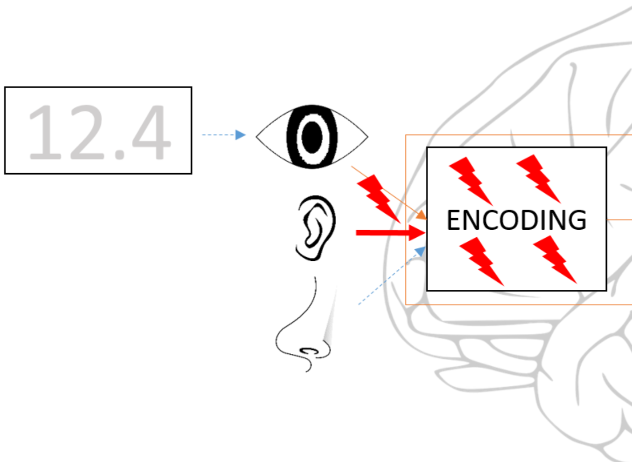


Figure 6.3: Audio distractions interfering with encoding processes

Interestingly *white-noise* did not seem to have the same interrupting effect. Maybe *white-noise* was not demanding enough to be processed at the same level as *narrative-speech* and participants could easily screen-out the distractor. Much more interesting is however, if *white-noise* is actually provoking or supporting the *DOV* effect. Alternatively, it could also have an entirely different effect by just cancelling out the general noise around a person which may contribute to focusing effects similar to previous work (Harrison & Kelly, 1989; Soederlund et al., 2010).

Nevertheless, Chapter Five has shown that the *DOV* effect can be seen once auditory distractions are present. There may be practical applications of the effect in real-world distracting environments where the *DOV* can increase accuracy by reducing the influence of distractions on the encoding of information in memory. Perhaps the *DOV* can even shield cognitive processes entirely from distractions as seen in figure 6.4.

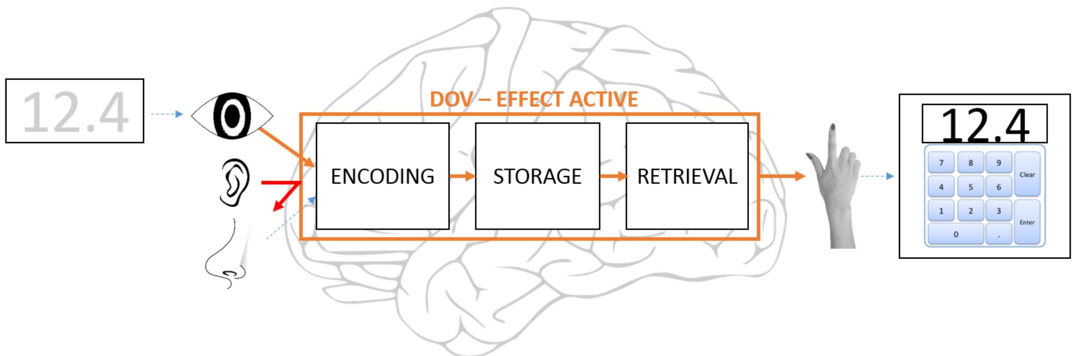


Figure 6.4: DOV effect prevents audio distractions from interfering with encoding

6.2.4 What are the implications of the *DOV* effect in context of transcription tasks?

Study Seven and Eight tried to measure which design element, either the source or the target display is responsible for the *DOV* effect to manifest itself. Study Seven failed due to what was identified as possible environmental effects. However, Study Eight strongly indicated that the source is responsible, which aligns with the theories surrounding the effect as discussed in previous sections and in the literature. Later studies (Nine and Ten) only used source manipulations therefore it can be seen that the *DOV* is source related. The implications of the *DOV* effect in context of transcription tasks would theoretically be enormous. As many different kinds of transcription tasks could be made more accurate with little effort. Designing an interface which is able to stimulate the *DOV* effect can essentially act as another safety layer in safety-critical environments such as aviation or healthcare. Moreover, as the stimuli itself to provoke the *DOV* effect can be as little as presenting the to-be-transcribed source in a grey (faded) font colour, implications for the industry might be very cost-effective as no extra equipment is needed. Apart from safety-critical domains transcription tasks occur in everyday situations and therefore, as the *DOV* effect is essentially an artefact in transcription tasks, possible applications could be forms in general. For example on-line banking forms or flight-ticket booking forms. How to do this in practice is unclear and yet need to be investigated. Additionally, as Study Nine and Ten show the *DOV* can be used regardless of an auditory distracting environment. Even while some distractors may result in an increase of errors the *DOV* still counteracts this effect to some degree which maybe due to the different theories surrounding auditory distractors (Lavie & De Fockert, 2003; Banbury et al., 2001; Tellinghuisen & Nowak, 2003).

6.3 Validity

6.3.1 Internal Validity

Transcribing text or numbers seems to be a trivial task that requires no specialised training. Natural variation is always present when entering numbers. Some people might be able to enter them faster some might be slower and therefore enter less numbers. This fact was present throughout the studies including the one focusing on text-entry. How many numbers or sentences people could transcribe was to some extent dependent on their previous experience with the interfaces that they encountered in the specific studies. For example, in the text-entry study results were hugely dependent on the typing ability of the individual participants. Indeed people entered more sentences in the *Clear* condition but not significantly. As previously mentioned, *5-Key* interfaces as the ones used in Study One, Two and Three may not be so common in daily life compared to *Serial* interfaces. In view of the *Serial* interface based studies, many participants stated that they had no previous experience with using a number-pad or even the iPad to enter numbers but this did not have a significant impact in the studies presented here. It is not known how this factor would have influenced the results with a greater sample size of people in either of the following extreme conditions: very experienced or not experienced at all. In general, people will always vary when transcribing information in view of accuracy and speed. Some people might have been slower when transcribing information by simple getting used to the interface while others did not have to make such efforts and therefore entered more numbers. There were people who made no errors at all while others made several in a row. Additionally, the majority of the participants in all of the studies were students from different disciplines such as Mathematics, Computer Science, Nursing and many others. They did not seem to perform better or worse due to that fact but they might still enter numbers at a frequent pace in their daily life; a fact that is hard to control for.

Furthermore, there could be a cultural factor that had an impact on the studies which was indicated in early Studies. In Study One and Two results might have been strongly dependent on what cultural background people had in view of the strategy that people entered the numbers. For example, left-to-right writing countries might have difficulties accepting an interface where they were required to enter the number from right to left. Additionally, several participants commented on having difficulties entering the correct number as they were confused by how the to-be-entered numbers were presented. For example, if they had to transcribe the number 1,023.00 people from countries like Germany or Austria would normally switch the comma and the decimal point in their meaning.

The visual representation itself could have been, to some degree, a confound for studies presented in Chapter Four. Although, all participants stated that they had normal vision and did not suffer from dyslexia the studies were not designed to measure the exact "sweet spot" of when the grade of obscurity begins to have an effect on the performance of people other than the simple attempt in Study Five which presented a possible range of where a "sweet spot" might be found.

Focusing on the error data which was produced in all of the studies it becomes clear that it was generally very low. The measures undertaken to increase the pressure and to resemble real-world settings might not have been sufficient for some participants.

As transcribing and entering more than 60 to 130 numbers can be seen as very tiring, therefore fatigue could have been another factor that had influenced the results of individuals. Additionally, it is yet unknown to what extent the visual representation had also contributed to people getting tired and becoming less engaged. This became clear especially in Study Seven where people due to the within-participants design were asked four times to transcribe numbers in a time-frame of five minutes for each task. The alternative hypothesis had to be rejected but results indicated that people began to show fatigue effects.

6.3.2 External Validity

Most of the participants used in the studies were students from the University of York. With some exceptions of people working in different professions. The students were not limited to a single discipline like computer science but from many diverse study backgrounds. Number transcription tasks can be seen as relatively simple tasks and therefore the same results may very likely be found by participants who were not students. Only one participant in all studies stated a disability. In this case dyscalculia. None was dyslexic, or had other visual or hearing impairments. At least not that they were aware of at the time of the specific study.

There are many more applications for the findings presented in this thesis. Applications which are not limited to the vast healthcare sector but other domains where accuracy is of great importance. For example, in banking procedures, aviation safety or manufacturing processes; simply everywhere where people need to transfer numerical information from one medium into the other. In banking procedures, people would benefit from a higher accuracy when filling out on-line transfer forms or interacting with a cash machine. In aviation check lists are the bread and butter for operating an aircraft. Being able to memorise information better by little more than making the to-be transcribed content harder to read, which then leads to a better encoding and retrieval from memory (and might even have a beneficial effect to reduce errors in a distracting environment or even counteract distractions) to avoid any errors is an inspiring and cost-effective way to avoid errors.

There were errors due to cultural effects discovered in Study One and Two. However, no such effects were detected in the other experiments. In general, there is a possible interference effect as for example participants in *5-Key* experiments saw numbers which had a comma and a decimal point in its structure. The decimal point is fixed in the *5-Key* interface therefore participants were possibly confused and fell back to their cultural background to enter numbers. *Serial* interfaces do not have a specific comma key (only the decimal point) participants can not enter a comma but have to interpret the comma and have to rethink of what the meaning of the comma is. However, later experiments in Chapter Four and Five had a number range of 0-999.99 with 5% of the numbers being randomly generated without any decimal places at all and therefore participants could not have experienced an interference in the first place as there were no numbers high enough to be presented with a comma. Additionally, collecting participants in an University one has to be pragmatic to get participants who are sensitive to the types of errors that might occur. The errors due to cultural effects were detected as a very small effect in the first studies but they disappeared partially due to the different task designs. In the first experiments (Chapter Three) participants had to transcribe numbers from cards but in later studies directly from the screen.

6.3.3 Ecological Validity

All experiments except Study Five took place in a controlled environment. In this case an office in the computer science building or a study room in the Morell library. This was done to make sure that no confounds were influencing the experiments and were subsequently altering the results. In case of Study Five, a classroom setting was chosen to collect the data as the experiment ran as part of a first-year undergraduate module. It can be argued that none of the studies represent a busy hospital environment, where nurses or doctors who need to transcribe information such as programming infusion pumps or complete patient data forms are under constant pressure or influenced by high cognitive load, stress or time pressure. However, this thesis presents studies to understand the basic science underpinning the interaction with number-transcription interfaces (including traditional or cognitive driven design features) in the context of errors. This is a novel approach and therefore revealed a translation problem towards the healthcare environment but not towards the data-entry environment.

Participants in all studies were therefore asked to enter the given numbers as fast as possible, as accurate as possible or both. This was being done to replicate a situation where people are required to transcribe information quickly and accurately. For example, a nurse is sometimes required to program multiple infusion pumps (a stack) mounted on one desk to serve multiple patients in a room (Back & Cox, 2013) while monitoring other inputs or reacting to other aspects in the environment. Study Five, does resemble a busy environment to some extend as students were chatting and introducing some noise during the study as they were asked to run studies on each other. Additionally, studies Nine and Ten investigated how the *DOV* effect would behave in the real-world by introducing distractions in the tasks. The literature shows that auditory distractions are very common in a nurses daily routine. The most severe distractions are by patients themselves. Although, the positive results from Study Nine and Ten it should be said that they are still laboratory studies and to some degree not representative to a real-world hospital ward.

In any case, none of the measures taken to move the laboratory studies close to a real-world healthcare transcription environment can be seen as equivalent to the latter. As previously outlined, many factors can influence errors. Especially, disastrous errors are mostly not due to a single wrong errors they are a chain of human errors and system failures that usually lead to a major adverse event as seen in the *Air Inter 148* incident (Reason, 1990). Although one step in such a chain of events may well be a simple transcription error all the experiments are specifically targeted to investigate how people perform in simple transcription tasks but still people can act differently in real-world settings.

All the interfaces used in the studies presented here are designed after commonly used real-world infusion-pump interfaces. Among them the *5-Key* interface designed after a B.Braun infusion pump and the *Serial* interface designed after the Baxter infusion-pump. Both the *5-Key* and the *Serial* interface used for studies in this thesis were initially implemented as a web application running in the browser. Participants were required to interact with the simulators via keyboards where the specific functionality behind the design features were mapped onto specific keys. This introduced an interference as people may interact with real-world devices differently. Oladimeji et al. (2013) found that people who use *Chevron* interfaces usually do not focus on the interface while people who use *5-Key & Serial* interfaces do but maybe not to that degree as in the experiments presented in this thesis. Study Nine and Ten, provided an interface closer to a real-world infusion pump display as participants were asked to transcribe numbers via touch interaction on an iPad.

Additionally, in all the number-transcription experiments participants had to transcribe a series of numbers, usually between 60-120 numbers each, which is not likely to happen in a hospital. However, to some degree nurses have to program several infusion pumps or fill-out several forms at the same time which requires them to sometimes interrupt one task (Back et al., 2010). It might be that the amount of errors is therefore higher than in a daily work schedule of a nurse in a hospital. On the other hand, general data entry tasks indeed require people to transcribe large amounts of numbers. Nevertheless, the tasks were designed with the purpose to increase the data to work with in this field.

6.4 Future Work

By looking at Chapter Three, future work can consist of studies on interface elements such as proposed in (Cauchi et al., 2012) to identify the impact of features other than the ones chosen in Study One, Two and Three. However, future studies need to consider the limitations of *5-Key* devices in terms of error-rate. As shown in this thesis the error rate is very low making possible empirical effects very hard to see. Moreover, this thesis revealed limitations in the analysis of Cauchi et al. (2012) which need to be taken into account for any future work based on their findings. The design features they present are (i) ad-hoc and (ii) the impact of those features might not come through in empirical studies as the keystroke approach by Cauchi et al. (2012) does not account for people's preference in strategies when entering numbers. Specifically, these strategies can also be exploited in further experiments.

By looking at Chapter Four and Five, this thesis has shown a series of experiments providing evidence that the *DOV* effect can be stimulated in interactive number- as well as text-transcription tasks. Moreover, it has been established that the *DOV* effect seems to be a source related effect which is in line with the surrounding literature of how the effect works (Kahneman, 2008; Diemand-Yauman et al., 2011; Yue et al., 2013). Additionally, *DOV* revealed beneficial characteristics in task performance of experiments once distractors were introduced in the study environment. Therefore it is only a natural choice to take these findings to the next level as future work may at some point involve real-world studies to establish how the results presented in this thesis perform in environments such as hospital wards or flight decks. Ideal laboratory studies are to a certain degree representative but there might still be specific aspects and confounds that laboratory studies can not account for. For example such studies could focus on obscuring drug dosage prescriptions in the same way as presented in this thesis. Two groups of nurses of a hospital ward would then be examined over a period of one week comparing the performance of the two groups to see how the DOV effect behaves. Additionally, as Chapter Four has shown the effect is not limited to the number-entry domain but can also be seen in text entry. Therefore, a similar study can evaluate the obscurity of forms used by nurses in their daily routine. Interesting would also be, in this context, to see how mixed obscured content behaves.

Another aspect is the amount of information that can be processed by the DOV. The studies in this thesis have only used short text phrases and individual numbers but it is not known if a larger obscured content such as an entire data in a form or list will have the same effect on improving accuracy and if there is a threshold of when the DOV comes into play in terms of the size of the to be processed information.

Moreover, future studies could involve applying the grade of obscurity onto real medical device interfaces to see how the people perform in terms of accuracy. It was not possible to get real infusion-pumps for the studies presented in this thesis. A study including read devices also requites a task design including transcribing VTBI and rate numbers similar to programming a real infusion pump. The entire programming process of a real-world pump would involve not only the volume number of the drug that should be administered but also the rate and other steps such as turning the device on and off which is missing in the studies presented here.
Interestingly, *white-noise* may be part of the DOV effect and therefore further studies on the interplay between the DOV and auditory distractors are required.

Another interesting aspect for future studies would be to focus on dyslexic participants as they show a preference for obscured information rather than clear presented black on white text or numbers. Further to that additional studies that look into luminance in particular of how the to be transcribed information is presented would be required to investigate if the detected *DOV* effect is dependent on a particular font colour only on a white background or if other background colours would also have an influence on it. Some student experiments have already been conducted which are not included in this thesis as they are considered mere scruffy pilots. Moreover, future studies could explore to what extent the *DOV* effect manifests itself if a second level of obscurity is present as in (Diemand-Yauman et al., 2011). This means that for example, not only an obscured font-colour can be used as manipulation but font-colour and font-type combined.

As the *DOV* effect aligns with DPT of *System 1 & System 2* and as System 2 is characterised to be a learning system although slow compared to *System 1*. Further studies are required to look into long time effects of how long does the *DOV* effect last. Moreover, to what extent can it be applied before it becomes too difficult to have an effect as indicated in Study Five. In relation to that, further studies could aim to narrow the range presented in Study Five or even pinpoint the exact location where a possible "sweet-spot" might be. An early attempt to explore when disfluency would be desired was made by Yue et al. (2013) but was not able to establish this fact.

Additionally, future studies investigating the *DOV* effect would need to include eye-tracking measures. Eye-tracking would provide further data about where do people actually focus on when transcribing numbers. It would also provide a further measure to what extend the participants focus more on the source when transcribing information. Eye-tracking data could further provide clear results as discussed earlier that people who are more experienced with specific interface designs would focus more on the display when entering data rather then the number pad itself. Most importantly, data from eye-trackers would be able to reveal if participants' pupils dilate. As previous research shows this is a strong indication for high cognitive activity which would further add to DPT (*System 1 & System 2*) that harder to read information can lead to people engage in deeper thought processes and therefore encode and retrieve information better resulting in fewer errors.

No eye-tracking data was available for the studies which would have provided valuable insights into where people focus during the task. Such data would confirm that, if participants focus more on the source display, target or the keypad, which could confirm the theories mentioned above. In other words it would reveal if the information processing strategies adopted by those in the two conditions (*Clear* and *Obscured*) result in differences in the number of fixations on, and the speed of reading of the target content. Specifically, it will be important to understand whether the enhanced performance of participants in the *Obscured* group is the result of slower more deliberate processes such as additional perceptual monitoring of the target content resulting in enhanced memory for the information or more careful execution of the motor movements required to enter the content (Salthouse, 1986; Inhoff et al., 1986; Gray et al., 2006; Kahneman, 2011). Moreover, eye-tracking data might also reveal in studies which contain distractors how participants behave in terms of collecting information and correcting themselves. For example, depending on the level of distractor the tracking data might reveal if participants focus multiple times on the source information or go back to it more than one time before entering the number or the data might reveal theories surrounding the corrected errors in Study Nine and Ten. If participants focus more on the source of the to-be-transcribed information while they correct themselves this might give evidence that comparisons are not made in memory as currently assumed but simply with the source. However, this might also depend on the interference of a specific distractor.

A substantial body of number and error data could be collected throughout the conducted experiments. Future work maybe involve the classification of the errors made and extend previous attempts or classifications as presented in (Wiseman, 2011).

At last, as results presented in this work provide evidence that indicates that the detected effect is bound to the source in transcription tasks further studies in other domains such as filling out web forms are important to see how the *DOV* unfolds in these areas.

6.5 Concluding Thoughts

Although the weight of the work presented in this thesis, is clearly on investigating the *DOV* effect (Chapter Four & Chapter Five) in interactive transcription tasks, much more work remains to be done to investigate both presented research avenues: traditional design features and design features based on cognitive theory. Despite the fact that no significant effects of the traditional design features evaluated in Chapter Three could be found, there are many more features proposed by the literature that can be investigated in future studies. However, this comes with the caveat that possible effects might be too small to be seen and depend highly on personal strategies. A fact that previous analytical approaches such as shown in Cauchi et al. (2012) did not address. Future analysis approaches should take this into account or if an analysis approach is similar to the one in Cauchi et al. (2012) it should perhaps only be used as first indicative investigation but clearly needs refinement and further empirical work.

The novel approach presented in this thesis has shown that there is indeed potential for increasing the accuracy in transcription tasks by counter-intuitively making information harder to read, which indicates a better memory encoding and retrieval. This view is supported by other studies in the field (Diemand-Yauman et al., 2011). However, as Daniel Kahneman (Kahneman, 2011) said: *"We have a very narrow view of what is going on"*, there is much more work that requires further investigation in order to make any clear, definitive and wider recommendations for future design applications in transcription tasks. Additionally, it is also too early at this stage to make any suggestions with the aim to change existing design methods and principles in the field. The results in this work however, provide a very promising approach to further understand how design features can influence accuracy in transcription tasks. While there is still much that is unknown about the full implications of this work, and in the words of Laura Nyro: *"As beautiful as simplicity is, it can become a tradition that stands in the way of exploration."*, what has been presented here has shown that traditional approaches may not always be the only option when it comes to increasing accuracy in transcription tasks.

Appendices

Appendix A

Chapter 3

A.1 Experiment 1

A.1.1 Experimental Instructions & Informed Consent

Number-Entry Experiment Informed Consent

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenski, Paul Cairns
Dept of Computer Science, York
March, 2012

1 Who is running this?

The study is being run by:

- Frank Soboczenski who is a research student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a senior lecturer in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a 5-key interface to enter numbers.

3 What will I have to do?

You will be asked to complete two number entry tasks. For this you will have a training session before each task starts, where you can get familiar with the controls and the interface.

For the first and the second task you will be given a list of 33 numbers on cards and asked to enter these numbers into the presented interface as quickly and as accurately as possible. You will only have a short timeframe of 3 Minutes to enter all the numbers in both tasks.

At the end, Frank will ask you to fill out a small questionnaire about yourself. He will then debrief you and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can only use the keyboard to enter the numbers!

Please, also only use your index finger to enter the numbers.

1

The interface which you will see consists of 7 components. A number display, a cursor, four control buttons and an ok button . You can move the cursor with the control buttons left or right and increase or decrease the selected number field with buttons up and down . By pressing the current displayed number will be saved and the display will be cleared.

Note: The keyboard is labeled accordingly to the buttons

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Feel free to ask Frank any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have about the purpose or background of the experiment until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

A.1.2 Demographic Questionnaire

Number-Entry experiment demographic questionnaire

Participant number:

What is your age?

Please tick as appropriate:
- ☐ 18-25
- ☐ 30-39
- ☐ 40-49
- ☐ 50-59
- ☐ 60 or above

Are you a student at the University of York?

(If not please state your profession:)

____

If you are a student, what course are you enrolled in?

____

What country are you from?

____

Are you familiar with a B.Braun infusion pump?

Please tick as appropriate: ☐yes ☐no

Do you have working experience in the healthcare sector?

____

Are you a right- or left-handed or both?

____

1

A.2 Experiment 2

A.2.1 Experimental Instructions & Informed Consent

2nd Number-Entry Experiment Informed Consent

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenski, Paul Cairns
Department of Computer Science, York
April, 2012

1 Who is running this?

The study is being run by:

- Frank Soboczenski who is a research student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a senior lecturer in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a 5-key interface to enter numbers.

3 What will I have to do?

You will be asked to complete a number entry task. For this you will have a training session before the main task starts, where you can get familiar with the controls and the interface.

For the main task you will be given a list of 40 numbers on cards and asked to enter these numbers into the presented interface as quickly and as accurately as possible. You will only have a short timeframe of 4 minutes to enter all the numbers.

At the end, Frank will ask you to fill out a small questionnaire about yourself. He will then debrief you and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can only use the keyboard to enter the numbers!

Please, also only use your index finger to enter the numbers.

1

The interface which you will see consists of 7 components. A number display, a cursor, four control buttons  and an ok button . You can move the cursor with the control buttons left  or right  and increase or decrease the selected number field with buttons up  and down . By pressing  the current displayed number will be saved and the display will be cleared.

Note: The keyboard is labeled accordingly to the buttons

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Feel free to ask Frank any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have about the purpose or background of the experiment until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

A.2.2 Demographic Questionnaire

2nd Number-Entry Experiment Demographic Questionnaire

Participant number:

What is your age?

Please tick as appropriate:
☐ 18-25
☐ 26-29
☐ 30-39
☐ 40-49
☐ 50-59
☐ 60 or above

Are you a student at the University of York?

(If not please state your profession:)

____

If you are a student, what course are you enrolled in?

____

What country are you from?

____

Are you familiar with a B.Braun infusion pump?

Please tick as appropriate: ☐yes ☐no

Do you have working experience in the healthcare sector?

____

Are you right- or left-handed or both?

____

1

A.3 Experiment 3

A.3.1 Experimental Instructions & Informed Consent High

Experiment 3 - Informed Consent

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenski, Paul Cairns
Department of Computer Science, York
March, 2015

1 Who is running this?

- Frank Soboczenski who is a PhD student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a reader in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a 5-Key interface to enter numbers.

3 What will I have to do?

You will be asked to complete two tasks (at the same time): a number-entry task and a listening task. For this you will have a training session before the start of the main experiment, where you can get familiar with the controls of the interface and the procedure of the tasks.

You will be asked to enter numbers into the presented interface as quickly as possible. You will only have a short timeframe of **6 minutes** to finish this task. While you enter the numbers you will be asked to listen to an audio track and to count the appearance of all the vowels **(A,E,I,O,U)**.

At the end, you will be asked to fill out a small questionnaire about yourself. Frank will then debrief you and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can use the 5-Key interface on the iPad to enter the numbers.

Please, only use your index finger to enter the numbers.

1

The interface which you will see consists of 7 components:

A number display, a cursor, four control buttons  and an ok button 

. You can move the cursor with the control buttons left  or right  and increase or decrease the selected number with buttons up  and down . By pressing 

the current displayed number will be saved and the display will be cleared.

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiment's data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

A.3.2 Experimental Instructions & Informed Consent Low

Experiment 3 - Informed Consent

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenski, Paul Cairns
Department of Computer Science, York
March, 2015

1 Who is running this?

- Frank Soboczenski who is a PhD student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a reader in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a 5-Key interface to enter numbers.

3 What will I have to do?

You will be asked to complete two tasks (at the same time): a number-entry task and a listening task. For this you will have a training session before the start of the main experiment, where you can get familiar with the controls of the interface and the procedure of the tasks.

You will be asked to enter numbers into the presented interface as quickly as possible. You will only have a short timeframe of **6 minutes** to finish this task. While you enter the numbers you will be asked to listen to an audio track and to count the appearance of the **letter A**.

At the end, you will be asked to fill out a small questionnaire about yourself. Frank will then debrief you and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can use the 5-Key interface on the iPad to enter the numbers.

Please, only use your index finger to enter the numbers.

1

The interface which you will see consists of 7 components:

A number display, a cursor, four control buttons     and an ok button . You can move the cursor with the control buttons left  or right  and increase or decrease the selected number with buttons up  and down . By pressing  the current displayed number will be saved and the display will be cleared.

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiment's data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

A.3.3 Demographic Questionnaire

Experiment 3

*Required

Participant Number *

Are you a student at the University of York (if not please state your profession)? *

Sex *

- male
- female

What is your age? *

- 18 - 25
- 26 - 29
- 30 - 39
- 40 - 49
- 50 - 59
- 60 +

What Country are you from? *

How many times did the letter "A" occur in the audio track? *

Are you familiar with syringe pumps? *

- Yes
- No

Do you have working experience in the healthcare sector? *

- Yes
- No

Are you familiar with 5-Key interfaces? *

- Yes
- No

How often do you use an iPad or other tablet? *

- never used
- rarely used
- sometimes used
- often used

How often do you use an iPad or other tablet to enter numbers? *

- ◯ never used
- ◯ rarely used
- ◯ sometimes used
- ◯ often used

How often do you use a touch based device in general? *

- ◯ never used
- ◯ rarely used
- ◯ sometimes used
- ◯ often used

Do you consider yourself to have normal vision (when corrected)? *

- ◯ Yes
- ◯ No

Do you consider yourself to have normal hearing (when corrected)? *

- ◯ Yes
- ◯ No

Did you speak the numbers in your mind during the task? *

- ◯ Yes
- ◯ No

Did you enjoy the task? *

	1	2	3	4	5	
Did not enjoy	◯	◯	◯	◯	◯	enjoyed very much

To what degree did you notice the audio track? *

	1	2	3	4	5	
barely noticed	◯	◯	◯	◯	◯	noticed very much

To what degree did the audio track make you focus more on the number entry task? *

	1	2	3	4	5	
did not make me focus more	◯	◯	◯	◯	◯	made me focus much more

To what degree did the audio track distract you from the task? *

	1	2	3	4	5	

did not distract me at all ○ ○ ○ ○ ○ did distract me a lot

To what extent did you feel the task was difficult? *

	1	2	3	4	5	
not at all	○	○	○	○	○	very difficult

Did you find it difficult to use the iPad to enter numbers? *

	1	2	3	4	5	
No, very easy	○	○	○	○	○	Yes, very difficult

To what extent did you feel focussed on the task? *

	1	2	3	4	5	
not at all	○	○	○	○	○	very focussed

How much effort did you put into the task? *

	1	2	3	4	5	
no effort	○	○	○	○	○	a lot

To what extent did you have the feeling that you made an error? *

	1	2	3	4	5	
no error	○	○	○	○	○	multiple errors

To what extent do you have the feeling that you would have made less errors without the audio track? *

	1	2	3	4	5	
would not have made less errors	○	○	○	○	○	definitely would have made less errors

To what extent did you memorise the numbers? *

	1	2	3	4	5	
did not memorise	○	○	○	○	○	memorised them a lot

Please indicate for each of the twelve statements which is

closest to how you have been feeling over the last two hours. Notice that higher numbers mean better well-being.

I felt cheerful and in good spirits. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt calm and relaxed. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt active and vigorous. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt alert. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt attentive. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt energetic. *

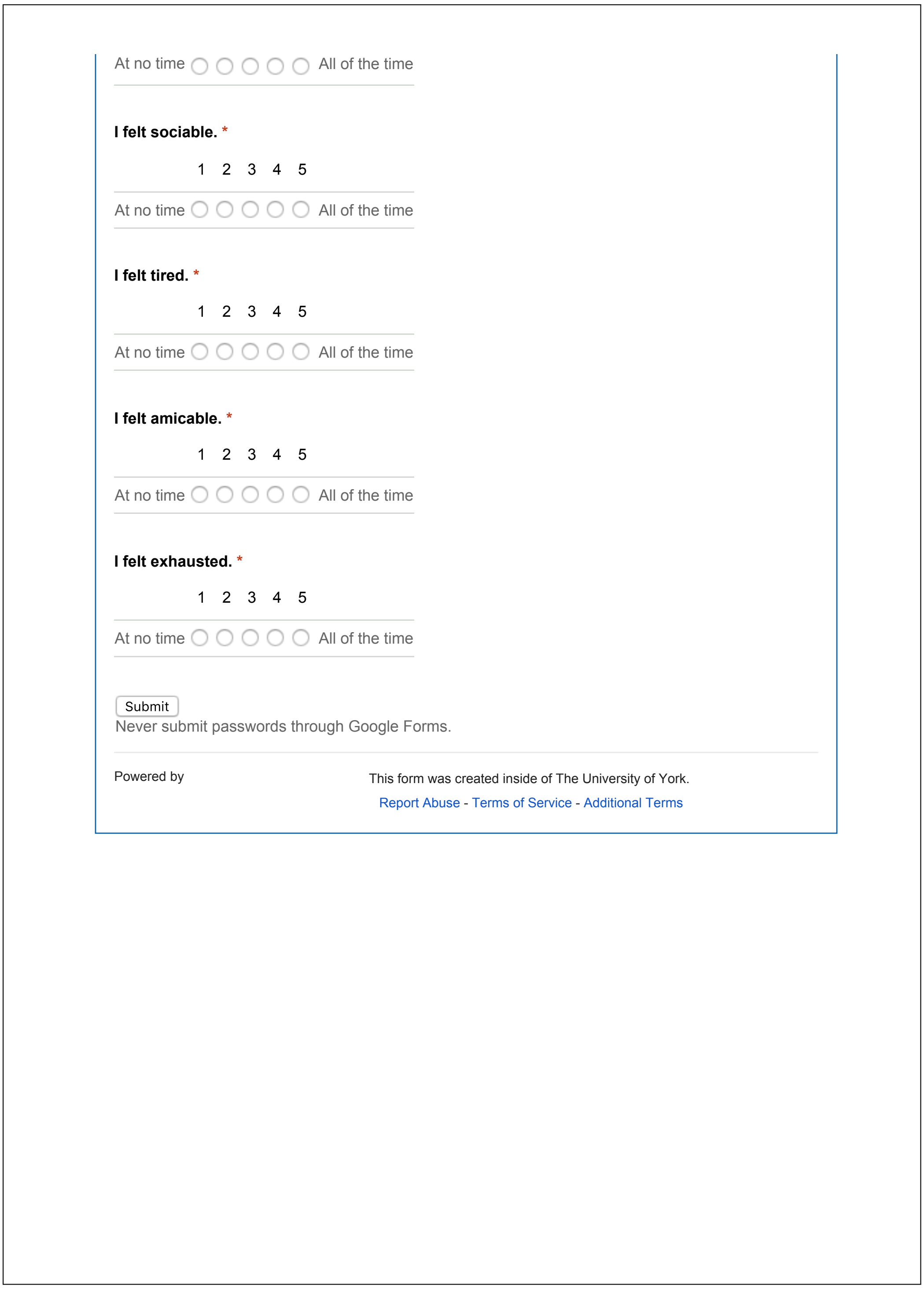
	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt excited. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt happy. *

	1	2	3	4	5	



At no time ○ ○ ○ ○ ○ All of the time

I felt sociable. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt tired. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt amicable. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt exhausted. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

Submit

Never submit passwords through Google Forms.

Powered by

This form was created inside of The University of York.

Report Abuse - Terms of Service - Additional Terms

A.3.4 Secondary task results Study Three

Participant	Correct A's	Participant	Correct vowels
1	20	21	48
2	42	22	40
3	10	23	180
4	5	24	157
5	17	25	35
6	12	26	85
7	16	27	64
8	11	28	37
9	16	29	287
10	14	30	71
11	13	31	87
12	18	32	18
13	18	33	47
14	6	34	89
15	22	35	36
16	18	36	26
17	18	37	94
18	10	38	51
19	18	39	26
20	20	40	16

Table A.1: Study 3: Secondary task results

Appendix B

Chapter 4

B.1 Experiment 4

B.1.1 Experimental Instructions & Informed Consent

Typo Madness Experiment: Informed Consent Form

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenski, Paul Cairns
Department of Computer Science, York
June, 2012

1 Who is running this?

The study is being run by:

- Frank Soboczenski who is a research student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a senior lecturer in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people perform in a typing game.

3 What will I have to do?

You will be asked to play the game *Typo Madness*! Typo Madness is a typing game where you will see a variety of text paragraphs which you then need to retype to get a score. The goal is to get a high score by finishing the game in the given time. You will only have a short timeframe of **15 minutes** to complete the game.

At the beginning you will have a training session before the game starts, where you can get familiar with the interface and how the game works.

At the end, Frank will ask you to fill out a small questionnaire about yourself (where you can leave questions blank if you feel uncomfortable in answering them). He will then debrief you and you may ask any questions about the study or give him any general comments you have.

4 How can play the game to get high score?

You can use the keyboard to type in each given paragraph. However, the *return key*, *backspace* and all *cursor keys* are disabled. Additionally, you can only use the mouse to click the *red* button which will save your text as well as forward you to the next paragraph.

1

Each character you type will be counted for the score. If the character matches with the original text it will increase your score. Therefore, each character that does not match with the original text will decrease your score.

Here is an example:

original text: There is no spoon!
retyped text: There is no Spon!
Score: 13 (15-2 mismatches namely uppercase 'S' and a missing 'o')

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and it will not be possible to identify you by your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will still not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Feel free to ask Frank any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have about the purpose or background of the experiment until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

B.1.2 Demographic Questionnaire

Typo Madness Experiment Demographic Questionnaire

(please feel free to leave questions blank if you feel uncomfortable in answering them)

Participant number:

What is your age?

Please tick as appropriate:

☐ 18-25 ☐ 26-29 ☐ 30-39 ☐ 40-49 ☐ 50-59 ☐ 60 or above

Are you a student at the University of York? ____________

(If not please state your profession:) ____________

If you are a student, what course are you enrolled in? ____________

What country are you from? ____________

How many words would you say can you type per minute?

Please tick as appropriate:

☐less than 20 ☐20-40 ☐40-60 ☐60-80 ☐80 or more ☐do not know

Were the given text phrases easy to read?

Please tick as appropriate:

☐not readable ☐less readable ☐readable ☐clearly readable

Did you find the task difficult?

Please tick as appropriate: ☐Yes ☐No

If *Yes* please state why: ____________

Do you consider yourself to have normal vision (when corrected)?

Please tick as appropriate: ☐Yes ☐No

Do you have dyslexia?

Please tick as appropriate: ☐Yes ☐No

1

B.1.3 Sentences for the Game TypoMadness

Starfleet Commander Benjamin Sisko arrives (with his young son, Jake) at Deep Space Nine, a space station formerly run by the Cardassians.

Targeting Trinity chips at touchscreen laptops provides AMD a wider opportunity to expand its presence in the tablet market.

The prototype device shown at the AMD press conference had a detachable touchscreen that could be used as a tablet when undocked from the laptop.

PC manufacturer Acer has announced a line-up of touchscreen ultrabooks, tablets and all-in one PCs running the eagerly awaited Windows 8 operating system (OS).

Hewlett-Packard is playing catch up, like everyone else in the cloudy infrastructure racket at the Discover 2012 event in Las Vegas.

Amazon figured out companies don't want to buy virtual servers, they want to buy multi-tiered virtual clouds consisting of a mix of servers.

Researcher Joseph Bonneau noted that bad password behaviour exists in every region on the web.

A report from Cambridge University Computer Laboratory has found that users continue to utilise easily-guessed passwords.

After a decade's worth of best-selling action games from Japan, Devil May Cry's Dante has headed West with a new master, Ninja Theory.

ICS appeared to be the most highly anticipated Android operating system, owing to its seamless user interface for multiple devices.

Barely had the dust settled on the dud Honeycomb version of the Android operating system then came the news of the promising Ice Cream Sandwich (ICS).

There were reports of many people dressed in red, white and blue left standing on station platforms across the Midlands.

In 2008, the BMA, other health unions and the government negotiated a major reform of the NHS scheme.

The government has begun to implement major changes to the NHS pension scheme despite widespread criticism.

CIWEM hopes that during this anniversary year politicians will realise that running economies in the way we've always done is clearly not working.

Producing a formula to accurately calculate the animal's dimensions, they found a Brachiosaur would actually have weighed a relatively light 23 tons.

Dinosaurs are traditionally portrayed as monstrously heavy, with thundering footsteps and a huge bulk adding to their fearsome appearance.

Archaeologists in Bulgaria have unearthed two medieval skeletons pierced through the chest with iron rods to keep them from turning into vampires.

The Chinese President greets Russian President Vladimir Putin for the 7th time in China, but this visit has a special significance.

Yesterday Prince Harry and the Duchess of Cambridge smiled and giggled their way through the people.

While Kate's husband Prince William stood formally, his cheeky brother whispered a series of humorous asides to his sister-in-law.

Sandra Piller, widow of beloved Star Trek head writer Michael Piller, will be appearing at the official Star Trek convention in Nashville.

Anticipation for the first Star Trek teaser is running unfortunately very high, the recent two where not officially declared real.

Eager skywatchers flocked to universities and observatories that scheduled viewings and astronomy talks.

The NASA Hubble Space Telescope cannot view the Sun directly to learn more about Venus's atmosphere.

B.1.4 Scores & Sentences by Condition

The total score (number of correctly transcribed characters) of the *Clear* group and of the *Obscured* group was structured into:

Clear group	**Sentences**	**Obscured group**	**Sentences**
1800	20	3117	26
2478	23	3040	26
1867	18	2754	22
3017	26	2979	24
3066	25	3077	25
2379	22	3171	26
3177	26	3001	24
2747	24	2968	24
2721	26	2827	21
1917	20	2932	24
$\sum$ 25,269	230	30,356	242

Table B.1: Study 4: Total scores & sentences transcribed

B.2 Experiment 5

B.2.1 Experimental Instructions & Informed Consent

HACS: Practical 1 Informed Consent

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

1 Who is running this?

The study is being run by Dr Paul Cairns who is a senior lecturer in the department of Computer Science at the University of York. Your experimenter is a student on his module, Human Aspects of Computer Science, and is a first year undergraduate at the department.

2 What is the purpose of this study?

The study aims to investigate how people use a serial (calculator based) interface to enter numbers.

3 What will I have to do?

You will be asked to complete a number entry task. For this you will have a training session before the main task starts, where you can get familiar with the controls and the interface.

For the main task you will asked to enter numbers into the presented interface as quickly as possible. You will only have a short timeframe of 5 minutes to finish the task.

At the end, you will be asked to fill out a small questionnaire about yourself. You will then be debriefed by the experimenter and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can only use the keyboard number-pad to enter the numbers!

The interface which you will see consists of 16 components. Three number displays, a decimal key, an enter key, a clear key and the keys 0-9. You can enter the displayed number by pressing the associated number-pad keys *Note: the '+' key is labelled as CLEAR*. By pressing ENTER the current displayed number will be saved, the displays be cleared and the process starts again. If you noticed that you made a mistake before you pressed ENTER, you can cancel the current input by pressing CLEAR.

1

5 Who will see this data?

Obviously the experimenter with you will see this data. Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed. Unusually for a study, this is a class exercise and Paul strongly advises that you take part to see how such experiments work from both the participant and experimenter perspective.

7 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to the experimenter and save any questions you may have until the end of the practical class. If you have any questions about the purpose or background of the experiment, please wait until the end of the class and you will have an opportunity to ask Paul your questions.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

B.2.2 Demographic Questionnaire

HACS: Practical 1 Questionaire

1 Some details about you

1. Sex (M/F): __________
2. Age (please circle): 17, 18, 19, 20, 21, 22-25, 25-30, 31+
3. What course are you enrolled in? ______________________________
4. What country are you from? ______________________________
5. Are you familiar with syringe pumps? ______________________________
6. Do you have working experience in the healthcare sector? ______________________________
7. Are you familiar with a calculator based interface? ______________________________
8. How often do you use the numpad of the keyboard to enter numbers (please circle) ?
   never used; rarely used; sometimes used; often used
9. Did you find the task difficult? ______________________________

1

B.3 Experiment 6

B.3.1 Experimental Instructions & Informed Consent

Experiment 6 - Informed Consent

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenkski, Paul Cairns
Department of Computer Science, York
November, 2012

1 Who is running this?

- Frank Soboczenski who is a research student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a reader in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a serial (calculator based) interface to enter numbers.

3 What will I have to do?

You will be asked to complete a number entry task. For this you will have a training session before the main task starts, where you can get familiar with the controls and the interface.

For the main task you will asked to enter numbers into the presented interface as quickly as possible. You will only have a short timeframe of 5 minutes to finish the task.

At the end, you will be asked to fill out a small questionnaire about yourself. Frank will then debrief you and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can only use the keyboard number-pad to enter the numbers!

The interface which you will see consists of 16 components. Three number displays, a decimal key, an enter key, a clear key and the keys 0-9. You can enter the displayed number by pressing the associated number-pad keys *Note: the '+' key is labelled as CLEAR*. By pressing ENTER the current displayed number will be saved, the displays be cleared and the process starts again. If you

1

noticed that you made a mistake before you pressed ENTER, you can cancel the current input by pressing CLEAR.

For every number that you enter correctly you will **increase your score by 100 points**. You can see your score displayed to the right of the interface.
For every incorrect number your score will **decrease by 100 points**. This also applies should you accidentally press 'enter' at any point without having typed the correct number. Basically you will decrease your score by any action other than entering the correct number.

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

B.3.2 Demographic Questionnaire

Experiment 6

Some details about you

*Required

Participant Number: *

[ ]

Sex *

- [ ] male
- [ ] female

What is your age? *

- [ ] 18-25
- [ ] 26-29
- [ ] 30-39
- [ ] 40-49
- [ ] 50-59
- [ ] 60+

Are you a student at the University of York (if not please state your profession)? *

please state the course you're enrolled in if you are a student

[ ]

What Country are you from?

[ ]

Are you familiar with syringe pumps? *

- [ ] Yes
- [ ] No

Are you familiar with a B.Braun infusion pump? *

- [ ] Yes
- [ ] No

Do you have working experience in the healthcare sector? *

- [ ] Yes
- [ ] No

Are you right- or left-handed or both? *

- [ ] Left
- [ ] Right
- [ ] Both

Did you find the task difficult? *

- [ ] Yes
- [ ] No

Do you consider yourself to have normal vision (when corrected)? *

- [ ] Yes
- [ ] No

Please indicate for each of the twelve statements which is closest to how you have been feeling over the last two hours. Notice that higher numbers mean better well-being.

I have felt cheerful and in good spirits. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt calm and relaxed. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt active and vigorous. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt alert. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt attentive. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt energetic. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt excited. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt happy. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt sociable. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt tired. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt amicable. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I have felt exhausted. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

Submit

Never submit passwords through Google Forms.

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

B.4 Experiment 7

B.4.1 Experimental Instructions & Informed Consent

Experiment 7 - Informed Consent

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenkski, Paul Cairns
Department of Computer Science, York
March, 2013

1 Who is running this?

- Frank Soboczenski who is a PhD student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a reader in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a serial (calculator based) interface to enter numbers.

3 What will I have to do?

You will be asked to complete four number-entry tasks. For this you will have a training session before the start of the main tasks, where you can get familiar with the controls and the interface.

For the four main tasks you will asked to enter numbers into the presented interface as quickly as possible. You will only have a short timeframe of 5 minutes to finish each task.

After you have finished a number-entry task you will be asked a number of questions about the task followed by a short break until the next task starts.

At the end, you will be asked to fill out a small questionnaire about yourself and Frank will ask you a few qualitative questions about the study which will be recorded. Frank will then debrief you and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can only use the keyboard number-pad to enter the numbers!

1

The interface which you will see consists of 16 components. Three number displays, a decimal key, an enter key, a clear key and the keys 0-9. You can enter the displayed number by pressing the associated number-pad keys *Note: the '+' key is labelled as CLEAR*. By pressing ENTER the current displayed number will be saved, the displays be cleared and the process starts again. If you noticed that you made a mistake before you pressed ENTER, you can cancel the current input by pressing CLEAR.

For every number that you enter correctly you will **increase your score by 100 points**. You can see your score displayed to the right of the interface.

For every incorrect number your score will **decrease by 100 points**. This also applies should you accidentally press 'enter' at any point without having typed the correct number. Basically you will decrease your score by any action other than entering the correct number.

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way. In addition, all the recordings of the qualitative questions will be deleted after the data is transcribed into the spreadsheet.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have until the end of the study.

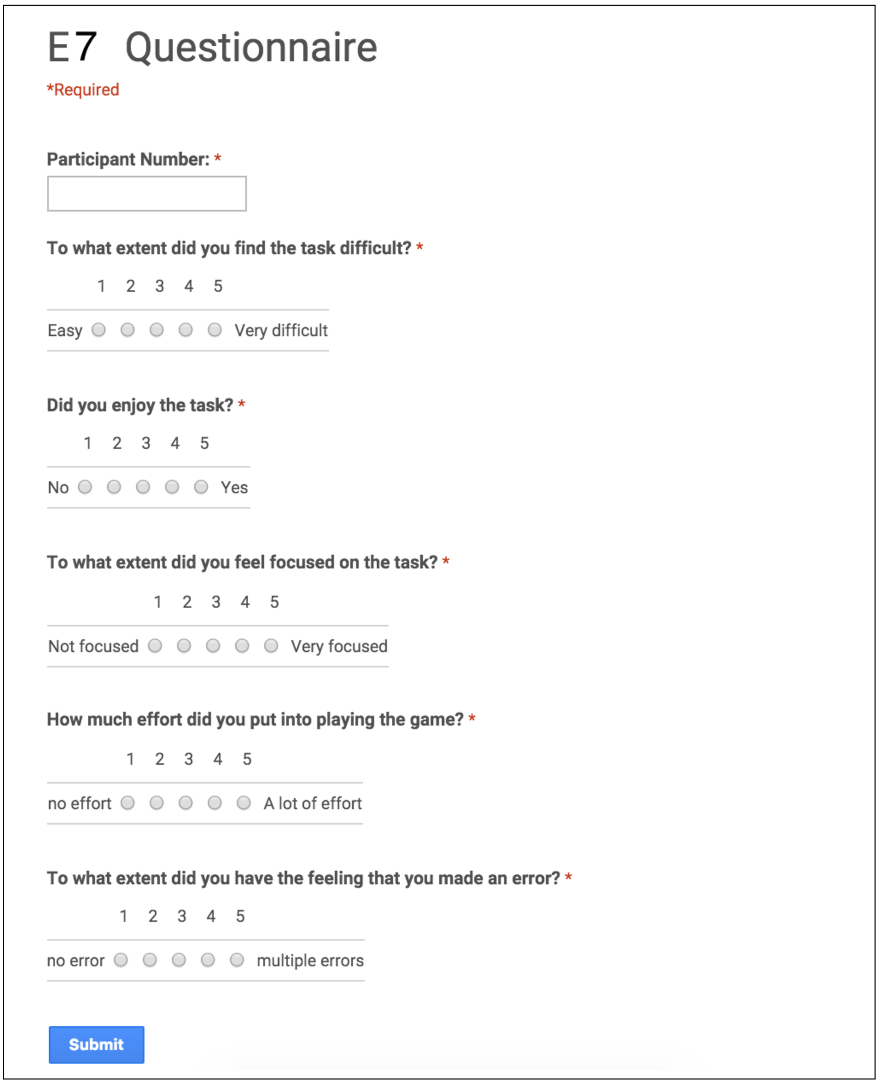
8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:
Signature:
Date:

2

B.4.2 Demographic Questionnaire 1


E7 Questionnaire

*Required

Participant Number: *

To what extent did you find the task difficult? *

	1	2	3	4	5	
Easy	○	○	○	○	○	Very difficult

Did you enjoy the task? *

	1	2	3	4	5	
No	○	○	○	○	○	Yes

To what extent did you feel focused on the task? *

	1	2	3	4	5	
Not focused	○	○	○	○	○	Very focused

How much effort did you put into playing the game? *

	1	2	3	4	5	
no effort	○	○	○	○	○	A lot of effort

To what extent did you have the feeling that you made an error? *

	1	2	3	4	5	
no error	○	○	○	○	○	multiple errors

Submit

B.4.3 Demographic Questionnaire 2

Experiment 7

Some details about you

*Required

Participant Number: *

Sex *

- [ ] male
- [ ] female

What is your age? *

- [ ] 18-25
- [ ] 26-29
- [ ] 30-39
- [ ] 40-49
- [ ] 50-59
- [ ] 60+

Are you a student at the University of York (if not please state your profession)? *

please state the course you're enrolled in if you are a student

What Country are you from?

Are you familiar with syringe pumps? *

- [ ] Yes
- [ ] No

Do you have working experience in the healthcare sector? *

- [ ] Yes
- [ ] No

Are you familiar with calculator-based interfaces? *

- [ ] Yes
- [ ] No

How often do you use the numpad of the keyboard to enter the numbers? *

- [ ] never used
- [ ] rarely used
- [ ] sometimes used
- [ ] often used

Do you consider yourself to have normal vision (when corrected)? *

- [ ] Yes
- [ ] No

Please indicate for each of the twelve statements which is closest to how you have been feeling over the last two hours. Notice that higher numbers mean better well-being.

I felt cheerful and in good spirits. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt calm and relaxed. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt active and vigorous. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt alert. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt attentive. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt energetic. *

	1	2	3	4	5	

At no time ◯ ◯ ◯ ◯ ◯ All of the time

I felt excited. *

1 2 3 4 5

At no time ◯ ◯ ◯ ◯ ◯ All of the time

I felt happy. *

1 2 3 4 5

At no time ◯ ◯ ◯ ◯ ◯ All of the time

I felt sociable. *

1 2 3 4 5

At no time ◯ ◯ ◯ ◯ ◯ All of the time

I felt tired. *

1 2 3 4 5

At no time ◯ ◯ ◯ ◯ ◯ All of the time

I felt amicable. *

1 2 3 4 5

At no time ◯ ◯ ◯ ◯ ◯ All of the time

I felt exhausted. *

1 2 3 4 5

At no time ◯ ◯ ◯ ◯ ◯ All of the time

Submit

Never submit passwords through Google Forms.

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

B.5 Experiment 8

B.5.1 Experimental Instructions & Informed Consent

Experiment 8 - Informed Consent

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenkski, Paul Cairns
Department of Computer Science, York
June, 2012

1 Who is running this?

- Frank Soboczenski who is a research student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a reader in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a serial (calculator based) interface to enter numbers.

3 What will I have to do?

You will be asked to complete a number entry task. For this you will have a training session before the main task starts, where you can get familiar with the controls and the interface.

For the main task you will asked to enter numbers into the presented interface as quickly as possible. You will only have a short timeframe of 5 minutes to finish the task.

At the end, you will be asked to fill out a small questionnaire about yourself. Frank will then debrief you and you may ask any questions about the study or give him any general comments you have.

4 How can I enter the numbers?

You can only use the keyboard number-pad to enter the numbers!

The interface which you will see consists of 16 components. Three number displays, a decimal key, an enter key, a clear key and the keys 0-9. You can enter the displayed number by pressing the associated number-pad keys *Note: the '+' key is labelled as CLEAR*. By pressing ENTER the current displayed number will be saved, the displays be cleared and the process starts again. If you

1

noticed that you made a mistake before you pressed ENTER, you can cancel the current input by pressing CLEAR.

For every number that you enter correctly you will **increase your score by 100 points**. You can see your score displayed to the right of the interface.

For every incorrect number your score will **decrease by 100 points**. This also applies should you accidentally press 'enter' at any point without having typed the correct number. Basically you will decrease your score by any action other than entering the correct number.

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

B.5.2 Demographic Questionnaire

Experiment 8

Some details about you

*Required

Participant Number: *

Sex *

- [ ] male
- [ ] female

What is your age? *

- [ ] 18-25
- [ ] 26-29
- [ ] 30-39
- [ ] 40-49
- [ ] 50-59
- [ ] 60+

Are you a student at the University of York (if not please state your profession)? *

please state the course you're enrolled in if you are a student

What Country are you from?

Are you familiar with syringe pumps? *

- [ ] Yes
- [ ] No

Do you have working experience in the healthcare sector? *

- [ ] Yes
- [ ] No

Are you familiar with calculator-based interfaces? *

- [ ] Yes
- [ ] No

How often do you use the numpad of the keyboard to enter the numbers? *

- [ ] never used
- [ ] rarely used
- [ ] sometimes used
- [ ] often used

Do you consider yourself to have normal vision (when corrected)? *

- [ ] Yes
- [ ] No

Did you enjoy the task? *

	1	2	3	4	5	
Did not enjoy	○	○	○	○	○	enjoyed very much

To what extent did you feel the task was difficult? *

	1	2	3	4	5	
not at all	○	○	○	○	○	very difficult

To what extent did you feel focussed on the task? *

	1	2	3	4	5	
not at all	○	○	○	○	○	very focussed

How much effort did you put into the task? *

	1	2	3	4	5	
no effort	○	○	○	○	○	a lot

To what extent did you have the feeling that you made an error? *

	1	2	3	4	5	
no error	○	○	○	○	○	multiple errors

Please indicate for each of the twelve statements which is closest to how you have been feeling over the last two hours. Notice that higher numbers mean better well-being.

I felt cheerful and in good spirits. *

	1	2	3	4	5

At no time ◯ ◯ ◯ ◯ ◯ All of the time

I felt calm and relaxed. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt active and vigorous. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt alert. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt attentive. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt energetic. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt excited. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt happy. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt sociable. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt tired. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt amicable. *

1 2 3 4 5

At no time ○ ○ ○ ○ ○ All of the time

I felt exhausted. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

Submit

Never submit passwords through Google Forms.

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

Appendix C

Chapter 5

C.1 Experiment 9

C.1.1 Experimental Instructions & Informed Consent

Experiment 9 - Informed Consent

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenski, Paul Cairns
Department of Computer Science, York
November, 2013

1 Who is running this?

- Frank Soboczenski who is a PhD student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a reader in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a serial (calculator based) interface to enter numbers.

3 What will I have to do?

You will be asked to complete a number-entry task. For this you will have a training session before the start of the main task, where you can get familiar with the controls and the interface.

For the main tasks you will asked to enter numbers into the presented interface as quickly as possible. You will only have a short timeframe of **5 minutes** to finish each task and you will also hear an audio track during the number-entry task.

At the end, you will be asked to fill out a small questionnaire about yourself. Frank will then debrief you and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can use the number-pad interface on the iPad to enter the numbers.

The interface which you will see consists of 16 components. Three number displays, a decimal key, an enter key, a clear key and the keys 0-9. You can enter the displayed number by pressing the associated number-pad keys. By pressing ENTER the current displayed number will be saved,

1

the displays be cleared and the process starts again. If you noticed that you made a mistake before you pressed ENTER, you can cancel the current input by pressing CLEAR *(Note: pressing clear will not decrease your score)*.

For every number that you enter correctly you will **increase your score by 100 points**. You can see your score displayed to the right of the interface.

For every incorrect number your score will **decrease by 100 points**. This also applies should you accidentally press 'enter' at any point without having typed the correct number. Basically you will decrease your score by any action other than entering the correct number.

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

C.1.2 Demographic Questionnaire

Experiment 9

*Required

Participant Number *

Are you a student at the University of York (if not please state your profession)? *

Sex *

- male
- female

What is your age? *

- 18 - 25
- 26 - 29
- 30 - 39
- 40 - 49
- 50 - 59
- 60 +

What Country are you from? *

Are you familiar with syringe pumps? *

- Yes
- No

Do you have working experience in the healthcare sector? *

- Yes
- No

Are you familiar with calculator-based interfaces? *

- Yes
- No

How often do you use the numpad of the keyboard to enter the numbers? *

- never used
- rarely used
- sometimes used
- often used

How often do you use an iPad or other tablet? *

- never used
- rarely used
- sometimes used
- often used

How often do you use an iPad or other tablet to enter the numbers? *

- never used
- rarely used
- sometimes used
- often used

How often do you use a touch based device in general? *

- never used
- rarely used
- sometimes used
- often used

Do you consider yourself to have normal vision (when corrected)? *

- Yes
- No

Do you consider yourself to have normal hearing (when corrected)? *

- Yes
- No

Did you speak the numbers in your mind during the task? *

- Yes
- No

Did you enjoy the task? *

	1	2	3	4	5	
Did not enjoy	◯	◯	◯	◯	◯	enjoyed very much

To what degree did you notice the audio track? *

	1	2	3	4	5	
barely noticed	◯	◯	◯	◯	◯	noticed very much

To what degree did the audio track make you focus more on the number entry task? *

	1	2	3	4	5	
did not make me focus more	◯	◯	◯	◯	◯	made me focus much more

To what degree did the audio track distract you from the task? *

	1	2	3	4	5	
did not distract me at all	◯	◯	◯	◯	◯	did distract me a lot

To what extent did you feel the task was difficult? *

	1	2	3	4	5	
not at all	◯	◯	◯	◯	◯	very difficult

Did you find it difficult to use the iPad to enter numbers? *

	1	2	3	4	5	
No, very easy	◯	◯	◯	◯	◯	Yes, very difficult

To what extent did you feel focussed on the task? *

	1	2	3	4	5	
not at all	◯	◯	◯	◯	◯	very focussed

How much effort did you put into the task? *

	1	2	3	4	5	
no effort	◯	◯	◯	◯	◯	a lot

To what extent did you have the feeling that you made an error? *

	1	2	3	4	5	
no error	◯	◯	◯	◯	◯	multiple errors

To what extent do you have the feeling that you would have made less errors without the audio track? *

	1	2	3	4	5	
would not have made less errors	◯	◯	◯	◯	◯	definitely would have made less errors

To what extent did you memorise the numbers? *

	1	2	3	4	5	
did not memorise	◯	◯	◯	◯	◯	memorised them a lot

Please indicate for each of the twelve statements which is closest to how you have been feeling over the last two hours. Notice that higher numbers mean better well-being.

I felt cheerful and in good spirits. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt calm and relaxed. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt active and vigorous. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt alert. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt attentive. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt energetic. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt excited. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt happy. *

1 2 3 4 5

At no time ◯ ◯ ◯ ◯ ◯ All of the time

I felt sociable. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt tired. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt amicable. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

I felt exhausted. *

	1	2	3	4	5	
At no time	◯	◯	◯	◯	◯	All of the time

Submit

Never submit passwords through Google Forms.

C.2 Experiment 10

C.2.1 Experimental Instructions & Informed Consent 1

Experiment 10 - Informed Consent & Instructions

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenkski, Paul Cairns
Department of Computer Science, York
July, 2015

1 Who is running this?

- Frank Soboczenski who is a research student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a reader in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a serial (calculator based) interface to enter numbers.

3 What will I have to do?

You will be asked to complete a number entry task while wearing headphones (to shield from outside noise). For this you will have a training session before the main task starts, where you can get familiar with the controls and the interface.

For the main task you will be asked to enter numbers into the presented interface as quickly as possible. You will only have a short timeframe of **6 minutes** to finish the task.

At the end, you will be asked to fill out a small questionnaire about yourself. Frank will then debrief you and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can use the number-pad interface on the iPad to enter the numbers.

The interface which you will see consists of 16 components. Three number displays, a decimal key, an enter key, a clear key and the keys 0-9. You can enter the displayed number by pressing the associated number-pad keys. By pressing ENTER the current displayed number will be saved,

1

the displays be cleared and the process starts again. If you noticed that you made a mistake before you pressed ENTER, you can cancel the current input by pressing CLEAR.

For every number that you enter correctly you will **increase your score by 100 points**. You can see your score displayed to the right of the interface.

For every incorrect number your score will **decrease by 100 points**. This also applies should you accidentally press 'enter' at any point without having typed the correct number. Basically you will decrease your score by any action other than entering the correct number.

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

C.2.2 Experimental Instructions & Informed Consent 2

Experiment 10 - Informed Consent & Instructions

The purpose of this form is to tell you about the study and highlight features of your participation in the study.

Frank Soboczenkski, Paul Cairns
Department of Computer Science, York
July, 2015

1 Who is running this?

- Frank Soboczenski who is a research student in the department of Computer Science at the University of York.
- Dr Paul Cairns who is a reader in the department of Computer Science at the University of York.

2 What is the purpose of this study?

The study aims to investigate how people use a serial (calculator based) interface to enter numbers.

3 What will I have to do?

You will be asked to complete a number entry task while wearing headphones and listening to an audio track. For this you will have a training session before the main task starts, where you can get familiar with the controls and the interface.

For the main task you will be asked to enter numbers into the presented interface as quickly as possible. You will only have a short timeframe of **6 minutes** to finish the task.

At the end, you will be asked to fill out a small questionnaire about yourself. Frank will then debrief you and you may ask any questions about the study or tell him any comments you have in general.

4 How can I enter the numbers?

You can use the number-pad interface on the iPad to enter the numbers.

The interface which you will see consists of 16 components. Three number displays, a decimal key, an enter key, a clear key and the keys 0-9. You can enter the displayed number by pressing the associated number-pad keys. By pressing ENTER the current displayed number will be saved,

1

the displays be cleared and the process starts again. If you noticed that you made a mistake before you pressed ENTER, you can cancel the current input by pressing CLEAR.

For every number that you enter correctly you will **increase your score by 100 points**. You can see your score displayed to the right of the interface.

For every incorrect number your score will **decrease by 100 points**. This also applies should you accidentally press 'enter' at any point without having typed the correct number. Basically you will decrease your score by any action other than entering the correct number.

5 Who will see this data?

Frank and Paul will see all of your data and will compile all of the experiments data into a single spreadsheet for further analysis. However, once it has been compiled, it will be completely anonymised and you will not be able to be identified with your data. The experiment may be published in an academic journal but the data will only be presented in summary form and you will not be directly identifiable in any way.

6 Do I have to do this?

Your participation is completely voluntary. You can therefore withdraw from the study at any point and if requested your data can be destroyed.

7 Can I ask questions?

Do ask the experimenter any questions you may have about the procedure that you are about to follow. However, during the study, please refrain from talking to Frank and save any questions you may have until the end of the study.

8 Consent

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that we will be obliged to treat your data as described.

Name:

Signature:

Date:

2

C.2.3 Demographic Questionnaire

Experiment 10

*Required

Participant Number *

Are you a student at the University of York (if not please state your profession)? *

Sex *
- male
- female

What is your age? *
- 18 - 25
- 26 - 29
- 30 - 39
- 40 - 49
- 50 - 59
- 60 +

What Country are you from? *

Are you familiar with syringe pumps? *
- Yes
- No

Do you have working experience in the healthcare sector? *
- Yes
- No

Are you familiar with calculator-based interfaces? *
- Yes
- No

How often do you use an iPad or other tablet? *
- never used
- rarely used
- sometimes used
- often used

How often do you use an iPad or other tablet to enter numbers? *

- never used
- rarely used
- sometimes used
- often used

How often do you use a touch based device in general? *

- never used
- rarely used
- sometimes used
- often used

Do you consider yourself to have normal vision (when corrected)? *

- Yes
- No

Do you consider yourself to have normal hearing (when corrected)? *

- Yes
- No

Did you speak the numbers in your mind during the task? *

- Yes
- No

Please give your opinion about the readability of the presented numbers. *

	1	2	3	4	5	
Hard to read	○	○	○	○	○	Easy to read

Did you enjoy the task? *

	1	2	3	4	5	
Did not enjoy	○	○	○	○	○	enjoyed very much

To what degree did you notice the audio track? *

	1	2	3	4	5	
barely noticed	○	○	○	○	○	noticed very much

To what degree did the audio track make you focus more on the number entry task? *

	1	2	3	4	5	
did not make me focus more	○	○	○	○	○	made me focus much more

To what degree did the audio track distract you from the task? *

	1	2	3	4	5	
did not distract me at all	○	○	○	○	○	did distract me a lot

To what extent did you feel the task was difficult? *

	1	2	3	4	5	
not at all	○	○	○	○	○	very difficult

Did you find it difficult to use the iPad to enter numbers? *

	1	2	3	4	5	
No, very easy	○	○	○	○	○	Yes, very difficult

To what extent did you feel focused on the task? *

	1	2	3	4	5	
not at all	○	○	○	○	○	very focussed

How much effort did you put into the task? *

	1	2	3	4	5	
no effort	○	○	○	○	○	a lot

To what extent did you have the feeling that you made an error? *

	1	2	3	4	5	
no error	○	○	○	○	○	multiple errors

To what extent do you have the feeling that you would have made fewer errors without the audio track? *

	1	2	3	4	5	
would not have made less errors	○	○	○	○	○	definitely would have made less errors

To what extent did you memorise the numbers? *

	1	2	3	4	5	
did not memorise	○	○	○	○	○	memorised them a lot

Please indicate for each of the twelve statements which is closest to how you have been feeling over the last two hours. Notice that higher numbers mean better well-being.

I felt cheerful and in good spirits. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt calm and relaxed. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt active and vigorous. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt alert. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt attentive. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt energetic. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt excited. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt happy. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt sociable. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt tired. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt amicable. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

I felt exhausted. *

	1	2	3	4	5	
At no time	○	○	○	○	○	All of the time

Submit

Never submit passwords through Google Forms.

Glossary

ART Aligned Rank Transformation test for nonparametric factorial analyses

CaCI Calium Cloride

CD Compact Disk

CRM Crew Resource Management

CSS Cascading Style Sheets

DOV Diemand-Yauman, Oppenheimer and Vaughn effect - named after the three researchers. In other places also known as perceptual-interference effect

DPT Dual-Processing Theory

e.g. Latin abbreviation for: for example

EHR Electronic Health Record

FAA Federal Aviation Administration

FDA Food and Drug Administration

GPWS Ground Proximity Warning System

HACS Human Aspects of Computer Science

HCI Human-Computer Interaction

HFACS Human Factors Analysis and Classification System

IAC Information Access Cost

i.e. Latin for "id est"; that is; in other words

ISMP Institute for Safe Medication Practices

ISO International Standardisation Organisation

JS Javascript

JSP Java Server Page

MS Microsoft

MySQL SQL stands for Structured Query Language; its named after co-founder Michael Widenius's daughter, My

NASA National Aeronautics and Space Administration

NHS National Health Service

OSX Apple Operating System version 10

PANAS Positive and Negative Affect Schedule

PC Personal Computer

PHP PHP originally stood for Personal Home Page but it now stands for the recursive backronym Hypertext Preprocessor

SD Standard Deviation

SNARC Spatial Numerical Association of Response Codes

SP1 Microsoft Windows Service Pack 1

UCD User-Centered Design

UCL University College London

UK United Kingdom

US United States

VTBI Volume to Be Infused

XCODE Apple integrated development environment for Objective-C and Swift

Bibliography

Abelson, R. P. (1995). *Statistics as principled argument.* Lawrence Erlbaum Associates, Hillsdale NJ.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology. General*, *136*, 569–576.

Alter, A. L., & Oppenheimer, D. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(2), 219–235.

Aspden, P., Wolcott, J. A., Bootman, L., & Cronenwett, L. R. (2007). *Preventing Medication Errors.* National Academies Press Washington, DC.

Atwal, A., & Caldwell, K. (2006). Nurses' perceptions of multidisciplinary team work in acute health-care. *International Journal of Nursing Practice*, *12*(6), 359–365.

Back, J., Blandford, A., & Curzon, P. (2007). Slip errors and cue salience. *W. Brinkman, D. Ham and B. L. W. Wong (Eds.), Proceedings of the 14th European Conference on Cognitive Ergonomics: Invent! Explore!*, 221–224.

Back, J., Brumby, D., & Cox, A. (2010). Locked-out Investigating the Effectiveness of System Lockouts to Reduce Errors in Routine Tasks. *CHI2010: 28th ACM Conference on Human Factors in Computing Systems.*

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*, pp. 1651–1654.

Back, J., & Cox, A. L. (2013). Artifacts for programmable devices: The good, the bad and the ugly. In *Proceedings of the 2013 Conference on Human Factors in Computing Systems (CHI-2013), Extended Abstracts*, pp. 1731–1736.

Baddeley, A. (2007). *Working memory, thought and action.* Oxford, England: Oxford University Press.

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80.

Banbury, S. P., Macken, W. J., Tremblay, S., & Jones, D. M. J. (2001). Auditory distraction and short-term memory: Phenomena and practical implications. *Human Factors*, *43*(1), 12–29.

BEA (1993). Rapport de la commission d'enquête sur l'accident survenu le 20 janvier 1992 près du mont sainte-odile (bas rhin) à l'airbus a 320 immatriculé f-gged exploité par la compagnie air inter (bea f-ed920120). Tech. rep., Bureau d'Enquêtes et d'Analyses (BEA).

Beaman, C. P. (2005). Auditory distraction from low-intensity noise: A review of the consequences for learning and workplace environments. *Applied Cognitive Psychology*, *19*, 1041–1064.

Blandford, A. (2010). When i get older: Interaction design for medical devices. *Interfaces - BCS Interaction specialist group*, pp. 12–13.

Blandford, A. (2013). Interactions "in the wild": Explorations in healthcare. In *M. Anderson (Ed.), Contemporary Ergonomics and Human Factors 2013: Proceedings of the International Conference on Ergonomics and Human Factors, Croydon, England: CRC Press.*, pp. 3–6.

Braun (2012). B.Braun Space - The new Standard. Tech. rep., B. Braun Melsungen AG.

Brunswik, E. (1955). Representative Design and Probabilistic Theory in a Functional Psychology. *Psychological Review*, *62*(3), 193–217.

Byrne, M. D., & Bovair, S. (1997). A working memory model of a common procedural error. *Cognitive Science*, *21*(1), 31–61.

Byrne, M. D., & Davis, E. (2006). Task structure and postcompletion error in the execution of a routine procedure. *Human Factors*, *48*, 627–638.

Cairns, P., & Cox, A. L. (2008). *Research Methods for Human Computer Interaction.* Cambridge: Cambridge University Press.

Casey, S. M. (1998). Set phasers on stun: and other true tales of design, technology, and human error. *Santa Barbara, CA: Aegean Publishing Company.*

Casey, S. M. (2006). The atomic chef: and other true tales of design, technology, and human error. *Santa Barbara, CA: Aegean Publishing Company.*

Cauchi, A., Gimblett, A., Thimbleby, H., Curzon, P., & Masci, P. (2012). Safer "5-key" number entry user interfaces using differential formal analysis. *Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers*, 29–38.

Clancy, C. M. (2012). Alleviating "second victim" syndrome: how we should handle patient harm. *Journal of Nursing Care Quality*, *27*(1), 1–5.

Colle, H., & Hiszem, K. (2004). Standing at a kiosk: Effects of key size and spacing on touch screen numeric keypad performance and user preference. *Ergonomics*, *47*(13), 1406–1423.

Conrad, R., & Hull, A. (1968). The preferred layout for numeral data-entry keysets. *Ergonomics*, *11*(2), 165–173.

Dalton, B. H., & Behm, D. G. (2007). Effects of noise and music on human and task performance: A systematic review. *Occupational Ergonomics*, *7*, 143–152.

Dalton, P., & Lavie, N. (2004). Auditory attentional capture: effects of singleton distractor sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(1), 180.

Dehaene, S. (1997). *The number sense: how the mind creates mathematics.* Oxford: Oxford University Press.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(371-396).

Dekker, S. (2006). *The Field Guide to Understanding Human Error*. Ashgate Publishing Company.

Diemand-Yauman, C., Oppenheimer, D., & Vaughan, E. (2011). Fortune favors the bold (and the italicized): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111–115.

Ellermeier, W., & Zimmer, K. (1997). Individual differences in susceptibility to the "irrelevant speech effect". *Journal of the Acoustical Society of America*, *102*, 478–487.

Evans, J. S. B. T. (2003). In two minds dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459.

Evans, J. S. B. T., & Stanovich, K. (2013). Dual process theories of cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223–241.

FAA (2010). Federal aviation administration. dot/faa/am-10/13, office of aerospace medicine, causes of general aviation accidents and incidents: Analysis using nasa aviation. safety reporting system data. Tech. rep., U.S. Department of Transportation Press, Washington, DC.

FAA (2013). Operational use of flight path management systems - final report of the performance-based operations aviation rulemaking committee. Report of the parc/cast flight deck automation wgcast flight deck automation wg, Federal Aviation Administration, Flight Deck Automation Working Group.

FAA (2014). Federal aviation administration (incident report air inter flight). `http://lessonslearned.faa.gov/ll_main.cfm?TabID=2&LLID=57&LLTypeID=2` (accessed January 2014).

FDA (2000). Medical device use-safety: Incorporating human factors engineering into risk management. Tech. rep., United States Food and Drug Administration.

Fitts, P. M., & Jones, R. E. (1947). Psychological aspects of instrument display: 1. analysis of 270 "pilot-error" experiences in reading and interpreting aircraft instruments. Tech. rep., Wright-Patterson Air Force Base, Ohio: Air Material Command, Aero Medical Laboratory, Memorandum Report TSEAA-694-12A.

Forster, S., & Lavie, N. (2008). Failures to ignore entirely irrelevant distractors: The role of load. *Journal of Experimental Psychology: Applied*, *14*, 73–83.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, *5*(10), 914–926.

Frankish, K., & Evans, J. S. B. T. (2009). *In Two Minds: Dual Processes and Beyond.* Oxford University Press.

Furniss, D., Blandford, A., Rajkomar, A., Vincent, C., & Mayer, A. (2011). The visible and the invisible Distributed Cognition for medical devices. In *Proceedings of the 1st International Workshop on Engineering Interactive Computing Systems for Medicine and Health Care (EICS4Med)*, pp. 1–6.

Gawande, A. (2011). *The Checklist Manifesto: How to Get Things Right.* Picador.

Gigerenzer, G., & Regier, T. (1996). How do we tell an association from a rule? comment on sloman (1996). *Psychological Bulletin*, *119*, 23–26.

Glass, J. M. (2007). Visual function and cognitive aging: Differential role of contrast sensitivity in verbal versus spatial tasks. *Psychology and Aging*, *22*(2), 233–238.

Goebel, S. M., & Rushworth, M. F. (2004). Cognitive neuroscience: Acting on dispatch numbers. *Current Biology*, *14*(13), 517–519.

Gomes, B., & Higginson, I. (2008). Where people die (1974-2030): past trends, future projections and implications for care. *Palliative Medicine*, *22*(33-41).

Gray, W. D., Sims, C. R., Schoelles, M. J., & Fu, W.-t. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*, 461–482.

Grissinger, M. (2003). The virtues of independent double-checks: They really are worth your time!. *Physical Therapy*, *31*(9).

Grissinger, M. (2011). Fluorouracil mistake ends with a fatality - applying the lessons learned can save lives. *Pharmacy and Therapeutics*, *36*(6), 313–314.

Halin, N., Marsh, J. E., Haga, A., Holmgren, M., & Sörqvist, P. (2014a). Effects of speech on proofreading: Can task-engagement manipulations shield against distraction?. *Journal of Experimental Psychology: Applied*, *20*(1), 69.

Halin, N., Marsh, J. E., Hellman, A., Hellström, I., & Sörqvist, P. (2014b). A shield against distraction. *Journal of Applied Research in Memory and Cognition*, *3*(1), 31–36.

Hammond, K. R., & Summers, D. A. (1972). Cognitive Control. *Psychological Review*, *79*(1), 58–67.

Harrison, D., & Kelly, P. (1989). Age differences in cardiovascular and cognitive performance under noise conditions. *Perceptual and Motor Skills*, 547–554.

Henriksen, K., Dayton, E., Keyes, M., Carayon, P., & Hughes, R. (2008). *Understanding Adverse Events: A Human Factors Framework*. Agency for Healthcare and Research Quality.

Hollnagel, E. (1993). The phenotype of erroneous actions. *International Journal of Man-Machines Studies*, *39*, 1–32.

Hollnagel, E., & Amalberti, R. (2001). The emperor's new clothes, or whatever happened to "human error"?. *Invited keynote presentation at 4th International Workshop on Human Error, Safety and System Development Linkoeping.*

Holzinger, A., Thimbleby, H., & Beale, R. (2010). Human–computer interaction for medicine and health care (hci4med): Towards making information usable. *Int. J. Human-Computer Studies*, *68*, 325–327.

Hughes, R., Vachon, F., Hurlstone, M., Marsh, J., Macken, W., & Jones, D. (2011). *Disruption of cognitive performance by sound: Differentiating two forms of auditory distraction.*

Inhoff, A. W., Morris, R., & Calabrese, J. (1986). Eye movements in skilled transcription typing. *Bulletin ofthe Psychonomic Society*, *24*(2), 113–114.

ISMP (2004). Misprogram a pca pump? it's easy!. *ISMP Medication Safety Alert!*, *9*(15), 1–2.

ISMP (2007). Fluorouracil incident root cause analysis: Follow-up. *ISMP Canada Safety Bulletin*, *7*(4).

ISO (2007). Medical devices – application of usability engineering to medical devices. Tech. rep. IEC 62366, International Standardisation Organisation.

Johnson, E. N., & Pritchett, A. R. (1995). Experimental study of vertical flight path mode awareness. Tech. rep., Massachusetts Institute of Technology - Department of Aeronautics andAstronautics.

Jones, D. M., & Broadbent, D. E. (1991). Human performance and noise. *Handbook of acoustical measurements and noise control. New York: McGraw-Hill.*, 1–24.

Jones, D. M., Madden, C., & Miles, C. (1992). Privileged access by irrelevant speech to short-term memory: The role of changing state. *Quarterly Journal of Experimental Psychology*, *44A*, 645–669.

Jones, D. M., Marsh, J. E., & Hughes, R. W. (2012). Retrieval from memory: Vulnerable or inviolable?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 905–922.

Jones, D. M., Miles, C., & Page, J. (1990). Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory?. *Applied Cognitive Psychology*, *4*, 89–108.

Kahneman, D. (2008). Explorations of the mind intuition. UC Berkeley Graduate Council Lectures, Humanities.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1982). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251.

Kantowitz, B. H., & Sorkin, R. D. (1983). *Human Factors: Understanding People - System Relationships.* New York: Wiley.

Keren, G. (2013). A tale of two systems: A scientific advance or a theoretical stone soup? commentary on evans and stanovich (2013). *Perspectives on Psychological Science*, *8*, 257.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*, 533.

Kohn, L., Corrigan, J., & Donaldson, M. (2000). To err is human: Building a safer healthcare system. *National Academy Press*, 28.

Kostopoulou, O., & Delaney, B. (2007). Confidential reporting of patient safety events in primary care: results from a multilevel classification of cognitive and system factors. *Qual Saf Health Care*, *16*, 95–100.

Kuehl, T., Eitel, A., Damnit, G., & Koendle, H. (2014). The impact of disfluency, pacing, and students' need for cognition on learning with multimedia. *Computers in Human Behavior*, *35*, 189–198.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 451–468.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in cognitive sciences*, *9*(2), 75–82.

Lavie, N., & De Fockert, J. W. (2003). Contrasting effects of sensory limits and capacity limits in visual selective attention. *Perception & Psychophysics*, *65*(2), 202–212.

Lavie, N., Lin, Z., Zokaei, N., & Thoma, V. (2009). The role of perceptual load in object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 42–57.

Lee, S., & Zhai, S. (2009). The performance of touch screen soft buttons. In *Proc. CHI 2009, ACM Press*, pp. 309–318.

Lesar, T. (2002). Tenfold medication dose prescribing errors. *Ann. Pharmacother*, *36*, 1833–1839.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Li, S. Y. W., Blandford, A., Cairns, P., & Young, R. M. (2008). The effect of interruptions on postcompletion and other procedural errors: An account based on the activation-based goal memory model. *Journal of Experimental Psychology: Applied*, *14*(4), 314–328.

Li, Y., & Thimbleby, H. (2012). Hot cheese: A processed swiss cheese model. *Submitted to British Medical Journal Quality and Safety*.

Logan, G., & Crump, M. (2009). The left hand doesn't know what the right hand is doing: The disruptive effects of attention to the hands in skilled typewriting. *Psychological Science*, *20*, 1296.

Macdonald, J. S. P., & Lavie, N. (2011). Visual perceptual load induces inattentional deafness. *Atten Percept Psychophys*, *73*(1780-1789).

Mack, A., & Rock, I. (1998). Inattentional blindness. *Cambridge, MA: MIT Press*.

Macken, W. J., Phelps, F. G., & Jones, D. M. (2009). What causes auditory distraction?. *Psychonomic Bulletin and Review*, *16*(1), 139–144.

Marteniuk, R., Ivens, C., & Brown, B. (1996). Are there task specific performance effects for differently configured numeric keypads?. *Applied Ergonomics*, *27*(5), 321–325.

Martins, I. T., Martins, E. T., Soares, M. M., & da Silva, A. L. G. (2013). Human error in aviation: The behavior of pilots facing the modern technology. In *HCI (11)*, pp. 150–159.

Mayer, R. E. (2009). *Multimedia learning* (2nd edition edition). Cambridge: Cambridge University Press.

McCloskey, M., Caramazza, A., & Basili, A. (1985). Cognitive mechanisms in number processing and calculation: evidence from dyscalculia. *Brain and cognition*, *4*, 171–96.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The psychological review*, *63*, 81–97.

Morgan, P., & Patrick, J. (2010). Designing interfaces that encourage a more effortful cognitive strategy. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 408–412.

Murphy, S., Fraenkel, N., & Dalton, P. (2013). Perceptual load does not modulate auditory distractor processing. *Cognition*, *129*(2), 345–355.

Nairne, J. S. (1988). The mnemonic value of perceptual identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 148–255.

NASA (2011). *NASA Space Flight Human-System Standard - Volume 2: Human Factors, Habitability, and Environmental Health*, Vol. 2. National Aeronautics and Space Administration.

NHS (2009). Saving lives in surgery. Guideline, NHS Institute for Innovation and Improvement and National Patient Safety Agency.

Nol, M.-P., Rouselle, L., & Mussolin, C. (2005). Magnitude representation in children: Its development and dysfunction. In *Handbook of Mathematical Cognition (J. I. D. Campbell, ed.)*. New York and Hove: Psychology Press.

Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review*, *75*(6), 522–536.

Norman, D. A. (1980). Errors in Human Performance. Tech. rep., University of California, San Diego, Center for Human Infromation Processing Report No. 8004.

Norman, D. A. (1981). Categorization of action slips. *Psych. Rev.*, *88*, 1–15.

Norman, D. A. (1983). Design Rules Based on Analyses of Human Error. *Communications of the ACM*, *26*(4).

Norman, D. A. (1990). *The Design of Everyday Things*. The Design of Everyday Things, Basic Books, New York.

Obradovich, J., & Woods, D. (1996). Users as designers: How people cope with poor hci design in computer-based medical devices. *Human Factors*, *38*(4), 574–592.

Oladimeji, P., Thimbleby, H., & Cox, A. L. (2013). A performance review of number entry interfaces. In *Lecture Notes in Computer Science*, Vol. 8117, pp. 365–382.

Oladimeji, P., Thimbleby, H., Curzon, P., Iacovides, J., & Cox, A. (2012). Exploring unlikely errors using video games: An example in number entry research. In *Workshop on Safety-Critical Systems and Video Games: Contradictions and Commonalities, Held at Fun and Games 2012, Toulouse, France.*

Oladimeji, P., Thimbleby, H., & Cox, A. (2011). Number Entry Interfaces and their Effects on Error Detection. *Human-Computer Interaction*, *1*(27), 12.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomatic Bulletin and Review*, *11*, 988–1010.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71.

Park, K. S. (2003). Is There No Such Thing As Human Error?. *Journal of the Ergonomics Society of Korea*, *22*, 95–99.

Rajkomar, A., & Blandford, A. (2012). A distributed cognition model for analysing interruption resumption during infusion administration. In *Proceedings of European Conference on Cognitive Ergonomics (ECCE 2012)*, pp. 108–111. New York: ACM.

Rasmussen, J. (1983). Skill, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models. *IEEE Transactions on Systems Man and Cybernetics*, *79*, 58–67.

Rasmussen, J., O., P., Mancini, G., Carnino, A., Griffon, M., & Gagnolet, P. (1981). Classification system for reporting events involving human malfunctions. Tech. rep., Riso National Laboratory, Roskilde, Denmark.

Reason, J. (1990). *Human Error*. New York Cambridge University Press.

Reason, J. (2008). The human contribution. *Farnham, UK: Ashgate.*

Reber, A. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology*, *137*, 615–625.

Rothblum, A., Wheal, D., Withington, S., Shappell, S. A., & Wiegmann, D. A. (2002). Improving incident investigation through inclusion of human factors. Paper, United States Department of Transportation Publications and Papers.

Rouse, W., & Rouse, S. (1983). Analysis and classification of human error. *IEEE Transactions on Systems, Man and Cybernetics*, 539–549.

Rumelhart, D., & Norman, D. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science*, *6*, 1–36.

Salame, P., & Baddeley, A. D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behaviour*, *21*, 150–164.

Salame, P., & Baddeley, A. D. (1989). Effects of background music on phonological short-term memory. *Quarterly Journal of Experimental Psychology*, *41A*, 107–122.

Salthouse, T. (1986). Perceptual, cognitive, and motoric aspects of transcription typing. *Psychol. Bull.*, *99*(3), 303–319.

Salthouse, T. A., & Saults, J. S. (1987). Multiple spans in transcription typing. *Journal of Applied Psychology*, *72*(2), 187.

Sawilowsky, S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, *60*, 91–126.

Schmidt, J., Schmorrow, D., & Figlock, R. (2000). Human factors analysis of naval aviation maintenance related mishaps. *Proceedings of the IEA 2000/HFES 2000 Congress*, *3*(7), 75–78.

Scott, L., Rogers, A., Hwang, W.-T., & Zhang, Y. (2006). Effects of critical care nurses' work hours on vigilance and patients' safety. *American Journal of Critical Care*, *15*(1), 30–37.

Shappell, S., & Wiegmann, D. (2001). Applying reason: The human factors analysis and classification system (hfacs). *Human Factors and Aerospace Safety*, 59–86.

Siegel, S., & Castellan, N. (1988). *Nonparametric statistics for the behavioral sciences* (Second edition). McGraw–Hill, Inc.

Smetzer, J., & Cohen, M. (2006). Preventing medication errors related to drug delivery devices. *In M. R Cohen (Ed.) Medication Errors, 2nd Ed.*

Smith, A. (1988). Acute effects of noise exposure: An experimental investigation of the effects of noise and task parameters on cognitive vigilance tasks. *International Archives of Occupational and Environmental Health*, *60*, 307–310.

Soederlund, G. B., Sikstroem, S., Loftesnes, J. M., & Sonuga-Barke, E. J. (2010). The effects of background white noise on memory performance in inattentive school children. *Behavioral and Brain Functions: BBF*, *6*(55).

Soerqvist, P. (2014). On interpretation and task selection in studies on the effects of noise on cognitive performance. *Frontiers in psychology*, *5*.

Soerqvist, P., Noestl, A., & Halin, N. (2012). Disruption of writing processes by the semanticity of background speech. *Scandinavian Journal of Psychology*, *53*, 97–102.

Stanovich, K. (1999). *Who is Rational? Studies of Individual Differences in Reasoning.* Erlbaum.

Stanovich, K., & West, R. (2002). Individual differences in reasoning Implications for the rationality debate. *Behav. Brain Sci.*, *23*, 645–726.

Sternberg, R. J. (1996). *Cognitive Psychology* (2nd edition edition). Harcourt Brace College Publishers.

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin and Review*, *18*(5), 973–978.

Swain, A., & Guttman, H. (1983). Handbook of human reliability analysis with emphasis on nuclear power plant applications. Tech. rep., NUREG/CR-1278 (Washington D.C.).

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.

Szalma, J. L., & Hancock, P. A. (2011). Noise effects on human performance: A meta-analytic synthesis. *Psychological Bulletin*, *137*(4), 682–707.

Tellinghuisen, D. J., & Nowak, E. J. (2003). The inability to ignore auditory distractors as a function of visual task perceptual load. *Perception & Psychophysics*, *65*(5), 817–828.

Thimbleby, H. (2007). User-centered methods are insufficient for safety critical systems. In *Holzinger, A. (Ed.), HCI and Usability for Medicine and Health Care: Third Symposium of the Workgroup Human–Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2007 Lecture Notes in Computer Science pringer, Berlin, Heidelberg, New York*, Vol. 4799, pp. 1–20.

Thimbleby, H. (2011). Don't use 7-segment displays. In *Proceedings BCS Conference on HCI Newcastle, England. Oxford University Press, 2011*, pp. 1–6.

Thimbleby, H., & Cairns, P. (2010). Reducing number entry errors solving a widespread, serious problem. *Journal of the Royal Society, Interface / the Royal Society*, *7*(51), 1429–39.

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.

Vincent, C., Neale, G., & Woloshynowych, M. (2001). Adverse events in british hospitals: preliminary retrospective record review. *British Medical Journal*, *322*, 517–519.

Vincent, C., & Blandford, A. (2011). Maintaining the Standard Challenges in Adopting Best Practice when Designing Medical Devices and Systems. In *Workshop on Interactive Systems in Healthcare (WISH), Workshop at AMIA 2011 Annual Symposium*.

Wallace, D. R., & Kuhn, R. D. (2001). Failure modes in medical device software: An analysis of 15 years of recall data. *International Journal of Reliability, Quality and Safety Engineering*, *08*(04), 351–371.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

Wiegmann, D., & Shappell, S. (2001). Human error analysis of commercial aviation accidents: Application of the human factors analysis and classification system (hfacs). *Aviation, Space, and Environmental Medicine*, *72*, 1006–1016.

Wiklund, M. (1995). *Medical Device and Equipment Design: Usability Engineering and Ergonomics*. Taylor & Francis.

Wiseman, S., Cox, A. L., & Brumby, D. P. (2013). Designing devices with the task in mind: Which numbers are really used in hospitals?. *Human Factors*, *55*(1), 61–74.

Wiseman, S. (2011). Digit distributions - what numbers are really used in hospitals. In *Proceedings of the 4th York Doctoral Symposium*.

Wiseman, S., Cairns, P., & Cox, A. (2011). A Taxonomy of Number Entry Error. *HCI2011: The 25th BCS Conference on Human-Computer Interaction*.

Wiseman, S., Cox, A., Gould, S., & Furniss, D. (2012). Errordiary: Support for teaching human error. *Paper presented at the Contextualised Curriculum Workshop at CHI 2012, Austin, Texas*.

Wu, C., & Liu, Y. (2004). Modeling human transcription typing with queuing network-model human processor (qn-mhp). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48, p. 381.

Wu, X., & Li, Z. (2013). Secondary task method for workload measurement in alarm monitoring and identification tasks. In Rau, P. (Ed.), *Cross-Cultural Design. Methods, Practice, and Case Studies*, Vol. 8023 of *Lecture Notes in Computer Science*, pp. 346–354. Springer Berlin Heidelberg.

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory and Cognition*, *41*, 229–241.

Zhang, J., Johnson, T. R., Patel, V. L., Paige, D. L., & Kubose, T. (2003). Using usability heuristics to evaluate patient safety of medical devices. *Journal of Biomedical Informatics*, *36*(1–2), 23–30.

Zhang, J., Shortliffe, E. H., Patel, V. L., Freed, M., & Remington, R. (2000). The Nature of Human Errors An Emerging Interdisciplinary Perspective. In *Proceedings of the 22th Annual Conference of the Cognitive Science Society (CogSci2000)*.