Automation Bias: Exploring Causal Mechanisms and Potential Mitigation Strategies

Marwa Gadala

Under the Supervision of
Professor Lorenzo Strigini and Professor Peter Ayton

PhD Thesis

City University, London
School of Mathematics, Computer Science & Engineering
May 2017

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AB	automation bias
AIC	Akaike information criterion
CAD	computer-aided detection
CCT	central corneal thickness
CDR	cup-to-disc ratio
CDSS	clinical decision support system
CI	computer interpretation
CT	computed tomography
DR	double reading
DSS	decision support system
ECG	electrocardiogram
EGPS	European Glaucoma Prevention Study
FN	false negative
FP	false positive
IOP	intraocular pressure
JAS	Judge-Advisory System
LOA	level of automation
NICE	National Institute for Health and Care Excellence
OCT	ocular coherence tomography
OHT	ocular hypertension
OHTS	The Ocular Hypertension Treatment Study
QMR	Quick Medical Reference
TM	traceability matrices
TP	true positive
TN	true negative
NHS	National Health System

NNT	number needed to treat
POAG	primary open angle glaucoma
PSD	pattern standard deviation
SEQK	sequential-know condition
SEQNK	sequential-not know condition
SIM	simultaneous condition
VIDI-SM	Variability in Diagnostic Interpretation-Screening Mammography

Acknowledgements

All Praise to God for giving me the strength and will to complete this work.

I am extremely grateful to my first supervisor, Professor Lorenzo Strigini, for his continuous guidance, understanding and insightful advice throughout this PhD research.

I am very thankful to my second supervisor, Professor Peter Ayton, Dr. Eugenio Alberdi and Dr. Andrey Povyakalo for all their valuable discussions and input to this thesis.

I am also very grateful to Dr. Waleed Tantawy and Professor John Lawrenson for their involvement in and contributions to the glaucoma experiment.

I am also very grateful to Professor Paul Taylor for making available the data from his UK clinical trial, which I was fortunate to use in the double-reading simulations.

I am forever indebted to my beloved parents, my loving husband and my dear children for their motivation, patience, and understanding throughout my studies.

This work would not have been possible without you. Thank you all.

Declaration

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes subject to normal conditions of acknowledgement.

Abstract

Automated decision support tools are designed to aid users and improve their performance in certain tasks by providing advice in the form of prompts, alarms, assessments, or recommendations. However, recent evidence suggests that sometimes use of such tools introduces decision errors that are not made without the tool. We refer to this phenomenon as "*automation bias*" (AB), resulting in a broader definition of this term than used by many authors. Sometimes, such automation-induced errors can even result in overall performance (in terms of correct decisions) which is actually worse with the tool than without it. Our literature review reveals an emphasis on mediators affecting automation bias and some mitigation strategies aimed at reducing it. However, there is a lack of research on the cognitive causal explanations for automation bias and on adaptive mitigation strategies that result in tools that adapt to the needs and characteristics of individual users. This thesis aims to address some of these gaps in the literature and focusses on *systems* consisting of a *human* and an *automated tool* which does not replace, but instead supports the human towards making a decision, with the overall responsibility lying with the human user. The overall goal of this thesis is to help reduce the rate of automation bias through a better understanding of its causes and the proposal of innovative, adaptive mitigation strategies. To achieve this, we begin with an extensive literature review on automation bias including examples, mediators, explanations, and mitigations while identifying areas for further research. This review is followed by the presentation of three experiments aimed at reducing the rate of AB in different ways: (1) an experiment to explore causal mechanisms of automation bias, the effect of the mere presence of tool advice before its presentation and the effect of the sequence of tool advice in a glaucoma risk calculator environment, (2) simulations that apply concepts of diversity to human + human systems to improve system performance in a breast cancer double reading programme, and (3) an experiment to study the possibility of improving system performance by tailoring tool setting (sensitivity / specificity combination) for groups of similarly skilled users and cases of similar difficulty level using a spell-checking tool. Results from the glaucoma experiment provide evidence of the effect of the presence of tool advice on user decisions - even before its presentation, as well as evidence of a newly introduced cognitive mechanism (users' strategic change in decision threshold) which may account for some automation bias errors previously observed but unexplained in the literature. Results from the double reading experiment provide evidence of the benefits of diversity in improving system performance. Finally, results from the spell checker experiment provide evidence that groups of similarly skilled users perform better at different tool settings, that the same group of users perform better using a different tool setting in difficult versus easy tasks, and that use of simple models of user behaviour may allow the prediction, among a subset of tool settings for a certain tool, the tool setting that would be most appropriate for each user ability group and class of case difficulty.

General Introduction

In the summer of 1995, *Royal Majesty*, a Panamanian cruise ship went aground when its satellite based navigation system failed to inform the crew, as it was expected to do, that the ship was drifting off-course. Crew members did not monitor other non-automated sources of information that would have warned them that they were in danger, and the accident resulted in $7 million in structural damages and lost revenue (Official Report: Grounding of Cruise Ship). Automated tools like this satellite based navigation system have been introduced into a range of environments: tools for assisting luggage screeners identify suspicious luggage, clinical decision support systems (CDSSs) to aid doctors in diagnosis or screening, and collision-detection alarms installed in modern-day vehicles. Despite clear differences between these tools, they are all designed with the common purpose of aiding the user and improving his/her performance on a certain task in the form of advice given as prompts, alarms, assessments, or recommendations.

At first glance, it would appear that such tools serve only as sources of information (Mosier, 2001), and that the user with their support is at least no worse than the user without their support (Thurfjell, 1998; Lawrence, 2010). Indeed, most research shows overall improved performance with the aid of automated tools. However, several studies prove otherwise (Thurfjell, 1998; Goddard, 2012a; Tsai, 2003; Taylor, 2008a; Berner, 2003; Zheng, 2001; Skitka, 1999). For example, Alberdi et al. (2010) reported that some clinicians, analysing X-ray images of patient's breasts, failed to successfully recall cancer cases when supported by a computer-aided detection tool and which they successfully recalled when reading the X-ray images without the tool. In an aviation context, Mosier et al. (1998) showed that pilots in an aviation monitoring task supported by an automated tool made more errors than pilots who did the task manually. Such empirical evidence suggests that automated tools can create a new class of errors, mostly because even the best tools occasionally give erroneous advice (Goddard, 2012a). The rate of this new class of errors is not insignificant (Friedman, 1999). Moreover, in some contexts, even a single such error can have severe consequences. In fact, the European Joint Aviation Authorities' Future Aviation Safety Team claims that "crew reliance on automation"' is the top safety risk facing future aircrafts (cited by Goddard, 2012b).

Definitions

Parasuraman et al. (1997a) classify inappropriate use of automated tools into three categories: (1) disuse: underutilization of automation such as ignoring prompts, (2) misuse: over-relying on automation even when it's wrong, and (3) abuse: automating functions without regard to their implications. Such inappropriate use of automated tools sometimes appears to make certain errors more likely and/or appears to introduce decision errors that are not made without the tool. We refer to

such automation-induced errors as "*automation bias*" (AB). Many times, AB can even result in overall performance (in terms of correct decisions) which is actually worse with the automated tool than without it. Our definition of AB is broader than that used by many authors. Often times, in the literature, automation bias is defined using only a single one of its potential causes, such as "complacency", "overtrust" or "overreliance". For example, in two of the most recent reviews on automation bias, authors define AB as a form of "overreliance" (Lyell, 2016; Goddard, 2012). One of our main contributions in this thesis is to emphasize and show that AB can be caused by one or a combination of numerous cognitive mechanisms. One such mechanism not previously studied in the literature is the homeostasis mechanism: users' change in internal threshold when responding to cues. Thus, we argue that to reduce the definition of AB to a single one of its potential causes is both simplistic and unfair. Instead, we propose this more encompassing definition that does not attribute AB to a single cause and which groups researchers' similar observations of automation-induced errors and sometimes reduced system performance with tool support.

One of the first studies documenting automation bias was by Mosier et al. (1998) when studying the behaviour of pilots in a simulated flight. Automation bias may be associated with two types of errors: (1) errors of omission, when a user fails to respond to a target because the tool fails to highlight the target, or (2) errors of commission, when a user follows inaccurate advice from the tool (Mosier, 1998). More recently, Alberdi et al. suggest that there are errors apparently due to automated support that are of neither kind, e.g. failing to respond to a target that the tool did highlight, for instance because the user is confused by abundant false positives by the tool. Hence, our more general definition of AB, which includes such errors. Also important is that our holistic definition includes cases where decision errors occur despite correct tool advice, which is a phenomenon noted by researchers, but not included in conventional definitions of AB.

In the spectrum of mitigation strategies aimed at reducing automation bias, there exist two extremes related to the general issue of automation. On one end of the spectrum there is the argument that engineers should not automate a task just because it can be automated (Azar, 1998) and that removing the automated tool from the overall system may be preferred. In general, however, computers are better at responding quickly to control tasks, reasoning deductively, handling many complex tasks simultaneously, and performing repetitive routine tasks (Cummings, 2004). In some applications, automated tools also reduce task time (Southern, 2009), and thus boost efficiency and throughput (Parasuraman, 2012). Sometimes, automated tools act only as a reassurance to the user and may not change the user's actual decision but only increase his/her confidence in the decision (Taylor, 2008a; Hartswood, 1998) (which although often advantageous may sometimes be disadvantageous if the user is reassured in the wrong direction). In other cases, automation offers features not offered by the human. For example, software algorithms designed to detect the quality of a mammogram can

discover common problems such as a patient moving too much, or not pulling in the breast sufficiently, in real-time so that the image can be re-taken in the same setting (Roehrig, 1999). On the other extreme of the spectrum, proponents of automation claim the solution may be to remove the human operator from the system. But humans are better at perceiving patterns, improvising and using flexible procedures, reasoning inductively, and exercising judgement (Cummings, 2004). Thus, although automated tools often outperform humans on certain tasks, the belief that humans are more flexible and better able to cope with unexpected conditions means human supervision and management will continue to be a crucial part of the overall system (Oron-Gilad, 2005). In fact, current recommendations from the American College of Physicians state that ‘‘as errors in computer interpretation are still common, computers should not replace a qualified physician in making patient management decisions’’ (Southern, 2009).

For many applications, the balance likely lies between these two extremes in the form of semi-automated systems composed of a human that uses a decision support tool. Such tools do not serve as a replacement to the human but as an adjunct that support the human (Southern, 2009) by providing alerts, prompts, assessments, or recommendations, but where responsibility lies with the human operator (Parasuraman, 1997a). Our literature review covers tools from this full range of varying automated support. However, the experiments planned as part of this thesis restrict the type of automated tools to those that provide an assessment (glaucoma risk calculator in Chapter 2) and those that provide prompts (computer assisted diagnostic tool in Chapter 3 and spell checking tool in Chapter 4). If we are to minimize the probability of AB errors without diminishing the benefits of such tools, then we need to consider the system as the combined performance of two detectors: the human and the automated tool (Meyer, 2001; Alberdi, 2005). And importantly, not just the behaviour, abilities, and limitations of each detector need to be considered, but also how the two detectors interact (Oron-Gilad, 2005).

To aid in the discussion that follows, we follow Alberdi et al.’s (2009) lead and define a few key terms. In this report, the word *system* is used to refer to both the *human/user* + the automated *tool/Decision Support System (DSS)*. In this overall system, the user sees the *raw data* and looks for certain *cues*/*signs* in the data that indicate the presence of a *target* in a particular *case*. When necessary, the user will then take appropriate *action* such as pull an alarm or recall a patient. To aid the user, the automated tool may provide *assessments, prompts, alarms* or even *recommendations*, but even with the most reliable tools, there will always be the possibility that the tool *misses a target* or gives a *false prompt*. Using this terminology, users may ignore true prompts, and false prompts may be understood as indicating a true target (Halligan, 2011). In the literature, many of the causal terms used such as “complacency”, “overtrust” and “overreliance” are not well-defined and in their respective sections in the literature review we present the diverse range of behaviours or concepts they

are used to describe. These differences in definitions between various authors have made our attempt to use consistent language difficult and not always completely successful.

Thesis Objectives

We note from our review of automation bias in Chapter 1 that although there is abundant literature on automation bias mediators and some literature on mitigation strategies. However, there is a lack of research on the cognitive causal explanations of automation bias and on adaptive mitigation strategies that result in tools that adapt to the needs and characteristics of individual users. The overall goal of this thesis is to help reduce the rate of automation bias through a better understanding of its causes and the proposal of innovative, adaptive mitigation strategies. We achieve this through the design, analysis, and discussion of the three experiments presented in Chapters 2 to 4.

The first experiment is designed using a glaucoma risk calculator tool which provides clinicians with a risk estimate that a given patient will develop glaucoma in the next five years. By manipulating both the knowledge that advice will be presented and the sequence of tool advice, the experiment aims to, for the first time in the literature:

(a) Provide evidence of a newly proposed cognitive mechanism (that users may alter the internal threshold they use to respond to cues) as a possible cause of some automation bias errors
    - To help design more effective mitigation strategies based on a better understanding of the causes of automation bias

(b) Provide evidence that the mere presence of advice before its presentation alters user behaviour
    - To design experiments that evaluate the efficacy of DSSs based on a comparison between the user's post-advice decision to a control group that does not receive advice rather than the user's pre-advice decision

(c) Explore the cognitive processes involved when the sequence of CDSS advice is varied
    - To design more effective protocols for the use of automated tools

(d) Evaluate the effect of the glaucoma risk calculator on the management decisions of ophthalmologists and optometrists when provided with realistic and informative patient cases
    - To assess the benefit of the glaucoma risk calculator in the management of glaucoma suspects

The second experiment is based on a retrospective investigation of the benefits of diversity in double-reading pairs in a breast cancer screening context. Simulations of these pairs aims to, for the first time in the literature:

(a) Provide evidence of the benefits of creating diverse reader pairs compared to random pairs matched based on convenience
    - To motivate tool designers to design tools that focus on the true needs of users and not just tools that exhibit high performance
    - Also, to motivate managers to consider concepts of diversity when forming double-reading pairs

(b) Suggest ways to create effective complementary double-reading pairs
    - To assist managers in deciding how best to form complementary pairs of readers in the specific breast cancer screening context

Although this second experiment is based on a system consisting of two humans rather than a human and tool, we argue that similar benefits could be gained from applying concepts of diversity to systems consisting of a human and an automated tool.

The final experiment is designed using a spell checking tool. This experiment takes a new view on the task of choosing an appropriate tool setting (sensitivity / specificity combination), and unlike previous methods takes into account user ability and case difficulty. By exposing participants to two different tool settings, this experiment aims to, for the first time in the literature:

(a) Provide evidence that each group of similarly skilled users will perform best using a different tool setting
    - To motivate designers to implement tools with variable settings and to motivate managers to choose different settings for each user ability group

(b) Provide evidence that for a single user ability group, the tool setting that results in better system performance for difficult cases may be different from the tool setting that results in better system performance for easy cases
    - To motivate designers to implement tools with variable and easily adjustable settings and to motivate managers to consider ways that case difficulty can be determined and used to change the tool setting

(c) Predict, among a subset of tool settings for a certain tool, the tool setting that would be most appropriate for each user group using modelling of user behaviour
    - To motivate managers and designers to suggest the appropriate tool setting for each user group and for a specific application based on modelling and/or pilot studies

Thesis Organization

This report is divided into four chapters. The first chapter presents a literature review on the topic of automation bias including examples, mediators, explanations, and mitigations while emphasizing areas where further research is required. The second chapter presents an experiment in glaucoma treatment aimed at exploring causal mechanisms of automation bias, the effect of the mere presence of advice before its presentation and the effect of the sequence of tool advice. The third chapter presents work performed on existing data to improve the performance of double-reading pairs in breast cancer screening programs by applying concepts of diversity. The final chapter proposes an experiment to evaluate the effect of varying a tool's sensitivity / specificity combination on the performance of users of varying abilities and cases of varying difficulty using a spell checking tool. Each chapter, after the literature review, aims to present brief background to the specific context, motivation for research, design and methodology, results and an overall discussion and contribution of the work. Each chapter also starts with an executive summary for quick reference.

Chapter 1: Literature Review on Automation Bias

This chapter presents a review of the literature on automation bias. For a holistic view of related ideas, we relied on publications from a number of important disciplines: human computer interaction, psychology and especially papers considering the judge-advisory paradigm (human-human teams where one team member acts as an advisor), aviation (where automation bias was first introduced and remains a pervasive problem), and medicine (where a range of decision support tools and their effects have been evaluated).

Importantly, sometimes studies, especially in healthcare, report errors of automation bias without using the term itself, but such studies can supply valuable insight into our topic. Some of these studies report results or effects on automation *use* rather than automation bias. This more general topic provides information on users' use of automated tools, which may or may not result in automation bias. However, to better understand and predict scenarios and contexts where AB may be more prevalent, it is necessary to consider this more general topic. This is especially important since reliance on tool advice may result in system errors, particularly when tools provide inaccurate advice, which is probably the most prevalent type of automation bias. We are careful to present results of automation bias using this term, but also to present the larger cohort of studies on automation use to add insight to our more specific topic.

This review aims to show that automation bias is a recurring phenomenon in various contexts and then answer a number of general questions. Are there factors that affect susceptibility or resistance to automation bias? What cognitive mechanisms cause automation bias? What interventions can reduce automation bias? What areas would benefit from further research?

This literature review does not restrict its view of automated tools and incorporates a range of different tools. Tool support can range from simple assessments of the raw data to explicit recommendations. For example, risk calculators that provide clinicians with an estimate of the risk of a patient developing a certain disease such as coronary artery disease are emerging in several medical areas. Some tools provide more than an assessment and instead alert the user of a potential target by providing prompts or alarms. For example, a luggage screening tool that highlights suspicious luggage contents. Other tools go a step further by providing an explicit recommendation to the user; for example, to shut down the plane's engine due to a malfunction. The literature review covers tools from this full range of varying automated support. Common to all these tools of varying levels of support is that they are part of a larger system that shares the same following setup:

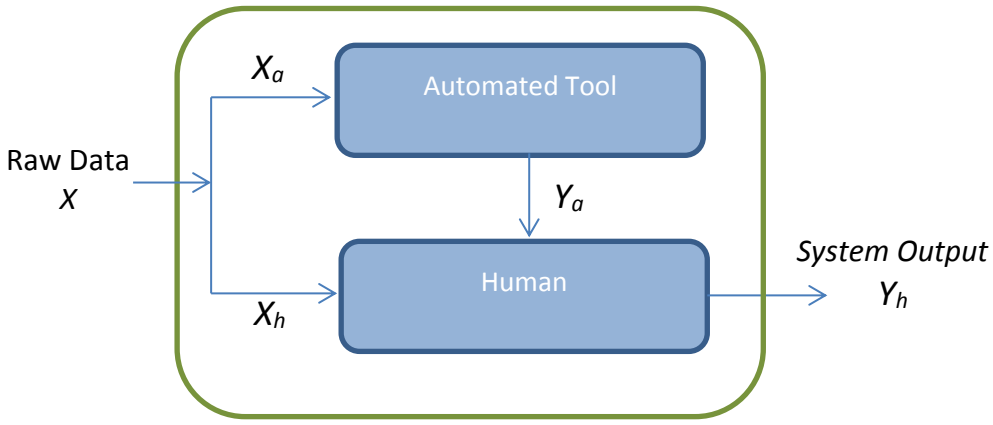


Figure 1-1: Human-Tool System Setup Common to Automated Tools Considered in This Report

Tool and human represent one system (green rectangle), composed of two components. Both components have access to the raw data X. The automated tool searches for a certain set of cues, X_a, and gives the human advice, Y_a. The human weighs this advice with their own set of cues, X_h, to yield the overall system output, Y_h.

From the figure, we see that both the automated tool and the human have access to the raw data X, although each will search for a certain set of cues in the raw data to evaluate whether there is a target. Sometimes these cues may be the same such that X_a and X_h are equal, but not always. The automated tool will process the input and give the user an output Y_a, which may vary in its level of support as previously discussed. The user will process the raw data and weigh it with the advice of the automated tool to give a final decision represented in the system output Y_h. In many systems, Y_h is a binary output representing the decision either to take or not take an action. Together, the tool and human represent one system (green rectangle), composed of two components.

We begin, in Section 1.1 with a few examples of automation bias in different contexts. Section 1.2 and 1.3 cover issues defining automated tools: a tool's setting (sensitivity / specificity combination) and a tool's false prompt rate, respectively. Following that, Section 1.4 presents numerous factors that have been documented in the literature as affecting automation use. These are divided into: personal factors, DSS characteristics, environmental factors, and attitudinal factors. Section 1.5 presents some of the explanations suggested for AB. The ultimate goal from studying automation-related errors is to mitigate and ideally prevent them. Thus, Section 1.6 looks at strategies proposed in the literature to reduce the rate of AB. Following these mitigation strategies is a short discussion, in Section 1.7, of important factors to consider when designing experiments in automation bias. Finally, Sections 1.8 identifies areas only briefly covered in the literature and which represent opportunities for future research and which we have chosen to address in our work presented in subsequent chapters.

1.1 Examples of Automation Bias in Different Contexts

Most research shows overall improved system performance with the use of automated tools, even when the tool provides inaccurate advice (Goddard, 2012a). However, there are numerous studies showing a decrease in overall performance, or in certain cases or by certain users (Goddard, 2012a). Performance degradation is often a result of incorrect tool advice (Parasuraman, 2010); however, there are instances in the literature where authors note degraded performance on certain cases, despite correct tool advice (Berner, 2003; Taylor, 2008a; Zheng, 2001).

Besides cases of automation bias manifested in omission and commission errors, the literature also repeatedly reports examples where overall system performance is lower than that of the tool alone (Cuddeback, 2011; Taylor, 2008b), or at least not superior to the more sensitive system component alone (Marten, 2004). For example, Taylor et al. (2008b) noted that average polyp detection was 51% despite the tool correctly prompting 89.5% of lesions, and Marten et al. (2004) noted that the performance of the inexperienced reader could be significantly improved with the tool, but that this difference was not significantly different from the performance of the tool alone.

Another observation is when users do not reap the full benefit of the tool as in Alberdi et al.'s study (2014) where users' overall sensitivity did not change after using the CAD tool, although if the tool's prompts were used appropriately, users' sensitivity could have increased by up to 10%. Taylor (2014) also noted that radiologists' decision-making generally failed to show significant improvements in cancer detection despite current impressive cancer detection tool capabilities. Such results emphasize that excellent tool performance alone does not necessarily translate into excellent system performance due to the underlining complex interactions between the user and the tool. The diagram in Figure 1-2 is a visual representation of the phenomena often observed in studies showing automation bias.

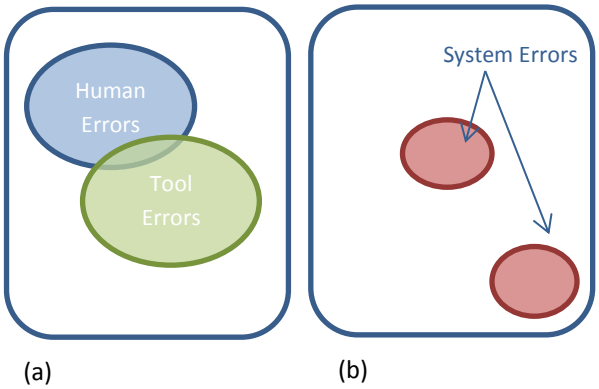


Figure 1-2: Visual Depiction of Automation Bias

In the two Venn diagrams, the universe set is the set of situations to which either a person or a human-tool system is called to respond. (a) Human and tool errors sets when the two perform independently. Overlap represents expected system errors when the human uses the tool. (b) System errors when the human uses the tool. Sometimes these system errors are not limited to the overlap in (a) but may also include: errors exclusive to the tool's error set, errors exclusive to the human's error set, and even errors outside both original error sets.

In this section, we focus on a few examples, in various contexts, of degraded human performance with the use of an automated tool. These represent a subset of the available studies, and other studies are presented throughout this chapter in sections where their results are relevant to a more specific discussion.

1.1.1 Examples of Automation Bias in Aviation

AB was first studied in the aviation context and many AB studies are still focussed in this area. One of the most famous of these studies was by Skitka et al. in 1996 who found that, in a lab simulating aviation monitoring tasks, students given unreliable automated advice made more errors than students who did the task manually. Even though students were informed that the tool was imperfect and even when they were provided with other non-automated sources of information that contradicted the tool's advice, they still chose to accept the tool's incorrect advice (Skitka, 1999). In a later study, pilots performed a similar amount of automation-related errors as the students in Skitka et al.'s experiment (Mosier, 1998). In a third follow-up study, researchers found that automation bias also persisted in the context of an aircraft cockpit and in teams of professional pilots (Mosier, 2001).

1.1.2 Examples of Automation Bias in Medicine

The other field where AB has been commonly studied is medicine in the use of several types of DSSs: computer interpretation in ECG, computer-aided detection (CAD) tools in mammography, CT colonography, and x-rays as well as diagnostic systems such as Quick Medical Reference (QMR) and ILIAD.

As reported by Goddard (2012), results in medical applications show mixed effects on overall system sensitivity and specificity. Some studies report a decrease in both sensitivity and specificity (Alberdi, 2004; Zheng, 2001), increased sensitivity and decreased specificity (Lawrence, 2010; Halligan, 2011), increased specificity with no effect on sensitivity (cited by Goddard, 2012b), and even decreased specificity with no effect on sensitivity (Petrick, 2008).

Computer Interpretation (CI) in aiding ECG interpretations is one area where several studies quoting AB have been published. In a study by Tsai et al., subjects correctly interpreted 48.9% of findings without CI, which increased to 55.4% with CI. This suggests that CI had an overall positive influence. However, when only cases of incorrect advice were included in the analysis, accuracy dropped from 57% to 48% (Tsai, 2003). This drop was masked by a higher benefit due to cases with correct CI (Tsai, 2003). In Southern et al.'s study (2009), the presence or absence of erroneous CIs did not significantly affect residents' ECG interpretations, but did affect their management recommendations as they showed a higher probability of recommending revascularization than those without CI.

While systems such as CAD are intended to aid in the detection - not classification - of cases, there are systems explicitly used for diagnostic purposes. For example, Ramnarayan et al. (2006) examined the effect of a tool that provided diagnostic advice in an acute paediatric setting and found that unsafe workups significantly decreased following tool consultation. Another recent study implementing a diagnostic tool is by Goddard et al. (2014), where practitioners were asked to write a prescription for 20 patient scenarios before and after receiving automated advice. Although there was an overall 8% improvement after seeing the tool's advice, authors recorded a 5% rate of negative consultations (where the correct diagnosis was there before using the DSS but not after) (Goddard, 2014). Earlier, Madhavan et al. (2007) examined physicians' diagnoses before and after using QMR, a DSS that generates an ordered list of diagnostic hypotheses and assigns a probability to each. Physicians were right 39.5% of the time versus 45.4% with the DSS. However, a deeper analysis of the data found positive consultations (where the correct diagnosis was not there before using the DSS but was there after) in 12% of the cases and negative

consultations in 6% of the cases – thus resulting in the 6% net gain in performance observed in the overall analysis (Madhavan, 2007). A study by Berner et al. also used QMR and found that the proportion of cases with correct diagnoses both prior to, and after using the CDSS, was 55%. Again, this overall analysis masked the real effect of the CDSS since there were 20 positive consultations and 21 negative consultations (8%) – where almost all of the latter were a result of the tool failing to display the correct diagnosis (Berner, 2003). Two studies also found that even physicians' electronic searches on sources such as PubMed, Medline, and Google can affect their answers to a set of standardized clinical questions. McKibbon and Fridsman recorded an 11% rate of negative consultations and Westbrook et al. recorded a 7% rate after electronic searches (cited by Parasuraman, 2010).

These studies help highlight the importance of looking not just at the overall effect of a DSS, but where reported, the type of decision changes in terms of positive and negative consultations. The rate of negative consultations in the CDSS literature is not insignificant. In fact, a recent systematic review showed that it can range from 6 to 11%, and that when in error, a CDSS can increase the risk of an incorrect decision by 26% (Goddard, 2014). Similar to these findings is one of the largest studies looking at the effect of CAD on radiologists' decisions when examining breast mammograms (Taylor, 2004). The first time the results were published from this large experiment, no statistically significant impact of CAD was reported (Taylor, 2004). Further analysis of the data from this study revealed that CAD improved the performance of some users and degraded the performance of others.

More specifically, researchers at City University London applied logistic non-linear regression to the data while focussing on the interactions between correctness of tool advice, difficulty of individual cases, and discriminating ability of users and found that "CAD tended to make cancers that were relatively 'easy' easier, and cancers which were relatively difficult even more difficult" (Alberdi, 2010). Interestingly, findings showed differences of the tool's effect depending on user discriminating ability and case difficulty. For the less discriminating readers (44 of the 50 readers), CAD increased their sensitivity by 0.016 for the relatively easy, mostly CAD-detected cancers. However, for the most discriminating users (6 of the 50 readers), CAD decreased their sensitivity by 0.145 for the relatively difficult cancers (Povyakalo, 2013).

Alberdi et al. also used these data to conduct a lesion-level rather than case-level analysis where not just the overall recall/no recall decision was analysed, but the individual features marked by the radiologists on each mammogram (Alberdi, 2008). The results corroborated the previous regression analysis results as they showed that the probability that a less effective reader missed a relatively easy lesion was reduced by 10% with CAD, but the probability of a more sensitive

reader missing a relatively difficult lesion increased with CAD (Alberdi, 2008). This lesion-level analysis also showed that "false prompts made non-cancer features more likely to be classified as cancer" when using CAD and that this effect "outweighed that of true prompts increasing the probability of readers correctly classifying cancer features" (Alberdi, 2008).

The original study by Taylor et al. (2004), which was further analysed by Alberdi et al., provided ample evidence for assessing CAD on cases where prompts were correctly placed, but not necessarily where CAD had failed. Alberdi et al. (2004, 2005) addressed this in two follow up studies (Study 1 with CAD and Study 2 without CAD) with significantly more cases where CAD provided incorrect advice both in the form of marking the incorrect feature on a cancer case or not prompting a cancer case at all. Overall, with CAD, users had an overall sensitivity of 61% compared to 73% without CAD. Moreover, the cases where CAD provided no mark at all showed an especially remarkable difference as sensitivity dropped from 54% without CAD to 33% with CAD (Alberdi, 2004).

1.1.3 Examples of Automation Bias in Other Fields

Outside both the aviation and medical contexts there are many signal detection tasks such as those sited in (Dzindolet, 1999; Meyer, 2001; Van Dongen, 2013) that document cases of AB. Meyer et al. (2001) noted that "with valid warnings participants reached a maximal sensitivity that was lower than the sensitivity of the warning[s]" alone, although higher than the sensitivity without any warnings. This suggests that although system performance increased with the tool, users were not reaping the tool's full benefits. In the software context, automated techniques have been developed to generate traceability matrices (TMs) between pairs of textual software engineering artefacts (such as between requirements and design). In the semi-automated scenario where the analyst is aided by an automated tool, researchers found that analysts tended to degrade high accuracy candidate TMs much more often than improve them (Cuddeback, 2011).

The fact that overall system sensitivity may be lower than that of the tool alone is interesting and has been studied by Wiegmann et al. (2002). When using an automated tool, users may adopt a maximization strategy meaning they always agree with the tool, or a probability-matching strategy meaning they agree with the tool at a rate equal to the tool's overall reliability (Wiegmann, 2002). Probability matching often yields a lower accuracy than does always agreeing with a highly reliable tool (Wiegmann, 2002). This can be seen in the example given by Wiegmann et al.: if a tool were 80% reliable and an operator agreed with it 80 out of 100 times, the overall system accuracy would be 68% versus 80% if they always agreed with the tool. Wiegmann et al. found

that participants periodically disagreed with the tool, "even though aided diagnosis was more accurate than unaided diagnosis and no other additional information was available upon which to base a diagnosis". They also found that fewer users adopted the maximization strategy even though, over several trials, this would have increased their performance. Other researchers have corroborated these results by finding that probability matching rather than maximizing is often the strategy employed by users (cited by Wiegmann, 2002). A recent study took these results further by exploring the dependence of user response strategies on workload, the availability to verify tool advice, and the cost associated with such verification (Manzey, 2014).

1.2 Setting the Threshold of an Automated Tool

The automated tools discussed in the examples in Section 1.1 are designed to make a positive or negative decision about the occurrence of a given target in a systematic way. This requires tool designers to select a threshold such that values above the threshold lead to a positive decision and those below the threshold lead to a negative decision (Swets, 1992). For any given threshold, four decision outcomes are possible as depicted in the table below:

Table 1-1: Decision Outcomes of a Tool

	Target Present	Target Absent
Alarm/Prompt	True Positive (TP)	False Positive (FP)
No Alarm/No Prompt	False Negative (FN)	True Negative (TN)

A fundamental point is that a trade-off exists among the probabilities of the four outcomes, and where the threshold is set determines the balance among these probabilities. For example, a lenient threshold will result in a large number of true positives and false positives, and a relatively small number of true negatives and false negatives (Swets, 1992). Using slightly different terms, there exists a trade-off between a tool's sensitivity (proportion of targets correctly identified) and its specificity (proportion of non-targets correctly identified as non-targets).

Where should a tool's threshold be set for optimal system performance? Swets (1992) defines The Optimal Criterion which takes into account both the prior probabilities of the two alternative conditions to be distinguished (*P(pos)* and *P(neg)*) and the benefits (B_{TP} and B_{TN}) and costs (C_{FP} and C_{FN}) of correct and incorrect decisions as shown in equation (1) below:

$$S_{opt} = \frac{P(neg)}{P(pos)} \times \frac{(B_{TN} - C_{FP})}{(B_{TP} - C_{FN})} \qquad \text{(Swets, 1992)} \qquad (1)$$

Estimates of the probabilities, benefits, and costs, as required to calculate the optimal threshold defined above, may, of course, be very difficult to make, especially when linked to human life (Swets, 1992). An alternative is to consider only true positives, without any regard to false positives (Swets, 1992). Such a lenient threshold derives from the engineering fail-safe approach and can be seen for example in devices designed to detect malfunctions in an airplane. A second approach is to primarily consider true positives and to consider false-positive decisions only qualitatively (Swets, 1992). A third is to consider false positives along with true positives, but to worry about false positives only if they exceed a certain limit (Swets, 1992). A fourth is a more symmetrical approach as it tends to produce a similar number of false negatives and false positives. This may be desirable when there are no strong differences between the benefits and costs of the outcomes or the prior probabilities of the conditions' occurrence (Swets, 1992).

1.2.1 Considerations Beyond Tool Sensitivity and Specificity When Setting the Threshold of an Automated Tool

The approaches used to choose a tool's threshold often focus on the sensitivity and specificity of the tool. But besides a high true positive rate and a low false positive rate, it is also important to consider the posterior true alarm probability, which is the probability that there is a target given that the tool prompts the user (Parasuraman, 1997b). Otherwise the user will be annoyed by the system and likely learn to distrust and disuse it (Parasuraman, 1997b). This same concept has recently resurfaced in the literature in Manzey et al.'s study (2014) where users' response strategies were studied under various tool posterior true alarm probabilities.

Using Bayes' theorem, if the base rate is very low (like in many systems), then even with a high true positive rate and a low false positive rate, the posterior probability can be very low (Parasuraman, 1997b). For example, consider the values in Parasuraman et al.'s example. For a very reliable tool with TP=0.99 and FP=0.01, if the a priori probability p=0.001, then only 1 in 11 alarms will represent a true alarm. If the threshold is set to be more lenient so that the TP=0.999, then when p=.0001, only 1 in 600 alarms will represent a true alarm. Such a sensitive tool would be crying wolf too often in certain applications. Thus, as Parasuraman et al. showed in their paper, setting the threshold for a very high TP rate and a very low FP rate may not be enough. Instead, considering the a priori probabilities and the posterior true alarm probability is also important.

1.2.2 Lower Limit of Tool Reliability

How unreliable can a tool be before neither benefits nor costs accrue? And what contextual factors influence such a threshold? If such a threshold exists, it could serve as a design guideline much

like the magic number 7 which represents the limits of working memory (Wickens, 2007). One study reported that information that was 71% accurate remained acceptable and useful to drivers, but that when information accuracy dropped to 43%, driver performance and tolerance of errors suffered. Wickens et al. (2007) set the threshold at 70% using a regression analysis on data from 20 different studies with varying levels of unreliability. Madhavan et al. (2007) also set the optimal threshold at 70% reliability before performance degrades. But these same authors also reported that when a DSS was only 41% reliable, users still showed a 6% gain in performance. These apparent discrepancies may be due to contextual factors (Wickens, 2007). Furthermore, many of these studies did not discriminate between degraded performance due to false negatives versus degradation due to false positives (Wickens, 2007). In truth, no single "lower limit of automation reliability below which automation complacency [does] not occur" has been found and this is an issue that requires further investigation (Parasuraman, 2010).

Other researchers have chosen to define the optimal threshold in terms of the tool's false positives. One study claimed that CAD schemes for nodule detection would be useful if the FP rate could be reduced to approximately one FP per radiograph and have a sensitivity of 75% (cited by Kobayashi, 1996). Hutt claimed that if the false positive rate was more than 1.5 times the true positive rate, the system would cease to be effective (cited by Hartswood, 1996). However, this was not a clinical study and was biased to true positive cases (Hartswood, 1996). Hartswood et al. (1996) built on this last result, and in a more realistic setting concluded that radiologists' tolerance level for FPs was higher than what was previously reported by Hutt. However, Hartswood et al.'s study did not contain enough targets in the data set to allow for the study of prompt rates on system sensitivity (only on recall rate), contained only four readers, and relied heavily on subjective assessments of the users.

1.2.3 System Performance at Varying Tool Thresholds

A group of interesting studies have evaluated the efficacy of automated tools under various tool settings. For example, in Zheng et al.'s study (2001), radiologists interpreted the same mammograms in the no CAD condition and then under five separate tool settings (combination of two sensitivities (90% and 50%) and two false positive rates (0.5 and 2.0 per image)). Using ROC analysis, the authors found that at 90% sensitivity and an average rate of 0.5 false-positives per image significantly improved system performance, but that the poor schemes adversely affected system performance (Zheng, 2001). In fact, at 50% sensitivity, "the average number of missed abnormalities in noncued areas increased significantly" (Zheng, 2001). Moreover, approximately 30% of these missed lesions were detected by the same radiologists without CAD (Zheng, 2001). It is important to note that in this study, authors maintained cuing rates for only a short duration,

recruited radiologists all with at least 3 years of experience (perhaps explaining why trends were consistent for all users), included only difficult cases, and had radiologists view cued images first.

In another study, also in mammography, authors subjected readers to varying FP tool rates but in a matched pair setup to observe changes in their recall rate and in their subjective assessment of the tool (Hartswood, 1996). They found no increase in recall rate due to higher FP tool rates, and the majority of readers showed a more positive assessment of the tool with increasing prompt rates (Hartswood, 1996). However, as mentioned in the previous section, this study by Hartswood et al. had several limitations.

In a third study, cases were read at different levels of CAD sensitivity and specificity, and the results were compared to those who read without CAD (Philpotts, 2009). Again, the high accuracy cuing was found to increase system performance, but lower tool reliability significantly decreased system performance to the extent that sometimes users would have been better off without CAD advice (Philpotts, 2009).

1.3 Tool False Prompts and their Effects

In even the most accurate DSSs, false prompts are inevitable (Parasuraman, 1997b). Often times, in order to avoid a very high cost of a miss (like the collision of two airplanes), the rate of these false prompts is high (Parasuraman, 1997a). Some researchers claim that false prompts can often be quickly dismissed by a user, and that users learn to habitually dismiss prompts of a certain pattern or frequency (Hartswood, 1998). For example, Taylor et al. (2008a) found no correlation between an increasing number of CAD FP marks and correct classification, and that as many as 25 CAD FP marks in a CT colonography data set did not affect radiologist specificity. Lawrence et al. (2010) also found that in CT colonography images, nearly 50% of marks could be easily dismissed by the user because they identified stool, the rectal catheter or a location outside the colon. In fact, the study by Hartswood et al. (1996) previously presented in Section 1.2, found that responses to a Likert questionnaire after using CAD at varying prompt rates showed more positive assessment of the tool with increasing prompt rates. On the other hand, other experiments do not support this expectation that users can readily discard most false positives, or that they do not find them distracting (Zheng, 2001; Alberdi, 2014). For example, Alberdi et al. (2008) found that falsely placed prompts significantly raised the probability by 12.3% that the prompted areas were marked as malignant compared with the unaided condition.

1.3.1 Upper Tool False-Positive Rate and Type of False Prompts

There may be several reasons for these seemingly conflicting results on the effects of tool FPs on system performance and user reactions: (1) it may be a context-specific problem, (2) there may be an upper limit of false positives after which system performance degrades, or (3) it may depend on the type of prompt and case difficulty (Zheng, 2001). As discussed in Section 1.2.2, Hutt and Hartswood et al. (1996) addressed this second question regarding an upper limit of false positives. However, more research is needed to study the effect of tool FPs on human FPs and *FNs* for *varying* user abilities and case difficulties.

Several researchers have also addressed the third question regarding the type of false prompts presented and how this may cause these seemingly conflicting user responses to false prompts. Taylor et al. (2008a) states that the effectiveness of a tool does not only depend on the *number* of FPs, but also, more importantly, on the *type* of FPs. More specifically, not all false prompts need be distracting; prompting candidate features can be useful in aiding a user's understanding of system rationale and capabilities, can be helpful in positively confirming system behaviour, and can be valuable in improving users' capacity for reflection and self-awareness (Hartswood, 1996; Hartswood, 1997). Radiologists have been quoted saying they prefer to have a prompt rather than guess why the tool did not prompt a certain feature, drawing analogies between the prompting of benign features and their colleagues' useful annotations (Hartswood, 1997). They have also been quoted saying they prefer a tool more sensitive (and less specific) to themselves but that did not prompt obviously benign features (Hartswood, 1996). Thus, even if there is no target, a prompt may sometimes be useful. A good example that illustrates this is provided by Parasuraman et al. (1997b). An ideal collision detection tool might alarm a user in collision-*possible* conditions even though he/she would probably not crash because if it never alarmed except in the case of an accident, then, when it did alarm, the driver would be surprised, disoriented and would likely not react appropriately (Parasuraman, 1997b).

1.3.2 Effect of Tool False Positive Rate on System Sensitivity

It is expected that a low specificity tool will decrease the overall specificity of the system, although at least one study showed users were more specific after using the tool (Alberdi, 2004). Most studies looking at the effect of false prompts were most interested in the effect of FPs on the specificity of the system. Importantly, the effect of FPs on system sensitivity was often ignored in the results of these studies or ignored in the study design itself (by including a low target rate in the data set). Of the few studies that did mention an effect, it is evident that FPs can have an effect on system sensitivity. Zheng et al. (2001) noted that increasing the FP rate from 0.5 to 2.0 per

image increased the number of missed abnormalities in noncued areas. Khoo et al. (2005) calculated that "with a false prompt rate of 1.59 per case for 6050 healthy women and a sensitivity of 84% for 62 cancers, readers will have to dismiss 180 false prompts for every true prompt" and suggested that such a rate may explain why readers sometimes ignore correct tool prompts.

Alberdi et al. (2010) also posited that when a large proportion of prompts are false, the absence of a prompt could become more informative than its presence. In turn, users' reliance on the absence of a prompt to indicate absence of a target could affect sensitivity and authors use this concept to explain some of the omission errors observed in their studies (Alberdi, 2010). Researchers highlight important scenarios where FPs can have significant effects on system sensitivity: when a lesion is surrounded by false prompts (Hartswood, 1998), when a target is in a location where false prompts habitually occur (ex, around the nipple in a mammogram), when excessive FPs distract the user from a target (Balleyguier, 2005), and when a true prompt's value is diminished when surrounded by obvious false prompts (Alberdi, 2010).

1.3.3 Indirect Effects of Tool False Prompts

Besides having an effect on system performance in terms of specificity and sensitivity, false prompts may also have indirect effects on the user such as increased workflow, increased cognitive load (Alberdi, 2010), and increased mistrust in the tool (Hartswood, 1996; Parasuraman, 1997b; Lawrence, 2010). Several studies also document an increase in interpretation time associated with an increase in FP rates (Meyer, 2001; Taylor, 2008a; Lawrence, 2010). This is probably inevitable since time is required to analyse each prompt (Hartswood, 2003). This need to account for the prompt exists even when the prompt is dismissed by the user because as Hartswood et al. (2003) note, users try to understand what caused each prompt. In fact, excessive marks can cause cognitive lockup (the sequential performance of tasks that should be done simultaneously) in the user (Moray, 2000). Users know this, and dangerously, sometimes "economise on the effort needed to dismiss a false positive" by anticipating where FPs might occur and by judging "the value of a prompt based on information in the prompt" rather than cues in the raw data (Hartswood, 1998). Furthermore, users' attention may be distracted towards unimportant areas prompted by the tool leaving them with less time to search for real targets (Alberdi, 2010). Such indirect effects are rarely included as a cost of these prompts (Parasuraman, 1997a).

1.4 Factors Affecting Automation Use

Several studies have developed models of factors that may affect how a DSS is used and how such factors may interact. For example, The Theory of Technology Dominance by Arnold et al. "predicts factors which lead to susceptibility to dominance by technology" and identifies DSS and task experience, task complexity and cognitive fit as key factors (cited by Goddard, 2012b). Principles of this theory were recently augmented with findings from the literature to create a model where confidence, trust, task experience, DSS experience, task complexity, time pressure, and cognitive fit are all connected to reliance or lack of reliance on DSSs (Goddard, 2010). Researchers also developed a Framework for Automation Use that groups factors involved in DSS reliance into cognitive, social, and motivational factors (cited by Van Dongen, 2013). One of the most recent frameworks for automation reliance was presented by Van Dongen et al. in 2013 and encompasses task environment and psychological processes, the latter which is divided into motivational, heuristic, and systematic factors.

In this section we present some of these results and other more specific findings that relate factors that affect automation use, and where available, affect automation bias. The more general topic of automation *use* provides information on users' use of automated tools, which may or may not result in automation bias. However, to better understand and predict scenarios and contexts where AB may be more prevalent, it is necessary to consider this more general topic. This is especially important since reliance on tool advice may result in system errors, particularly when tools provide inaccurate advice, which is probably the most prevalent type of automation bias. We are careful to present results of automation bias using this term, but also to present the larger cohort of studies on automation use to add insight to our more specific topic. The factors identified in this section are categorised into four broad areas: (1) personal, (2) environmental, (3) DSS, and (4) attitudinal factors and are summarized in the diagram on the next page.

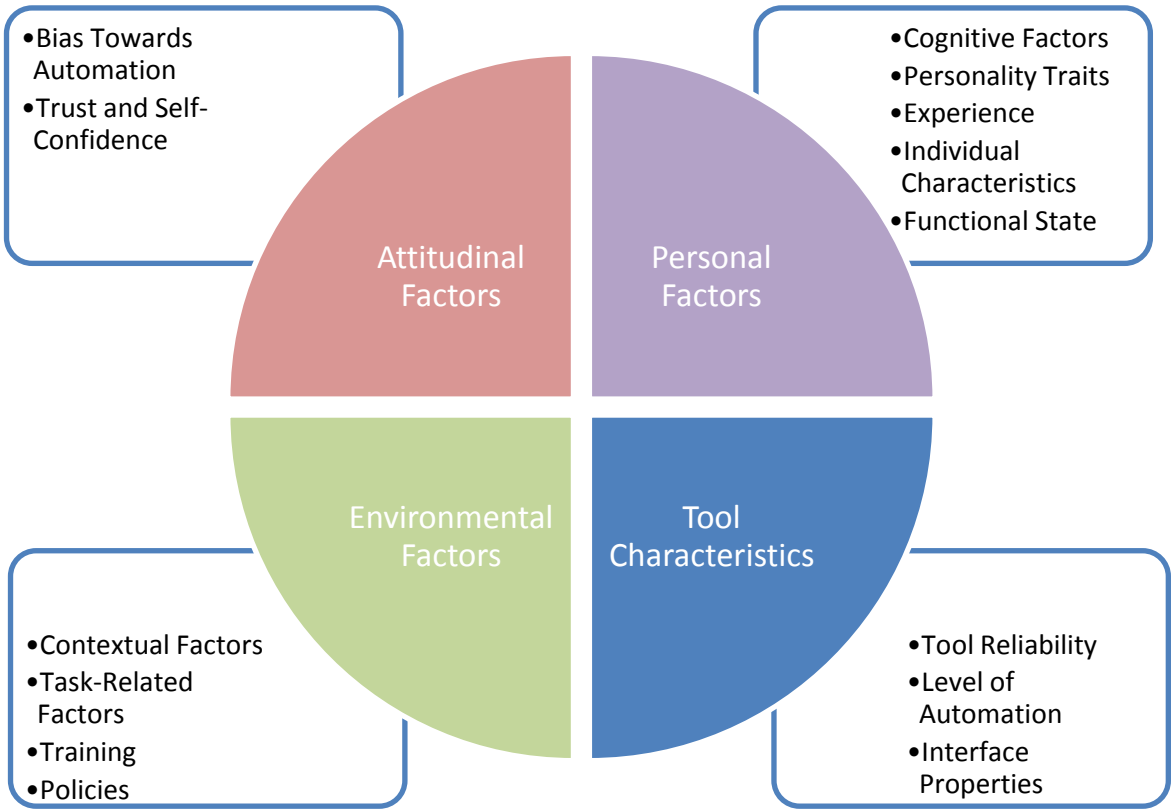


Figure 1-3: Summary of Factors Affecting Automation Use

1.4.1 Personal Factors Affecting Automation Use

The user, a key component of the overall system, naturally determines use of automated advice (Parasuraman, 2010). As suggested by Parasuraman (1997b), the relation between individual characteristics and interaction with automation is complex and requires further research. Individual differences in this category of factors include: (1) cognitive factors, (2) personality traits, (3) experience, (4) individual characteristics, and the (5) functional state of the user.

1.4.1.1 The Effect of User Cognition on Automation Use

Cognitive style refers to "the way individuals think, perceive and remember information" (Goddard, 2012b), need for cognition refers to people's "tendency to enjoy and engage in effortful thought" (Oron-Gilad, 2005), and cognitive capacity refers to "the amount of information that a person can retain and process at any particular time" (Goddard, 2012b). All have been cited to affect automation use. For example, people with a low need for cognition may use simple heuristics more often and thus be more prone to AB (Oron-Gilad, 2005). As another example, users with innovative cognitive styles may "perceive a new technology as useful and easy to use" compared to those with adaptive cognitive styles (Goddard, 2012b).

1.4.1.2 The Effect of Users' Personality Traits on Automation Use

Personality traits may also contribute to individual differences in automation use. For example, Ho et al. (2005) cited older users' risk-aversion character as a factor in determining reliance in their own abilities versus the tool's advice. Prinzel et al. also found a negative correlation between automation complacency and self-efficacy (cited by Parasuraman, 2010). On the other hand, Singh et al. found no relationship between automation complacency and differences in extraversion-introversion personalities (cited by Parasuraman, 2010). Prinzel et al. also found no relationship with boredom proneness or absentmindedness (cited by Parasuraman, 2010). Overall, strong associations between personality traits and automation complacency have not been found, but it is important to note that the small size of the current database does not warrant decisive conclusions (Parasuraman, 2010).

1.4.1.3 The Effect of User Experience on Automation Use

Unlike personality traits, experience is a factor with ample evidence associating it with automation use. It is important to distinguish between effects stemming from insufficient experience with the tool or from inadequate experience in the task itself (Parasuraman, 2010).

Thus, experience is often classified into DSS experience and task-specific experience. Logically, longer exposure to a DSS gives the user a richer context to better judge its advice (Alberdi, 2010). On the other hand, DSS experience may lead to desensitization and habituation, especially for highly reliable tools (Mosier, 1998). As Bailey and Scerbo noted, specific DSS experience decreased users' monitoring (cited by Goddard, 2012a), but other studies found no association between specific DSS experience and decision switching (Goddard, 2014).

Studies that compare performance of users of varying task experience with and without DSSs are relatively common. Some researchers found that AB occurs more often with task inexperienced users (Taylor, 2008a; Tsai, 2003; Goddard, 2012a), but others also showed that it can occur with more experienced users (Balleyguier , 2005; Galletta, 2005, Alberdi, 2014). In most of these studies, a common trend is evident: the gain in performance from using the DSS is greater for the less experienced user than for the more experienced user (Balleyguier, 2005; Madhavan, 2007, Petrick, 2008; Kobayashi, 1996; Taylor, 2008b; Alberdi, 2008), and sometimes the gain in performance is limited only to the inexperienced users (Walsham, 2008).

The rate of change in a user's decision has also been linked to the experience level of the user. Dreiseitl et al. (2005) found that less experienced users were slightly but significantly more susceptible to changing their opinions after using a DSS while Goddard noted a significant negative correlation between susceptibility to change and users' task-specific experience (2012, 2014).

As we mentioned, one trend is that DSSs tend to aid less experienced users more than experienced users (Lawrence, 2010; Philpotts, 2009). On the other hand, a second trend is that inexperienced users often commit the highest number of automation-related errors (Goddard, 2012a). At least two explanations for this exist: (1) experienced users rely less on DSS advice (Goddard, 2012b) because they perceive it as less credible after experiencing tool errors (Tseng, 1999) or because they generally have more self-confidence in their decisions, and (2) experienced users have a more complete knowledge base with which to compare DSS advice (Alberdi, 2010). Both increased trust in the tool and/or lack of experience in the task make it harder for inexperienced users to distinguish tool errors. For example, Taylor et al. (2008a) showed that, in general, CAD FP marks were easily dismissed by *trained* radiologists. In Walsham et al.'s study (2008), the most experienced user was not affected by CAD while the less experienced reader increased their subjective confidence in assessing a case as malignant after it was falsely prompted by the tool. Another study in cardiology showed that internists were much less precise than cardiologists in detecting tool errors (Tsai, 2003).

Although lower task experience has been related to an increase in automation-related errors, a few studies have reported slightly different results. Madhavan et al. (2007) found that negative consultations were equally prevalent across levels of experience, and Dreiseitl et al. (2012) found no significant differences in performance between three different experience groups. Balleyguier et al. (2005) and Taylor et al. (2008b) found that the experienced radiologist actually fell into more FPs than the junior radiologist and studies in the aviation context found that trained pilots showed more automation reliance than students (Skitka, 2000; Mosier, 2001). Galletta and colleagues (2005) also showed, using a spelling and grammar checking software, that compared to less experienced users, the performance of more experienced users was more damaged by the software's false negatives.

We propose to explain some of these seemingly contradictory results. In one case, it may seem that the experienced user's performance is degraded but only because they started with perfect sensitivity, such as what occurred in Petrick et al.'s study (2008), and where it has also been argued by Lawrence et al. (2010), that certain tools such as "CAD may add little to overall CT colonographic interpretation given the high baseline sensitivities when state-of-the-art technique[s]" are used by expert readers. More importantly, a possible explanation to the studies reporting a stronger decrease in performance for the experienced users than the less experienced users lies in the type of cases where incorrect tool advice occurred. These could be difficult cases that experienced users were initially uncertain about and which the tool's incorrect advice provided a false reassurance for. The less experienced users, on the other hand, would have already erred in these difficult cases prior to seeing the tool's incorrect advice and thus not been affected by the tool. This is most exemplified by the statistical analyses summarized by Alberdi et al. (2010) where, as we reported in Section 1.1, researchers found that when using CAD, less effective readers showed a greater benefit in performance after using the tool compared to the benefit of the more effective users, but that this was true mostly for the easier cases. On the other hand, for difficult cases, the more skilled users exhibited degraded performance with CAD compared to without CAD. Their results showed that the difficulty of the tool error dictated the effect exhibited by different users and that this effect varied with user ability. This important observation is visually depicted in Figure 1-4 using a plot from Alberdi et al.'s chapter (2010) in "The Handbook of Medical Image Perception and Techniques".

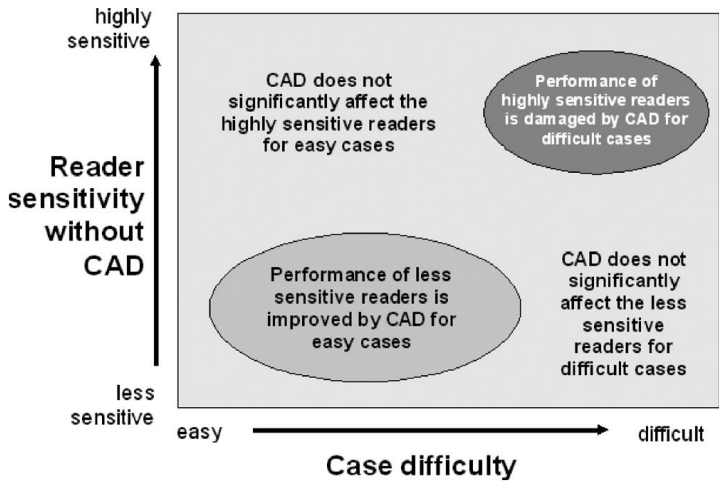


Figure 1-4: Performance Effects of CAD According to User Ability and Case Difficulty (Alberdi, 2010)

Besides the difficulty of the case where the tool erred, differences in the effect of a DSS according to user experience may also be attributed to the unique manner that experienced users use the tool or perform the task. Dreiseitl et al. (2005) used eye-tracking to observe dermatologists of varying experiences diagnose a set of lesion images. Although authors did not find significant differences in the experience groups in terms of sensitivity and specificity, the expert group exhibited visibly lower ratings for all eye measures. They spent a disproportionately lower time in fixations, took a more holistic approach, and did not scan for the presence or absence of features as much as the less experienced radiologists (Dreiseitl, 2005).

1.4.1.4 The Effect of Individual User Characteristics on Automation Use

Besides user experience, another factor that has been noted when observing varying effects of a DSS on different users is users' characteristics. Of notice in this field is the work by Oron-Gilad et al. (2005) which aims to link trait reactivity with attentional indices that predict performance in automated systems. A few of the individual variables the authors list to predict performance in automated environments are: the need for control, locus of control, extraversion, leadership style, need for cognition, and decision-making style (Oron-Gilad, 2005). Recently, McBride et al. (2014) also included complacency potential ("attitudes that influence complacent monitoring behaviour") in their extensive review of variables influencing the management of automation errors.

One would expect that increased accountability (the obligation to accept responsibility for one's actions by justifying them) would result in more vigilant information seeking and avoid

premature closure of decisions and thus help reduce the rate of AB (Skitka, 2000). Skitka et al. (2000) explored this idea in an aviation context, with student participants, and found that students who were instructed to be accountable for overall performance committed significantly less omission and commission errors and also showed more automation verification behaviour. However, an earlier experiment by Mosier et al. (1998) with pilot participants did not find a difference in automation-related errors between the two experimentally manipulated accountability groups, even though experimentally manipulated accountability demands did make pilots double-check automated cues. Instead, Mosier et al. found that, from the questionnaire, those who had a higher sense of *perceived* accountability were less likely to commit omission errors. A recent study also found that users' perceived accountability caused them to be more likely to accept (more reliable but) conflicting advice from the tool - associated with increased mental effort to changes one's initial opinion (Van Dongen, 2013).

Another study by Mosier (2001) looked at the effects of the presence of a second crewmember and found that it was not enough to eliminate, or even significantly reduce, automation bias. When two people are monitoring system events, it would seem to double the chances that they would detect a target, but many other factors complicate this interaction such as individuals' expected competency of a teammate (Mosier, 2001), social loafing (Mosier, 2001), staff hierarchy (Mosier, 2001) and lack of diversity (Alberdi, 2010).

Age differences in complacency have also been reported, "with older adults exhibiting greater automation-related complacency [compared to younger adults], but only under very high workload conditions" (Parasuraman, 2010). Ho et al. (2005) found that when using a medicine management system older users were more likely to trust the aid, were less confident in their performance, were more reliant on the aid, committed more automation related errors, and were less sensitive to automation failures. However, Goddard et al. (2014) found no such association between age and automation reliance. It is important to note that age may encompass effects due to other important influencing factors such as experience and sometimes a higher probability of cognitive overload. With respect to other sources of individual differences, gender differences have not been reported (Parasuraman, 2010).

Finally, in one of the most recent and unprecedented studies in AB, Parasuraman et al. (2012) looked at the effect of certain genes in identifying individuals' susceptibility to automation bias. Results from this study indicated that individuals with low levels of DBH enzyme activity, "which is associated with higher dopamine to norepinephrine levels in cortex", exhibit superior decision making compared to those with high DBH enzyme activity (Parasuraman, 2012). Specifically, the low DBH enzyme activity group, compared to the high DBH enzyme activity

group, was more accurate, was speedier in making decisions, verified inaccurate automated advice at more than twice the rate, and reported lower subjective trust in the tool (Parasuraman, 2012).

1.4.1.5 The Effect of the Functional State of a User on Automation Use

In many alerting systems, the extra layer of detection redundancy aims to reduce the likelihood of missing an event (Thurfjell, 1998; Hartswood, 1998). This can be especially important when compensating for fatigue, or human distractions (Walsham, 2008). Interestingly, although introduced with the goal of combatting fatigue in human operators, the use of some automated tools increases interpretation time as a result of the attention needed to assess tool prompts. "Could the use of [automated tools] actually increase fatigue? It is certainly possible" (Philpotts, 2009).

A recent study considered the functional state of users, and specifically sleep loss and fatigue, on automation use (Reichenbach, 2011). Previous research in "supervisory control tasks suggests that operators suffering from sleep loss tend to apply less cognitively demanding strategies and reduce their information sampling" (cited by Reichenbach, 2011). This would imply that users suffering from sleep loss will likely rely more on the tool and thus increase the rate of automation-related errors. On the other hand, users suffering from sleep loss might adopt a different strategy by which they protect their performance, show increased concern to compensate for the elevated risk from their fatigue and increase their information sampling (Reichenbach, 2011). Authors found support for the latter conjecture as participants suffering from sleep loss were less complacent, more attentive, and verified advice more often (Reichenbach, 2011).

Slightly contradictory results can be found in the Judge-Advisory System literature. A Judge-Advisory System (JAS) refers to human-human teams where one team member acts as an advisor. Research from this area will be referred to often in this chapter as in many ways, human-automation teams function similarly to human-human teams (Madhavan, 2007). From the JAS literature, Danziger et al. found that judges' decisions were negatively influenced by the length of time since a work break: "the percentage of favourable rulings drop[ped] gradually from ≈65% to nearly zero within each decision session and return[ed] abruptly to ≈65% after a break" (cited by Goddard, 2012b).

1.4.2 DSS Characteristics Affecting Automation Use

Just as the user is an integral part of the system, the automated tool itself is too, and evidence shows that certain characteristics of the tool can influence its use. These characteristics can be categorized as: (1) tool reliability, (2) level of automation, and (3) interface properties.

1.4.2.1 The Effect of Tool Reliability on Automation Use

Probably the most important DSS characteristic, even though it is often not reported (Kawamoto, 2005), is the tool's level of reliability. It seems natural that the number of errors committed by the tool will affect the number of automation-related system errors and indirectly affect user trust. Interestingly, several studies reported that decreased system performance, in cases of inaccurate tool advice, were most pronounced for tools with higher reliability (Parasuraman, 2010; Goddard, 2012b). Thus, sometimes, lower levels of tool reliability can paradoxically inspire better system performance due to decreased complacency levels (Goddard, 2012b).

Some researchers criticise experiments examining the rate of AB as suffering from a high error rate - contrary to some real world settings. To address this criticism, Molloy and Parasuraman conducted a study in which the DSS failed only once during the session and found that in the single-task, "most participants detected the automation failure, whether it occurred early or late" (cited by Parasuraman, 2010). However, "under multitask conditions, only about half the participants detected the automation failure, and even a smaller proportion detected the failure if it occurred late than if it occurred early" (cited by Parasuraman, 2010). These results show that automation complacency can occur when using highly reliable tools (Parasuraman, 2010).

Interestingly, it is not just the number of errors, but also the type of errors which can affect automation use. This has been touched upon in Section 1.3.1 in the discussion of types of false prompts and will be discussed again in Section 1.4.4.2 with respect to trust and confidence.

Besides the tool's actual level of reliability, the *perceived* level is also important (Parasuraman, 1997a). Misuse occurs when users overestimate the perceived reliability of a tool, and disuse occurs when users underestimate the perceived reliability of a tool (Dzindolet, 1999). Moreover, as Wiegmann et al. (2002) reported, users' perceived reliability is usually less than the true reliability of the DSS. Often, the perceived reliability is not directly linked to the tool's actual reliability and may instead depend on the user's bias toward automation or superficial factors such as the interface of the tool (Dzindolet, 1999). In fact, increasing the automated

tool's reliability increases automation use *only* if the change increases the perceived reliability of the tool (Dzindolet, 1999).

Also well cited and related to a tool's reliability is the idea of variable versus constant reliability. In one study, Parasuraman et al. (2003) found that individuals with a higher complacency potential, "were particularly poor in monitoring automation under constant than under variable-reliability automation" (cited by Parasuraman, 2010). In fact, the mean detection rate of automation failures showed a dramatic difference in magnitude of 149% between the variable-reliability and the constant-reliability conditions (cited by Parasuraman, 2010). Certainly, how this variation in reliability occurs, would influence the results.

1.4.2.2 The Effect of the Level of Automation on Automation Use

Several studies have shown that the level of automation (LOA) of a tool can affect AB, complacency and situation awareness (Parasuraman, 2010). Level of automation refers to which functions are automated and to what extent. Various LOAs can be introduced in a DSS, "from fully automated where the operator is completely left out of the decision process to minimal levels of automation" where the DSS only provides an assessment of the relevant data (Cummings, 2004). In general, performance impairments in cases of inaccurate advice are most pronounced at higher levels of automation (Parasuraman, 2010). Results by Cummings et al. (2004) corroborate this as they show that although pilots working with a command display performed better than those working with a status display in cases of accurate tool advice, when the computer erred, AB was stronger for command than for status displays. This effect, however, could not be replicated in a follow-up study (cited by Parasuraman, 2010), and one study even showed contradictory results (Kawamoto, 2005). Kawamoto et al. (2005) built a regression model to find CDSS characteristics that determine better system performance. In this review, authors concluded that "provision of recommendations rather than just assessments", "provision of decision support at the time and location of decision making", and "*computer* based decision support" were key factors in predicting better CDSS performance (Kawamoto, 2005). Of 32 systems with all four features 94% significantly improved system performance (Kawamoto, 2005).

This apparent controversy may be explained, as suggested by Cummings et al. (2004), as a context-specific issue. Cummings et al. (2004) suggest that high LOAs are often better for highly reliable tools that require little flexibility in decision-making, while lower LOAs may be useful for redundant tasks or time-critical tasks with changing constraints. Importantly, high

levels of automation may induce complacency, skill degradation, loss of situational awareness, and a lack of accountability by taking too much control away from the user (Cummings, 2004).

1.4.2.3 The Effect of Interface Properties on Automation Use

Besides tool reliability and LOA, the way the tool's advice is presented has also been linked to automation use. Visual design factors such as colours, increased scene realism, reduced screen-detail, a balanced layout, and consistent presentation can affect a user's perceived reliability of the tool, their trust in the tool, and in turn their reliance on its advice (Goddard, 2012b).

The location of tool advice with respect to raw data is also important. Research in human factors shows that decision makers may focus on salient cues and ignore other critical but less obvious information (Parasuraman, 1997a; Skitka, 2000). Automated advice not prominently displayed may lead to failure to heed the tool's advice (Berner, 2003). In one study, researchers found that display prominence increased AB such that incorrect advice was more likely to be followed when presented in a prominent location (Berner, 2003). However, Singh et al. found that moving the automated task from the periphery to the centre of the display did not prevent complacent behaviour (cited by Parasuraman, 2010). This consideration "that users may only attend to the most prominent suggestions" (Berner, 2003) is especially important in the design of tools, such as QMR, that return a long list of suggestions to the user (Kawamoto, 2005).

Meyer et al. (2001) addressed whether the proximity of the warning to the raw data affected users' acceptance of the tool's advice. In the low-proximity condition, the warning was shown in a separate field from the raw data while in the high-proximity condition, the warning was shown as the colour of the continuous stimulus. Results showed no effect of the presentation of the warning when the tool's advice was valid. However, "with non-valid warnings, the participants' sensitivity was higher with separate displays than with integrated displays". Authors concluded that it may be more difficult to ignore inaccurate advice when integrated or positioned in close proximity to the raw data (Meyer, 2001).

Many studies have also cited the importance of providing the DSS's decision rationale and reliability information (such as confidence levels) in influencing appropriate automation use (Parasuraman, 1997a; Hartswood, 1998; Alberdi, 2010; Kawamoto, 2005; de Vries, 2003). Madhavan et al. (2007) found that providing reliability information led to more appropriate reliance. This topic will be covered in more detail in Section 1.6.2.5 when presented as a mitigation strategy.

1.4.3 Environmental Factors Affecting Automation Use

As seen in Section 1.1, AB is not limited to lab environments or to a certain context. Instead it is a phenomenon found in a range of different settings. To better understand and predict scenarios and contexts where AB may be more prevalent, it is necessary to consider the environmental factors that affect how users rely on automated advice. This category of factors encapsulate a list of objective characteristics concerned with: (1) contextual factors describing the general environment a decision is being made in such as other competing tasks and time constraints, (2) the task itself such as task difficulty and nature, (3) training procedures, (4) organizational policies, and (5) sequence of advice.

1.4.3.1 The Effect of Contextual Factors on Automation Use

Complacency is primarily found in multitasking environments (Parasuraman, 2010). This can be easily understood by Parasuraman et al.'s attention model (see Section 1.5.5) because multitasking may cause attention allocation away from the automated task (Parasuraman, 2010). Researchers have also shown that users "faced with more than one task at a time in a complex system are very reluctant to time share control between them and prefer to perform them sequentially rather than simultaneously" (Moray, 2000). This effect is sometimes called "cognitive lockup" or "cognitive tunnel vision" (Moray, 2000).

The physical environment that the user works in can also have an effect on automation use. For example, heating, lighting, ventilation, and distractions can add to environmental stress (Goddard, 2012b). Increased reliance has also been associated with increased time pressure which can be explained by users increased satisficing behaviour and reduced vigilance (Goddard, 2012a). Evidence also shows that higher workloads often increase reliance (Parasuraman, 1997a). These loads can be caused either by the difficulty of a single task (discussed further in the next section), or the concurrent task load (Wickens, 2005). Surprisingly, there is little evidence that people prefer to use automation more under higher workload conditions or when the task is more difficult (Parasuraman, 1997a). Winere found that only a fraction of pilots agreed that "automation reduces workload", and a minority even felt that it increased their workload (cited by Parasuraman, 1997a).

There is disparate evidence on whether use of a DSS increases (Madhavan, 2007; Petrick, 2008; Halligan, 2011), decreases (Southern, 2009), or makes no difference (Taylor, 2009; Kobayashi, 1996) in the time a user spends making a decision. It would seem natural that additional time will be required simply because of the need to assess all tool prompts (Philpotts, 2009). Indeed

this was reported in several studies (Madhavan, 2007; Petrick, 2008; Halligan, 2011). In fact, Madhavan et al. (2007) found that physicians using QMR took 22 minutes to make a decision compared to 8 minutes without the tool. In some contexts, such an increase and its impact on system throughput could be substantial and could affect DSS implementation (Petrick, 2008). Halligan et al. (2011) and Taylor et al. (2008b) found an increase with CAD, but longer times for second-reader versus concurrent reading. This highlights an important point: that protocol and context are likely to affect interpretation time and help explain some of the variance in the results.

Recently, McBride et al. (2014) documented the importance of the cost of verifying tool advice on users' management of automation errors. This cost of verification can be in the form of increased time or resources. In a study by Ezer et al. (2008), reducing the time required to verify a tool's suggestions resulted in better system performance (cited by McBride, 2014). Also, Manzey et al. (2014) found that increasing the cost of verification increased users' reliance, but did not affect their compliance to the same extent.

1.4.3.2 The Effect of Task-Specific Factors on Automation Use

Evidence shows that more complex tasks increase automation reliance (Goddard, 2014). In fact, a first attempt at proving complacency exists failed and later studies attributed this failure to the task being too easy (cited by Parasuraman, 2010). This is related to a user's cognitive capacity because "as task difficulty increases to reach the user's cognitive capacity", aid from external resources increases (Goddard, 2012b). Also, from the JAS literature, judges appear to accept advice more often when dealing with complex compared to easier tasks (Bonaccio, 2006).

Repetitive tasks have also been shown "to elicit habituation effects" which may increase the rate of AB (Goddard, 2012b). Risk and urgency of the task have also been documented as increasing aid reliance (Parasuraman, 1997a). McBride et al. (2014) also mention the consequence of an automation error undetected by the human as an important mediator. For example, although Driestel et al. (2005) found no tendency to follow tool suggestions more willingly if they led to a safer diagnosis (i.e., recalling a patient), Palmer et al. (1991) found that pilots were less likely to detect automation errors that did not influence flight safety (cited by McBride, 2014).

1.4.3.3 The Effect of Training on Automation Use

Besides contextual factors and factors related to the specific task, research has also assessed the effect of training on appropriate tool reliance. Training aims to reduce the effect of automation bias by altering the schema that automated tools are perfect (Dzindloet, 1999), by ensuring users understand when a tool is likely to err (Dzindolet, 1999; Philpotts, 2009), and by explaining, to some extent, how the tool works. Of the most notable studies looking at the effect of training is the study by Bahner et al. (2008) which investigated the effect of exposing participants to rare automation failures. Researchers found that users who were *exposed* to automation failures showed a better cross-checking behaviour than users who were just *informed* about the possibility of such failures. This distinction between experienced and informed training may explain why not all researchers have recorded a benefit from training. For example, Mosier et al. (2001) explored the extent to which AB is reduced "when participants [are] given explicit training about automation bias as a potential problem" and found that training significantly reduced commission but not omission errors. Training will be discussed further in Section 1.6.3.2 when presented as a mitigation strategy.

1.4.3.4 The Effect of Policies on Automation Use

There is considerable attention to the effects of training in the literature, but the effects of policies are much less covered. There are differences in the expectations of managers, operators and designers of how a tool should be used (Parasuraman, 1997a), and when designing a new tool, designers need to consider organizations and managers and their effects. As Mosier et al. (2001) state, "Organizations may be fostering tendencies toward automation bias through their policies and training." For example, in Mosier et al.'s experiment (2001), many pilots committed the fire commission error, but "at least one of the airlines whose pilots were involved in the study trains crews to assume a fire is present whether or not all the indicators are consistent" (Mosier, 2001). In another example, a train crashed when the driver was not allowed to take manual control (Parasuraman, 1997a). Thus, there are examples where policies could help increase automation-related errors or prevent users from taking what they believe is the better course of action (Mosier, 2001). In fact, even "the importance placed on the use of the DSS by [an] organisation" can affect users' acceptance of a tool's advice (Goddard, 2012b).

1.4.3.5 The Effect of the Sequence of Tool Advice on Automation Use

Related to policy making is the proper integration of a DSS into a user's workflow, which is often dictated by policies. As the review by Kawamoto et al. (2005) highlighted, the "automatic provision of decision support as part of clinical workflow", is an essential factor in the success of a CDSS. This was also noted by Hartswood et al. (2003) as they argued that the introduction of computer-based tools into users' work environments should take into account sociality of work and implications for users' work practices. In fact, lack of integration can discourage use of a tool because sometimes when the switch to automation interrupts the workflow or requires extra effort, people choose manual control (Parasuraman, 1997a). Policies, especially certain examples associated with DSS use, will be discussed further in Section 1.6.3.1. Here, we focus on the sequence in which the tool's advice is presented.

Few studies have examined the effect of the sequence of advice, and even fewer have presented differences in results without a discussion of the cognitive effects that could cause these effects. In what follows, we define two modes in which advice from a tool may be incorporated: (1) *simultaneous* mode: the user sees the tool advice and the raw data at the same time, and (2) *sequential* mode: the user sees the raw data first, forms a pre-decision and then sees the tool's advice before giving a final decision.

The study by Kobayashi et al. (1996) is one of the first studies tackling this issue and likely one of the first studies to introduce the sequential mode. The purpose of this study was to evaluate the effect of CAD on radiologists' detection of lung nodules. Based on the findings, system performance was rated, according to the area under the ROC curve, and from worst to best as the: decision without CAD, pre-advice sequential decision, decision with advice presented simultaneously, and finally the post-advice sequential decision.

Surprisingly, authors concluded that both modes were equivalent due to the lack of significant statistical differences in the area under the ROC curve between the post-advice sequential decision and the simultaneous decision. Although overall results didn't reveal significant differences, the authors reported different behaviours in the different modes which could indicate cognitive changes that were not studied. Kobayashi et al. (1996), unlike other studies looking at this issue, provided valuable details on the type of errors committed:

- *Confidence changes:* More cases were affected by more than 30 units on the 0-100 confidence rating scale in the simultaneous mode than in the sequential mode.

- *True Positives:* When CAD correctly prompted a nodule, the net effect of CAD was significantly beneficial in both modes, but less beneficial in the sequential mode (2.9 cases vs 6.3).
- *False Positives:* When CAD falsely prompted a nodule, the average numbers of cases affected detrimentally were more in the simultaneous mode compared to the sequential mode (2.0 cases vs 0.6).
- *False Negatives:* When CAD failed to prompt a nodule, the average numbers of cases affected detrimentally were more in the simultaneous mode than in the sequential mode (1.8 cases vs 0.1).

This added detail suggests that users in the simultaneous mode were helped more by true positives but hurt more by false positives and false negatives. This may indicate that the weight of advice in this mode is greater than in the sequential mode. This may also point to whether asking users to provide a provisional decision before presenting advice affects their utilization of advice and specifically makes them more resistant to it (Bonaccio, 2006).

1.4.3.5.1 Studies from the JAS Literature Looking at the Effect of the Sequence of Advice

A basic dilemma when presented with advice is the amount of weight to place on it relative to the weight placed on the user's initial decision. A user's weighting policy might involve completely ignoring the advice, some adjustment of their pre-decision towards the advice, or complete adoption of the advice (Yaniv, 2004). From the JAS literature, evidence shows that this weighting is a function of the user's knowledge and of the distance of the advice from the user's pre-advice decision (Yaniv, 2004). Specifically, Yaniv's experiment (2004) showed that more knowledgeable users discounted advice significantly more than less knowledgeable users. Moreover, as the distance between their pre-advice opinion and the advice increased, the weighting of the advice decreased. This could be because the more knowledgeable a user is, the more evidence they have to back their own opinion, and thus the more weight they attribute to it (Yaniv, 2004). The less knowledgeable users did not exhibit such an association with distance, "perhaps because they felt they could benefit even from distant advice" (Yaniv, 2004). Bonaccio et al. (2006) showed that the weighting is also related to the post-advice accuracy of the decision.

Another study from the JAS literature compared performance in the sequential and simultaneous reading modes and showed that order manipulation affected judges' final

decision accuracy and confidence (Sniezek, 1995a). When there was no conflict, judges using the sequential mode had a higher final decision accuracy than those who saw advice simultaneously, but judges in both modes still performed better than those in the control group working without advice (Sniezek, 1995a). In fact, when advisors gave conflicting advice, performance in the simultaneous mode became equal to that of the control group; almost as if advice prevented judges from retrieving their own internal knowledge (Sniezek, 1995a). Dangerously, judges in the simulatneous mode were less accurate, but not less confident than those in the sequential mode. Authors concluded the paper by recommending that in order to optimize system performance, users should make an independent decision first. Also from the JAS literature, researchers found that participants' answers matched advisor's answers more often when they received advice first compared to when they made a preliminary decision before seeing the advice, possibly because when advice was presented first, judges' knowledge was mentally less available. (Van Dongen, 2013)

In a final important study in the JAS literature, participants were asked to compare bike attributes and choose the bike that gave the best value for money (Sniezek, 1995b). Advice could be solicited only once, but at any time before a final decision was made. Judges chose to access advice very late (on average, after they had completed over 75% of their own information search) (Sniezek, 1995b). Authors noted that the pre-advice search could be conceptualized as a hypothesis generating phase, involving the search of significant amounts of information relatively quickly to efficiently reduce the number of alternatives (Sniezek, 1995b). On the other hand, post-advice search could be conceptualized as a hypothesis testing phase characterized by less depth of search and more alternative-based search (Sniezek, 1995b). This conceptualization was also mentioned by Gale who posited that radiologists select from a pool of possible hypotheses based on their expectations about a given case, and then each considered hypothesis is confirmed or refuted based on cues gathered when visually inspecting the mammogram (cited by Hartswood, 1999).

What is striking about Sniezek at al.'s study (1995b) is that it did not support the idea that the availability of advice is taken as an indication by users that they can exert less effort than unadvised decision-makers (i.e., ideas of complacency, social loafing, or reduced vigilance which will be discussed in Section 1.5). Instead, in the study, judges outperformed unadvised decision-makers but did not differ significantly from them in terms of information search or effort (Sniezek, 1995b).

1.4.3.5.2 Studies from the Medical Literature Looking at the Effect of the Sequence of Advice

Several medical studies have examined the efficacy of the simultaneous and sequential reading paradigms. In CT colonography, two studies reported that relative to unassisted read, polyp detection was greater in the sequential mode than in the simultaneous mode, although both modes showed an improvement compared to without CAD (Halligan, 2011; Taylor, 2008b). In Taylor et al.'s study (2008b), there was added detail that 40 polyps correctly identified without CAD were later missed by at least one reader in the simultaneous mode. Authors also reported that readers in the simultaneous mode ignored more CAD-prompted polyps. On average, readers committed at least one false-positive in 18.8%, 26.4%, and 27.2% of data sets by using unassisted read, sequential mode, and simultaneous mode, respectively (Taylor, 2008b).

In another medical paper, radiologists read mammograms in six different display conditions (Zheng, 2004). Authors noted that advice in the sequential mode had little effect on sensitivity, and that with the least specific cuing rate, viewing the advice simultaneously significantly reduced performance (Zheng, 2004). Viewing CAD cues during the initial decision also resulted in fewer abnormalities being identified in noncued regions (Zheng, 2004). Other studies reported that the simultaneous mode reduced observer vigilance (Van Dongen, 2013; Taylor, 2008b) and sensitivity (Taylor, 2008b), although Halligan et al. (2011) suggested that users may not be vigilant before they are assisted because they prematurely activate the tool.

1.4.3.5.3 The Sequence of Advice and Automation Bias

In the latest reviews in automation bias, the sequence of advice was not mentioned or recognized either as a mitigation strategy or as a mediator of automation reliance (Lyell, 2016; Goddard, 2012; Parasuraman, 2010). Only a very recent study by Van Dongen et al. (2013) discussed the issue of the sequence of advice in specific relation to automation bias. In a simple pattern recognition task, authors found that the percentage agreement with the tool did not differ between the two modes. However, when the results were linked to subjective measures of trust and self-confidence, authors were able to suggest that users seem to be "cognitively engaged in making reliance decisions" when advice is presented sequentially and that users heuristically rely on the tool when advice is provided simultaneously (Van Dongen, 2013).

Some researchers argue for use of the sequential mode in tool evaluation studies as the two separate settings required in simultaneous mode, or the matching of similar cases is not always ideal or practical (Kobayashi, 1996). The main disadvantage researchers simplistically cite in using sequential mode is the inherent bias that the without advice condition is always presented first (Kobayashi, 1996).

1.4.4 Attitudinal Factors Affecting Automation Use

Attitudinal factors are related to personal user factors but are specifically concerned with the user's attitude. Factors under this category are divided into: (1) bias toward automation, and (2) trust and confidence.

1.4.4.1 The Effect of the Bias Toward Automation on Automation Use

In cinema, literature, and the consumer realm, computers are often portrayed as infallible, "better, more reliable, and more credible sources of information" compared to humans – almost as if they were "magical", having an "aura of objectivity", having a "scientific mystique", having "superior wisdom", or "faultless" (Tseng, 1999). Thus, the assumption is that users generally assign more credibility to automated tools than to humans (Tseng, 1999). However, empirical research does not support this notion (Tseng, 1999). In the following, at least three empirical studies that discuss this belief are presented. The studies show that even when initial trust in automated tools is stronger than that in a human partner, this positive bias leads users to be more sensitive to tool errors and to have a sharper drop in trust and tool reliance following tool faults.

In a route planning experiment, authors examined the effect of error feedback on trust, self-confidence, and allocation of control (de Vries, 2003). Results showed that those who experienced lower reliability from the tool were less likely to choose it in subsequent trials. The same was only marginally significant for the manual case. Thus, results showed a strong preference for manual mode and a bias to trust one's own abilities over the abilities of the automated tool (de Vries, 2003). There was also a higher susceptibility of trust in the automated tool to errors as they were more easily noticed and remembered (de Vries, 2003). This may be because people have access to the causes of their errors but not the tool's errors, which makes it easier for them to find excuses for their errors compared to those of the tool (Van Dongen, 2013). Furthermore, these excuses are more likely to be related to temporary and uncontrollable causes rather than permanent and internal factors as in the tool's case (Van Dongen, 2013). Users may also be less forgiving of tool errors because they have high

expectations of the tool's reliability which leads them to pay more attention to errors that contradict this high reliability schema (Van Dongen, 2013).

Madhavan et al. (2005) compared human-human interaction to human-automated interaction using a luggage screening task that varied in reliability and expertise. The experiment contained eight conditions of varying advice: (1) novice human exhibiting low reliability, (2) expert human exhibiting low reliability, (3) novice tool exhibiting low reliability, (4) expert tool exhibiting low reliability, and vice versa for high reliability. Results showed that agreement increased with increased reliability. Interestingly, however, at 70% reliability, this agreement depended on the source and expertise of the advice. When advice was described as coming from a novice, user reliance on the tool was more than that on human advice. When advice was labelled as coming from an expert, reliance on the tool was the same until the last experiment block when it was significantly lower than reliance on human advice. Authors concluded that tools are judged more by "situational actors than dispositional traits", and that users were more forgiving of human errors than tool errors (Madhavan, 2005).

These results were replicated in yet another study by Dzindolet et al. (1999) in a signal detection task. Again, an overwhelming majority of students did not rely on the tool (84%), even among participants provided with feedback that the tool's performance was far superior to their own. This "disuse was greater among participants working with a superior automated aid than among those working with a superior human aid". However, rather than a bias against automation as would be expected from the results, a bias toward automation was found in the questionnaire results. Students rated the expected performance of the automated tool more favourably than that of the human aid (Dzindolet, 1999). Given the apparent bias toward automation, why did participants ignore advice from the superior automated aids? The justifications provided by the study participants explain that this was largely due to errors committed by the tool which the users were confident the tool had erred in. Thus, in this experiment, incorrect tool advice influenced participants' perceived tool reliability to such an extent "that users would choose to rely on a manual system known to be half as good". Some participants also justified this disuse and self-reliance with "a moral obligation to rely on oneself". Finally, results from a recent study corroborate these results as when subjects were in disagreement with the aid, they did not rely more often on it although they perceived it to be 30% more reliable than themselves (Van Dongen, 2013).

1.4.4.2 The Effect of Trust and Self-Confidence on Automation Use

Riley suggested that "reliance, trust and confidence act as the primary cognitive mediators for human-computer interaction" (cited by Goddard, 2012a). Of these, trust, "the belief that an automated DSS provides reliable information", is arguably, the most researched mediator in the automation bias literature (Goddard, 2012b). In fact, one of the three main factors attributed to the occurrence of AB is the "the perceived trust of humans in automated aids as powerful agents with superior analysis capability" (Parasuraman, 2010).

Subjective measures of user trust in an automated tool have been strongly associated with users' frequency of use of the tool (de Vries, 2003) and inversely related to users' monitoring behaviour. Some studies did not find such a link and this potential discrepancy highlights the differences between subjective and objective measures of trust (Parasuraman, 2010). This discrepancy could also be due to differences in trust in the specific DSS in the experiment versus a general level of trust in automation (Goddard, 2012b).

An important relationship published in the literature is that relating trust, self-confidence, and automation use. In several papers cited by Goddard (2012), increased self-confidence decreased tool reliance, whereas trust in the tool increased reliance. As Moray et al. note (2000), it's not trust alone but the difference between trust and self-confidence that affects automation use. Parasuraman (1997a) reported that if trust > self-confidence then automation was engaged. This was also confirmed by de Vries et al. (2003) in a route planning experiment, where this "trust – self-confidence" variable was significant in the regression of determining which mode (manual or automated) was chosen by a user.

Thus, automation is influenced by the comparison between the perceived reliability of the automated tool and the perceived reliability of one's own abilities (Dzindolet, 1999). Unfortunately, the accuracy of this comparison is likely to be low since this trade-off between trust and confidence "is tempered by individual predispositions in terms of cognitive styles and technology acceptance" (Goddard, 2012b) and sometimes tempered by a lack of evidence of high/low reliability in the given set of cases. Self-confidence itself can be affected by experience (Goddard, 2012b) , and humans' tendency to exaggerate their abilities (Dzindolet, 1999; Van Dongen, 2013).

1.4.4.2.1 Factors Affecting Trust and Self-Confidence

Trust itself is also influenced by several factors, and when defined as a dimension of credibility can be seen as being influenced by four key stages of credibility: reputed, surface, experienced, and presumed credibility (Tseng, 1999). Relying solely on any one of "these ways of accounting can lead to mistaken views" about how tool advice should be interpreted (Hartswood, 2003).

- *Reputed credibility* describes how much a user believes something because of how a third party presents it (Tseng, 1999). For example, Dzindolet et al. (1999) found that users working with automated tools whose performance was framed negatively were more likely to rely on the tool's decisions than other users, even though both were presented with equal quantitative statistics of tool performance.
- *Surface credibility* describes credibility based on simple inspection and early impressions (Tseng, 1999). Effects of interface design would, for example, affect this type of credibility.
- *Presumed credibility* describes how much a user believes advice "because of general assumptions in [their] mind" (Tseng, 1999). This type of credibility is affected by a user's prior knowledge about the tool and preconceived notions concerning the performance of automated tools in general; the latter which leads many people to expect high reliability from tools (Dzindolet, 1999). This has been discussed in the previous section.
- *Experienced credibility* depends on a user's accumulated experiences with the tool and is likely the most complex type of credibility because it includes a chronological component that leads to dynamic credibility (Tseng, 1999). This type of credibility will be discussed further in the next section.

Moray et al. (2000) applied thirteen variables to a multiple regression analysis of trust and confidence. Trust was found to depend on tool reliability, the occurrence of a false diagnosis, and the occurrence of a disagreement between the user and the tool. Self-confidence was found to depend on user's rating of self-confidence, the occurrence of an error in the system and very slightly on tool reliability. An interesting property of the two resulting regression equations is that they converge. In the limit, "subjective trust would be self-calibrating to a level slightly below the objective reliability of the [tool]", while self-confidence would converge to an asymptotic value (Moray, 2000). The equations also suggest that trust is reduced by tool *properties*, whereas self-confidence is reduced by user *experiences*, and specifically that trust, but not self-confidence, changes dramatically as a function of tool reliability (Moray, 2000).

Other factors that have been empirically investigated as possible influences in people's trust in automation include: cultural differences associated with dependence on authority (Goddard, 2012b), the consistency of the tool's reliability (de Vries, 2003; Moray, 2000), interface features (Van Dongen, 2013), the user's tendency to trust (Van Dongen, 2013), and the invasiveness or intrusiveness of the tool's advice (Goddard, 2012b) (solicited advice compared to unsolicited advice is better received (Bonaccio, 2006)). However, trust does not seem to differ based on type of information: harmful or unharmful (de Vries, 2003).

As highlighted by Moray et al. (2000), an important factor affecting trust is the number of errors committed by the tool. Also important is the type of error committed, although literature on this is contradictory. Madhavan et al. (2007) found that trust degraded more quickly when the DSS made errors on easier tasks. In another study, large errors seemed to hurt credibility more than small errors (cited by Tseng, 1999). While in a third study, no difference was recorded in credibility between small and large errors (cited by Tseng, 1999). Muir even found that signs of incompetence in the tool which do not affect overall tool performance significantly decreased trust (cited by Goddard, 2012b). This relationship is further complicated by the fact that users may sometimes be unaware of a tool's failures as indicated in the follow-up questionnaire conducted by Alberdi et al. (2004). From the JAS literature, a good reputation is often difficult to gain, while it is easily lost with decreased advice quality (Bonaccio, 2006).

1.4.4.2.2 Changes in Trust Over Time

It is important to recognize that trust is dynamic (Van Dongen, 2013). General findings suggest that users initially trust automated tools and that trust declines only after tool failures are noticed (Wiegmann, 2002). Mistrust and disuse of the aid then occur (Wiegmann, 2002). Results that show that a single automation failure can significantly reduce trust in a tool suggest that a negative feedback loop that reduces complacency potential is initiated whenever a user experiences a tool failure (Parasuraman, 2010).

Research has shown that the reduction in operator trust following an automation failure is gradually regained when the tool presents accurate information, but not always to its initial level (Parasuraman, 2010; Wiegmann, 2002; de Vries, 2003). Sometimes this recovery takes time (Parasuraman, 1997a), but others have also found that occasional automation failures were not a deterrent to use automation and that students did not hesitate to reopen an automated tool after recovering from a tool failure (Parasuraman, 1997a). Sometimes, however, regaining trust may be difficult because the user may stop using the tool and the

only way to regain trust is to use the tool and find it reliable. Furthermore, Moray et al.'s (2000) equations previously discussed in the previous section, suggest that trust changes slowly and steadily in the absence of tool errors and disagreements but is significantly affected by the occasional tool error (Moray, 2000). The equations also suggest that only recent faults contribute significantly to the current level of trust, possibly because rapid discounting of past events is common (Moray, 2000).

If the reduction in user trust following a tool failure is followed by a slow recovery in trust, "then one would expect the complacency effect to be high before the first failure" due to an initial high-level of trust in the tool, "but to dissipate thereafter, a phenomenon referred to as the *first-failure effect*" (Parasuraman, 2010). Thus, exposure to automation failures may allow for more accurate calibration to the tool's true reliability, so that tool errors are better detected (Parasuraman, 2010). Some evidence for the first-failure effect was reported in a recent study, but other studies using different tasks have not found consistent evidence for it (cited by Parasuraman, 2010).

1.5 Explanations for Automation Bias

Designers often focus on the capabilities and cost of a new DSS and give little attention to the human cognitive factors involved with its use. It is important to consider how the information is processed by the user, because this can help predict and avoid automation-related errors (Tsai, 2003).

Several researchers have attempted to explain the cause of automation-related errors. Many of the explanations provided are inter-linked, complementary, and sometimes difficult to separate. An important idea presented repeatedly in the literature is that both AB and complacency errors seem to stem from the user's reallocation of attention (Parasuraman, 2010). The high workload and time pressure conditions associated with many of these tools and the environments they are used in also point to users heuristically using DSS output (Skitka, 2000), or over-relying on automation (Goddard, 2012b) so that attention can be channelled towards other tasks. However, as will be seen in this section, these are not the only explanations, and it is necessary to decompose the underlying causes of commission and omission errors while taking into account that there is likely not just one cause that explains all automation-induced errors. As mentioned in the introduction of this thesis, that is also why we have chosen to use a more general definition of AB than used by many authors, rather than defining it, as many authors choose to do (Lyell, 2016; Goddard, 2012), using a single one of its potential causes (for example: overreliance).

1.5.1 Changes in a Decision due to Uncertainty About the Original Decision

It is interesting to note that many of the cognitive mechanisms and changes in a decision after using a DSS occur in cases where the user exhibits uncertainty (Hartswood, 1998). In fact, evidence from the JAS literature shows that a judge's uncertainty in his/her initial decision predicts advice seeking (Bonaccio, 2006). This same trend is noted in the AB literature: users are more likely to accept DSS advice when they are less confident in their own decision (Dreiseitl, 2005; Goddard, 2014). When uncertain, users may also rely on simple heuristics (Detmer, 1978). Moreover, since the user's need for information in a decision affects how receptive they are of advice (Tseng, 1999), users may use any available advice to resolve uncertain cases (Alberdi, 2010). It is in this vicinity, around the user's internal threshold, that the strongest effects can be expected from tool advice (Hartswood, 1997). For example, Alberdi et al. (2014) conducted a unique ethnographic study whose think-aloud design gives insight on how a tool's advice can shift a user's decision when examining an uncertain case. They quote one radiologist from their study: "This is a case where without the prompt I'd probably let it go …but seeing the prompt I'll probably recall ... it doesn't look like a mass but she's got quite difficult dense breasts … I'd probably recall".

Although often used interchangeably, it is important also to distinguish between uncertainty and difficulty which may not always represent the same phenomena. Interestingly, lack of uncertainty in a user's decision could be a cause of disuse or some system false negatives. For example, Taylor et al. (2005) noticed that 22% of breast cancer cases correctly prompted by a computer-assisted diagnostic tool were not recalled and that there was a clear correlation between user confidence in their own pre-advice decision and failure to recall the correctly prompted case after advice.

1.5.2 Automation Bias due to Heuristic Processing

In the literature, a factor repeatedly posited as being a cause of automation bias is the tendency of humans to choose the road of least cognitive effort in decision making (Parasuraman, 1997a; Dzindolet, 1999; Skitka, 2000). One way humans minimize this cognitive effort is by substituting compensatory (comparing and weighing multiple cues) with non-compensatory decision strategies (simple heuristics that require fewer cues) (Goddard, 2012b). In fact, cognitive dual process theories posit that decisions can be based on one of two systems: logical effortful Bayesian processing or faster rule-of-thumb heuristics (or cognitive shortcuts) (Goddard, 2012a). The

former encompasses the process of analytical intelligence while the latter, is often quick, inflexible, and spontaneous (Goddard, 2012a).

As first proposed by Skitka et al. (2000), automation bias may be a case of substituting comprehensive information analysis and vigilant information seeking with applying the heuristic use of automation as the accurate answer. In this way, AB is categorized similar to other heuristics and biases in human decision making (Dzindolet, 1999; Parasuraman, 2010). This idea helps explain why increased workload, increased task difficulty, or tighter time constraints are related to higher levels of AB: they can all cause users to shift from more effortful processing to heuristic-based decision-making. Dangerously, users may adapt to rely on the tool's advice in this way, and may not realize they are doing so (Alberdi, 2005).

However, one of the main justifications for use of automated tools, for example in a cockpit, is to lessen the cognitive load on the user. Therefore, it can be argued that "heuristics are an adaptive method of saving cognitive and time resources" (Goddard, 2012b). When advice is reliable, heuristics are useful, but their indiscriminate use when the tool is imperfect could explain some of the commission and omission errors documented in the literature (Mosier, 1998).

1.5.3 Automation Bias Due to Overreliance

One simple explanation of users' responses to inaccurate tool advice is that it is a form of overreliance (Meyer, 2001). In fact, Goddard et al. (2011) define AB as "the tendency to over-rely on automation", and Lyell et al. (2016) state in the most recent review on AB that "automation bias happens when users become overreliant on decision support". However, this single, wide definition overlooks other causes and likely encompasses other key ideas.

Overreliance may itself be fostered by an attitude of trust toward automation (Parasuraman, 2010; Alberdi, 2009). Dzindolet et al. (1999) describe the authority hypothesis where people rely on an automated tool because they believe it to be superior, and thus, assign greater authority and trust to it compared to other information sources. This idea is also closely linked to the concept of attention allocation (see Section 1.5.5) because users use an attention allocation strategy to sample tasks, with attention to each task being driven in part by trust (Parasuraman, 2010). Besides linking overreliance to the authority hypothesis and attention allocation, overreliance may manifest itself in users becoming complacent (Parasuraman, 2010, Alberdi, 2009), which is another common explanation for automation-related errors (see Section 1.5.4).

One important distinction in the literature is that between reliance and compliance, especially that experimental results can sometimes show different results in terms of the two terms (Meyer, 2001). *Compliance* refers to following the tool's advice when it indicates there is a target (Meyer, 2001). On the other hand, *reliance* refers to a user assuming there is no target when the tool suggests this (Meyer, 2001). Using this terminology, unwarranted *compliance* leads to *errors of commission* and unwarranted *reliance* leads to *errors of omission.*

1.5.4 Automation Bias Due to Complacency and Reduced Sampling

One of the most common explanations for automation-related errors in the literature is complacency (Parasuraman, 2010; Hartswood, 1997; Mosier, 1998). In fact, "the National Aeronautics and Space Administration Aviation Safety Reporting System includes complacency as a coding item for incident reports" (Parasuraman, 2010). However, complacency is not well defined in the literature, although it appears to be linked to: deficient cross-verification (Parasuraman, 2010), "unjustified assumption of satisfactory system state" (Billings cited by Parasuraman, 2010), insufficient monitoring of automation output (Goddard, 2012a), a low index of suspicion (Parasuraman, 2010), unawareness of dangers of failures (Goddard, 2012b), reduced vigilance (Alberdi, 2009), and a lower effort to engage (Goddard, 2012b). Common to most of these definitions is a sense of user contentment, unawareness of dangers, and failure to carefully examine the raw data. Interestingly, Alberdi et al. (2009) show that AB can exist even when a user does not trust the tool.

It is important to note that concluding that an automation-related error was due to complacency is warranted only if it could be shown that the error was related to the user "monitoring automation less frequently or less keenly than the optimal value for a particular system" (Parasuraman, 2010). Perhaps the clearest such link between automation bias errors and complacency was documented by Bahner et al. (2008). In a lab setting, an automated tool provided advice for fault diagnosis and management, and complacency was measured by the participants' information sampling behaviour. Results showed that 21% of participants committed easily recognizable commission errors, and analyses of information sampling behaviour revealed that these errors were related to a higher level of complacency in these participants compared to those who did not commit the error (Bahner, 2008). In another study, Kobayashi et al. (1996) noted that radiographs were scanned more rapidly with CAD advice than without it. In another study, users who committed automation bias errors cross-checked, on average, only 61.5% of the necessary system parameters (Reichenbach, 2011). Finally, Metzger et al. found that air traffic controllers who missed a conflict among aircrafts had "significantly fewer fixations of the radar display under automation

support than under manual control" (cited by Parasuraman, 2010). This last finding provides convincing evidence linking automation complacency to reduced visual attention of the raw data (Parasuraman, 2010) – an idea discussed in the following section.

Parasuraman et al. (2010) also differentiate between complacency and complacency potential where complacency potential is conceived as "a behavioural tendency to react in a less attentive manner in interacting with a specific automated system", while complacent behaviour is reflected in poor monitoring of automation. Using this distinction, complacent behaviour may occur only when complacency potential coexists with other conditions such as the mediators previously discussed in Section 1.4.

There are important problems with categorizing all automation-related errors as effects of complacency and as Alberdi et al. (2014) state, doing so would be "simplistic and unfair". Firstly, complacency suggests that the user is at fault (Alberdi, 2014). More so, complacency does not explain why experts that appear attentive and thorough still commit omission and commission errors, a concept documented in several papers (Alberdi, 2005; Alberdi, 2014). For example, Bahner et al. (2008), whose study provided clear evidence linking AB to complacency, also reported that 20% of participants followed incorrect tool advice despite sampling sufficiently. The results by Alberdi et al. (2014) that presence of CAD prompts in an area made it more likely to be marked by a clinician using CAD, also suggest that users are paying attention to tool prompts, and comments from their ethnographic study support this. Besides not explaining all automation-related errors, many studies attributing AB to complacency also fail to provide an explanation of the cognitive processes involved in complacency (Ho, 2005).

1.5.5 Automation Bias Due to Attentional Effects

Both AB and complacency remain ill-defined and further research is needed to identify the differences and relationships between them (Goddard, 2012a). Parasuraman et al. (2010) addressed this need in a recent publication proposing that AB and complacency represent "different manifestations of overlapping automation-induced phenomena, with attention at the centre". In this publication, automation complacency is described as an active (not passive) redirection of attention from the automation to the manual tasks in a way that errors committed by the tool go unnoticed by the user (Parasuraman, 2010). Automation bias, on the other hand, is described as a manifestation of omission and commission errors due to errors committed by the tool (Parasuraman, 2010). This explanation helps explain why complacency is more easily prevalent in a multitask environment.

Two different types of attention can be described. The first, as explained above, represents an active reallocation of attention manifested in reduced information sampling. However, sometimes there may be a subtler effect reflected in less attentive processing of this information (Parasuraman, 2010). At the heart of this second type of attention allocation are some earlier findings. Researchers found that commission errors seemed to be attributable to a type of looking-but-not-seeing effect. In this effect, pilots (Mosier, 1998), and in a later experiment students (Skitka, 2000), who committed commission errors recalled in the final questionnaire at least one cue in the raw data indicating a fire. No such cue existed in the raw data, and the faster users shut down the engine, the more cues they recalled. Thus, the commission errors in this case could not be explained by reduced sampling or by discounting of contradictory information in the raw data. Instead, this "phantom memory" (Skitka, 2000) or "inattentional blindness" (Parasuraman, 2010) seems to suggest either a covert attention effect reflected in a withdrawal of attention from processing the looked-at information (Parasuraman, 2010); or a form of confirmation bias suggesting users believed they were seeing information that verified the automated advice (Mosier, 1998).

As researchers highlight, attentional factors contribute to some but not all forms of automation bias (Parasuraman, 2010; Wickens, 2007; Ho, 2005). For example, attentional effects may explain why users miss automation faults because they learn to pay less attention to certain tool prompts based on experienced reliability of the tool (Alberdi, 2003). However, they do not explain why a rational user, paying full attention to all evidence in the raw data and tool prompts errs with the tool compared to without the tool. For example, Ho et al. (2005) found that older adults failed to detect many automation failures despite sufficient monitoring, and that it appeared that "many missed signals were [instead] the result of perceptual and/or cognitive factors". Such errors where users err despite sufficient monitoring suggest the need for future research on the relative importance of attentional effects versus other explanations for automation bias (Parasuraman, 2010). Importantly, Lyell et al.'s (2016) recent review on automation bias revealed, contrary to the dominant human factors view, that AB occurs in single task environments, especially in the healthcare context. This finding "challenges the theoretical proposition put forward by Parasuraman that automation bias occurs when multiple tasks compete for the user's attention" (Lyell, 2016).

1.5.6 Automation Bias Due to Normative Reliance

Some of the terminology used to describe the root of automation-related errors conveys a value judgement on the user. However, user behaviour may actually represent a normative and rational

rather than careless adaptation. As Wickens et al. (2007) state, the user may depend on the tool to "preserve available processing resources for other tasks", and this may be why AB is more prevalent in high workload conditions. In light of this explanation, what has been characterized as "complacent" behaviour, may in fact represent a rational strategy. When users have several tasks to attend to, their monitoring of automation should be in proportion to its perceived reliability. Given that highly reliable tools fail only very rarely, then the rational strategy would be to cross-check their advice infrequently. A natural consequence of such a monitoring strategy is that users will occasionally miss automation failures because their attention will be allocated elsewhere. As Alberdi et al. (2004) state in the context of reliable CAD tools, considering the low specificity of the tools, using the absence of prompts to diagnose patients as observed by the authors, may in fact be termed "normative reliance". Such behaviour would be especially rational if detailed analysis of every prompt were too demanding, if in fact feasible (Alberdi, 2009). Wickens et al. (2007) question the notions of complacency or reduced vigilance, and argue that users are aware of the unreliability of the automated tools (unlike complacency as an explanation which implies that users are unaware or unsuspicious), but that they *choose* to depend on the imperfect tool to save their cognitive processing resources, particularly in high workloads and when using highly reliable tools.

1.5.7 Automation Bias Due to Social Loafing and other Group Processing Factors

Another explanation of automation bias comes from applying some of the cognitive, social, and motivational processes that have been implicated as causes of reduced productivity in human groups (Dzindolet, 1999) to the human + automated tool system. An important behaviour noted in human groups is social loafing (Dzindolet, 1999; McKinlay, 1999): the tendency for people to work less hard in groups, largely because responsibility is diffused among all group members. As Mosier et al. (2001) note, in a group, an individual may alter his/her behaviour or vigilance according to expectations of the other team member(s)' capabilities. In this setting, a highly reliable automated tool may thus induce non-vigilance in a user (Mosier, 2001). This is corroborated by comments by Alberdi et al. (2009) that abdicating responsibility to the tool depends on self-confidence and trust in the tool such that the user may be reliant on the tool for certain tasks they know they are poor at.

Literature on how individuals in a team function shows that several factors can affect team members' behaviours and efforts.

- *Team informity:* the degree to which a team is aware of all the relevant information (Mosier, 2001)
- *Staff validity:* the degree to which each team member can make a valid decision (Mosier, 2001)
- *Hierarchical sensitivity:* "the degree to which the team leader effectively weights team member judgments in arriving at the team's decision" (Mosier, 2001)
- *Individual motivation* (Dzindolet, 1999), which in turn is based on:
    - *Expectancy*: the degree that members feel their efforts are necessary for the group's success. Thus, misuse of an automated tool is more probable when interacting with a highly reliable tool. (Dzindolet, 1999)
    - *Instrumentality*: the degree that members feel the group's success will lead to a positive outcome (Dzindolet, 1999)
    - *Outcome value*: the result of the comparison between the cost of working hard and the importance of the group's outcome. Thus, members will reduce their effort and increase their reliance on the tool when costs are increased or rewards are decreased (Dzindolet, 1999).

In a human + automated tool system, these factors mean that system performance increases when: all members are aware of the relevant data (Mosier, 2001), all members have sufficient expertise (Mosier, 2001), the user is capable of effectively weighing tool advice and cues in the raw data (Mosier, 2001), the user understands the importance of the task outcome and their individual role in the system. For example, McKinlay et al. (1999) examined "the effects of computer-mediated communication on social loafing in brainstorming tasks". The nominal group (individuals who worked in isolation and whose ideas were then mixed) generated more ideas than the face-to-face or computer-mediated brainstorming groups. More importantly, the experiment showed that the computer-mediated group's performance was worse than that of the face-to-face group, most likely due to some of the ideas of expectancy and outcome value discussed here (McKinlay, 1999).

1.5.8 Automation Bias Due to the Homeostasis Mechanism

A more recent explanation for automation bias, not yet well adopted in the literature, is a concept introduced by researchers at the Centre for Software Reliability at City University London and which they refer to as the *homeostasis mechanism*. In a fairly recent publication, these researchers use a comprehensive diagram to describe possible sources of false negative errors in automated systems (Alberdi, 2009). This diagram highlights cases where a false negative can occur without complacency or reduced vigilance, without trust in the tool, even without the user relying on the

tool, and despite cautious protocols and training. Instead, it can be caused by users simply altering their decision threshold. More specifically, users may raise the threshold above which they respond to cues; for example, by not recalling a borderline case and seeing the tool's advice before committing to a decision (Alberdi, 2009). With the homeostasis mechanism, a tool could have an effect on a users' decision but this would not be observed by positive and negative consultations in the decision, but by differences in the pre-advice decisions between groups that know they have access to a tool compared to those who do not know.

Two other studies seem to explore similarly related, although different, concepts. Elvers and Elrif conducted a small study, which provided evidence that in a signal detection task, users' criterion setting was affected by the warning tool's criterion and that with a liberal tool setting, users tended to adopt a more conservative criterion (cited in Meyer, 2001). The homeostasis mechanism goes beyond this and suggests that tool presence alone can alter a user's criterion. In 1979, Parasuraman et al. explored whether vigilance decrement in a visual/auditory task could be due to sensitivity decreasing over time or a more stringent user criterion. Although their study did not involve automated tool advice, just a user, their analysis of the results is relevant. To distinguish between the two hypotheses, authors plotted the ROC of different groups over time, and noticed that when the task loads memory and when events occur rapidly, vigilance decrement is due to decreased user sensitivity/detection ability. Otherwise, the change is due to a difference in the user's criterion. In other words, when the ROC moves up and down then it's a change in sensitivity, and when the ROC is the same then change is due to the criterion.

Alberdi et al. (2009) note that homeostasis is likely affected by both the need to keep the alarm rate below a certain limit and further exacerbated by protocols that do not allow users to change their pre-advice decision from target to no target. Whether the homeostasis mechanism exists in different domains with different costs and protocols is certainly a question of interest.

1.5.9 System Errors Due to Other Explanations

Besides the explanations covered in the previous sections, a mix of other unrelated reasons which may cause system errors, which may sometimes be attributed to AB, have also been provided in the literature. Kobayashi et al. (1996) mention subjective factors such as distractions, and the varying decision criteria of users. Sometimes, an error may simply be due to active discounting of information that contradicts the recommendation of the tool, especially in difficult cases (Skitka, 2000; Parasuraman, 2010). Also, in CAD reading, Hartswood et al. (1997) mention missing features because they are so faint or errors due to misclassification rather than detection. In such

cases, the user sees the feature, acknowledges the tool's prompt with sufficient attention but decides that it is benign not malignant (Hartswood, 1997; Zheng, 2001).

Egocentric bias whereby users prefer their own opinions because they believe them to be superior to others' opinions is a bias described in several studies (Sniezek, 1995b; Roehrig, 1999; Bonaccio, 2006). Egocentric bias exists even for decisions about novel situations and is different from anchoring bias in that it refers to the long-term influence of one's personal opinion (Bonacci, 2006). Finally, there may be other psychological biases which may be involved in the decision-making process of automated systems. One important bias is confirmation bias which describes people's tendency to seek information that confirms a prior opinion and refute information that disconfirms it (Croskerry, 2002). With so many other explanations for automation-related errors observed in the literature, it is important when designing an experiment exploring AB to appropriately measure the outcome as some errors may be attributed to AB and in fact be the result of some of these other unrelated factors. This will be discussed in more detail in Section 1.7.

1.5.9.1 Automation Bias and Anchoring Bias

Anchoring is a cognitive bias that describes the tendency, when making decisions, to rely too heavily on the first information offered – termed the "anchor". In other words, under this bias, users exposed to identical information in different orders will make significantly different decisions (Friedlander, 1984). Two main explanations exist for anchoring errors. The first explanation is that people "may not adjust sufficiently" from their initial anchor in order to maintain consistency with their initial opinion (Friedlander, 1984). The second explanation is that a user's "processing of new information may be unduly influenced by his or her preconceived notions" and "integrated into the existing cognitive representation" (Friedlander, 1984). Thus, when advice warrants a significant change, the user may exaggerate features related to their first impression but downplay features contradicting their first impression (Tsai, 2003; Friedlander, 1984). Consequently, anchoring bias can both modulate how users interact with a DSS and also influence how they interpret DSS advice.

Anchoring bias is documented in several studies evaluating automated tools, and more extensively in the JAS literature. Tsai et al. (2003) found that subjects were unable to sufficiently adjust from the initial computer interpretation when interpreting ECGs . Berner et al. (2003) also found that physicians were strongly anchored by their initial diagnoses prior to viewing QMR's suggested diagnoses. Teich showed that users were more receptive to advice that did not require a significant change in their initial decision (cited by Goddard, 2012b).

From the JAS literature, Yaniv (2004) found that "the weight of advice decreased as its distance from the initial opinion increased". Yaniv (2004) also cited a study exploring changes in stereotypes where Sherif et al. noticed that a slightly different opinion caused people to shift their attitude, but significantly different opinions had a reduced effect on people's attitude as they fell outside their ''latitude of acceptance''.

Anchoring bias is an important effect, and is more pronounced in certain situations. For example, understanding this effect is especially important when advice is presented sequentially (compared to simultaneously) as is the protocol when using many modern CDSSs. It may also be affected by the extent that a user is held accountable to their initial decision (i.e., do they have to write it down, and are they allowed to change it in either direction?). It is also especially influential when interacting with a CDSS like QMR where what the user selects from the raw data as input into the tool dictates the tool's output. For example, Berner et al. (2003) showed that besides the fact that physicians were strongly anchored to their initial diagnoses, the presence of the correct diagnosis in a prominent position by QMR depended on the user's initial diagnosis.

The importance of the anchoring effect helps explain why simply presenting advice to users does not guarantee that it will be attended to and utilized appropriately (Berner, 2003). This is important to recognize when evaluating DSSs and perhaps also in understanding some of the cases of disuse documented in the literature. For example, anchoring bias could explain cases where a DSS provided useful information that was not acted upon, such as in Taylor et al.'s study (2008a) when two of four readers missed the same polyp despite a correct CAD prompt – although these may also be cases of discounted information. A serious danger is that advice that is closer to a user's original opinion is more likely to be accepted, even if it is wrong. It is important to clarify that anchoring bias does not cause automation bias but that some system errors such as when a user fails to heed appropriate DSS advice, could be explained by anchoring bias.

1.5.10 Summary of Automation Bias Explanations

In what follows, we use "AB" as we previously defined it: automation-induced errors that sometimes result in performance which is actually worse with the automated tool than without it. We emphasize that limiting the definition of AB by attributing it to just a single one of the causes presented in Section 1.5 is both simplistic and unfair. Even where there is evidence in an experiment that AB is caused by one of these explanations, as in Bahner et al.'s (2008) experiment

linking AB to "complacency", researchers still find that not all the errors in their study can be attributed to this single cause (for example, Bahner et al. reported that 20% of participants followed incorrect tool advice despite sufficient sampling of the raw data). Researchers that propose models of the cognitive mechanisms involved in the decision-making process in cases of automation bias also suggest the need for future research on the relative importance of such effects versus other explanations for automation bias (Parasuraman, 2010); thus emphasizing that their explanation alone does not explain all automation bias errors. Furthermore, many of the explanations provided are inter-linked, complementary, and sometimes difficult to separate. For example, overreliance can decrease users' "attention" to raw data (readiness to detect incorrect tool advice) or can result in users becoming "complacent", two essentially similar phenomena which are not clearly differentiated by authors using them as explanations for automation-induced errors.

As will be discussed in Section 1.8.1.1, this thesis focuses on providing empirical evidence of the homeostasis mechanism as one cause of automation bias. The choice to focus on this mechanism is because it has not been previously studied in the literature, despite the fact that it can account for some of the errors unexplained by conventional explanations for AB. However, the intention is not to deny/negate the validity of some of the other explanations, but to add another necessary explanation. Future research should elucidate the relative importance of the different explanations, in different contexts and in relation to different mediators. Finally, the holistic view we have adopted of the potential cognitive mechanisms that can cause AB is very important, especially when implementing effective design and protocol decisions aimed at mitigating automation bias.

1.6 Strategies Used to Mitigate Automation Bias

Mitigation strategies proposed and used to combat automation bias are closely linked to the ideas presented in Section 1.4 "Factors Affecting Automation Use", as many of them aim to appropriately decrease/increase factors that have been proven to affect automation use. Many of these strategies aim to reduce errors due to the user perceiving the tool as credible when it is not (gullibility errors), but the wider view should, in fact, be to reduce errors due to the user perceiving the tool as not credible when in fact it is (incredulity errors) without increasing gullibility errors (Tseng, 1999). Also important is that many of these strategies are aimed at combatting automation-related errors due to complacency or over-reliance. However, as seen in Section 1.5, these may not be the only causes of such errors, and future strategies should aim to mitigate these errors based on a new, wider perspective of their causes.

The mitigation strategies presented in this section are divided into: (1) strategies linked to user characteristics, (2) strategies linked to tool properties, and (3) strategies linked to environmental factors, and are summarized in Figure 1-5.

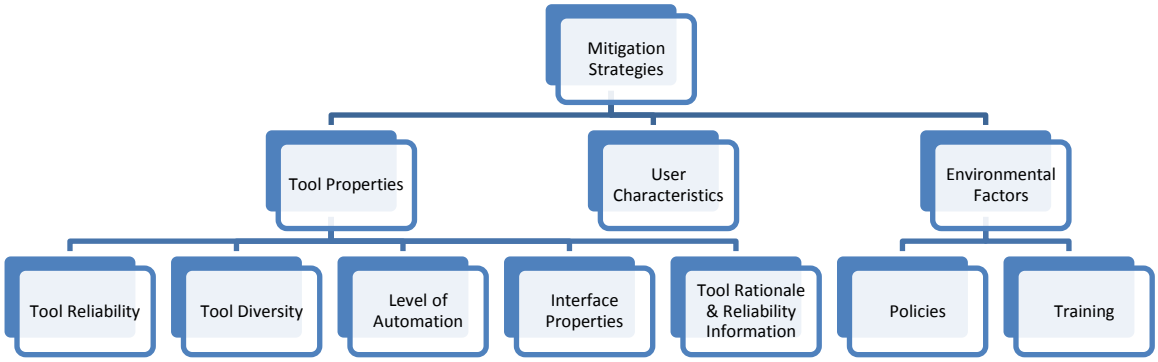


Figure 1-5: Summary of Strategies Used to Mitigate Automation Bias

1.6.1 Mitigation Strategies Linked to User Characteristics

Increasing accountability for decisions may prevent AB (Parasuraman, 2010), especially because it may discourage users from taking the road of least cognitive effort (Van Dongen, 2013). However, as Alberdi et al. (2009) note, protocols such as requiring that users reach an initial decision and take responsibility for it (e.g. by recording it) before seeing a tool's prompts might work to fix errors due to complacency or heuristic reliance. But interactions are more complicated, and such protocols may cause increased time pressure which could in turn increase cognitive load and decrease vigilance. Furthermore, if the error is due to other reasons such as the homeostasis mechanism discussed in Section 1.5.8, then such a protocol would not only be ineffective, but would exacerbate this mechanism, possibly causing more automation-induced errors.

When combatting social loafing which has been documented as another cause of some automation-related errors, the psychology literature suggests that increasing "the extent to which the group seems meaningful, coherent and cohesive" helps reduce this effect (McKinlay, 1999). Another mitigation strategy slightly touched upon by Philpotts et al. (2009) and Goddard (2012) is producing a DSS which provides good cognitive fit with the user.

1.6.2 Mitigation Strategies Linked to Tool Properties

Rather than mitigation strategies aimed at manipulating certain user behaviours or characteristics, other strategies focus on the design of the DSS itself. Strategies in this category are divided into:

(1) tool reliability, (2) tool diversity, (3) level of automation, (4) interface properties, and (5) tool rationale and reliability information.

1.6.2.1 Manipulating Tool Reliability to Mitigate Automation Bias

As seen in Section 1.4.2.1, automation complacency is often reduced when the automation failure rate is increased. Increasing tool unreliability also avoids the risk of a decline in manual skills. However, excessively low tool reliability can result in decreased trust and disuse of the tool (Goddard, 2012b). As discussed in Section 1.2, whether there is a threshold reliability level below which automation complacency does not occur, and the optimal setting for a given tool, remain unresolved issues (Parasuraman, 2010).

Based on proposals that repeated exposure to the same experience leads users to engage in mindless behaviour, Parasuraman et al. (2010) reason that automation with variable reliability as opposed to constant reliability could be used to combat complacency. Their results support this notion as was seen in Section 1.4.2.1.

1.6.2.2 Manipulating Tool Diversity to Mitigate Automation Bias

Alberdi et al. (2010) raise an important issue: in some systems, it is not the tool's overall reliability that may be most important, but the diversity between the tool and the user. Diversity is used to describe an intuitive notion of how focussed the automated tool is on the true needs of the users (Alberdi, 2010), or in other words, how different the human and tool errors are. Improvements in the tool's algorithm aimed at increasing sensitivity at the cost of decreasing specificity are often expensive, subject to the law of diminishing return and sometimes ineffective (Alberdi, 2005). For example, if the tool is improved such that it prompts targets that the user already detects on their own, then little overall benefit will be observed in system performance (Strigini, 2003; Alberdi, 2010). In fact, researchers argue that a poor stand-alone tool performance is permissible as long as the tool prompts targets typically missed by the human (Strigini, 2003; Lawrence, 2010). Diversity can also help reduce system false prompts caused by automation. For example, Kobayashi et al. (1999) report that, in their study, it was not very difficult for users to disregard tool false positives because they were generally different from those of the users. Moreover, diversity could help decrease users' cognitive load, for example, by designing tools that only prompt features users have not already noticed in their initial decision (Alberdi, 2014). In fact, a study from the JAS literature cited by Bonnacio et al. (2006), suggests that users would prefer such advice as judges sought more

advice from advisors with unique information compared to advisors with information redundant to that of the judges.

Currently, many tools seem not to be very diverse from their users, suggesting that this, rather than the tool's stand-alone accuracy, is an area for improvement. For example, Alberdi et al. (2005) reported, when analysing a CAD tool used to detect breast abnormalities, that the tool and user errors were strongly correlated. Philpotts et al. (2009) also stated that CAD does not do well with subtle lesions, although these are the lesions that radiologists need the most help with. Likewise, others report that cases that were hard for users were also hard for the tool (Madhavan, 2007; Petrick, 2008).

Building tools that aid users in cases where they need the most help is not necessarily easy, but "building the tools on principles different from replicating human behaviour" is important (Alberdi, 2005). Some tool designers have already begun incorporating concepts of diversity in their designs. For example, Balleyguier (2005) described a CAD tool designed to specifically identify lesions that are difficult to detect with conventional radiological procedures and which are smaller than 3 cm (does not prompt larger lesions as they are usually correctly diagnosed by radiologists). This limits the rate of tool false positives and addresses the needs of the user by helping the user detect those lesions that are most difficult - although it does limit the use of CAD alone. Such pursuit of diversity is promising and has been effective in other fields (Alberdi, 2005).

1.6.2.3 Manipulating the Level of Automation to Mitigate Automation Bias

Another way that has been suggested to increase monitoring performance and which is linked to the previous discussion of LOA in Section 1.4.2.2 is to ensure that the user stays actively involved in the system, even if this incurs a higher economic cost (Kawamoto, 2005). For example, this can be achieved by passing on manual control to the user from time to time (Parasuraman, 1997a). Adaptive automation that gives users control to change the level of automation or where the tool changes the level of automation in response to how the human is working, has also been proposed (Goddard, 2012b). For example, the proposal to design tools that monitor the user (for example their brain activity) and take control when the user is over-loaded, and then later hand over control (Goddard, 2012b).

1.6.2.4 Manipulating Interface Properties to Mitigate Automation Bias

Since the position of advice on the screen can affect the likelihood of AB (as shown in Section 1.4.2.3), care should be taken when designing DSS interfaces. Visual design factors discussed in Section 1.4.2.3 such as reduced screen-detail, a balanced layout, consistent presentation, advice saliency, and separation/integration of the advice and raw data should be considered.

The incorporation of verification procedures into DSSs has also been suggested as a mitigation strategy. For example, users may be provided with a prompt that asks them to verify automated functioning at appropriate points. Airlines have already introduced such verification procedures for some functions, such as after an altitude change (Mosier, 2001). This is an interesting precaution especially that Mosier et al. (2001) found that display prompts that reminded crews to verify automated tool suggestions had no effect on automation-related errors. Related, Kawamoto et al. (2005) found that systems that asked doctors to record a reason for not following the tool's advice were significantly more likely to succeed. A more sophisticated version of this verification is suggested by Cuddeback et al. (2011) who propose designing tools in such a manner that makes it more difficult for the user to commit certain errors. Since authors noticed that students inspecting traceability matrices (TMs) from the high recall, high precision region reduced the overall accuracy of the TM by adding bad links and removing good links, they suggested making it harder for the analyst to reject a candidate link than to add a new link in high recall, low precision candidate TMs (Cuddeback, 2011).

1.6.2.5 Providing Tool Rationale and Reliability Information to Mitigate Automation Bias

Perhaps one of the most common suggestions in the literature aimed at decreasing the rate of automation bias is to ensure users are aware of the tool's reasoning process and scope (Parasuraman, 1997a; Hartswood, 1998; Alberdi, 2010; Kawamoto, 2005; de Vries, 2003). This is especially important because users often have difficulty understanding the operational scope of the tool (Hartswood, 1998; Skitka, 2000) and researchers have found discrepancies between the ability of automated tools compared to users' understanding of their ability (Parasuraman, 1997b). For example, in one study, researchers discovered that users' explanations of the tool's behaviour were based on false notions of its capabilities (for example, that it could detect asymmetries in the breast) (Hartswood, 1997). In fact, in a recent experiment, Van Dongen et al. (2013) showed that "the more understandable participants thought their own decision making process was compared to that of the decision aid, the more

they relied on their own initial prediction.". However, these results were based on users' perceived understanding of the tool, and not actual differences in how the tool presented the advice.

The importance of this mitigation strategy is supported by results from the JAS literature which show that one reason that judges discount information is that they have access to their internal reasons for a judgement, but not to the advisor's internal reasons (Van Dongen, 2013). Furthermore, the weight assigned to a judgement depends on the evidence that is recruited to support it, and so advice is often disused when a judge does not have access to the underlying reasons for this advice (Van Dongen, 2013). That is one reason why tools designed to present explanations, such as that suggested by Taylor et al. (2003), warrant more attention.

Even if it is agreed that users should be provided with information regarding the tool's capabilities and algorithm, exactly how much information should they be given? There is a trade-off between providing too much information and ease of use, especially that providing too much information for certain users may lead to unnecessary complexity and jeopardize their acceptance of the tool (Ho, 2005). Furthermore, it is necessary to ensure that these explanations provide understanding for the areas where confusion does indeed arise and that they do not themselves become a source of further confusion (Alberdi, 2014). In fact, in some cases, good explanations may sometimes make inaccurate advice more convincing to the user and thus increase the chance of automation bias. Such cases seem to be ignored in the literature.

One form of information that has been cited as useful to provide to the user is the tool's confidence in each advice it provides. This information helps users assess the validity of the tool's single recommendations and to make less biased decisions. In some systems, the tool merely presents binary information. However, as Lawrence et al. (2010) suggest, "If a priority score or ranking were assigned to each CAD mark, the end user would perhaps have more insight into the 'thoughts' of the CAD system, which may serve to limit reductions in specificity". In one experiment, pilots presented with indications of system confidence in the display exhibited less automation bias than those who were simply given information about the overall tool reliability (Cummings, 2004).

Some of the mitigation strategies presented in this section have been recently introduced to modern DSSs. For example, one of the latest CAD tools offers options to help radiologists understand why the CAD algorithm marked a particular region by indicating what tissue characteristics caused the flag (The Smart Choice 2). Marks by this tool are also displayed in

variable sizes correlating to the prominence of the features and as a type of confidence assessment (The Smart Choice 4).

1.6.3 Mitigation Strategies Linked to Environmental Factors

Besides strategies addressed specifically at the two main components of the system: the user and the tool, there are also strategies concerned with the environment the tool is used in. Strategies in this category are divided into two groups: policies and training.

1.6.3.1 Constructing Effective Policies to Mitigate Automation Bias

While the mitigation strategies in the previous section focus on how the tool should be designed, strategies in this section are more concerned with how the tool should be used. The user's role should be defined not as a by-product of how the tool is implemented, but based on their capability, and managers who dictate policies and training should be more aware of automation errors (Parasuraman, 1997a).

Kawamoto et al.'s review suggests that performance is improved when CDSS advice is provided automatically as part of clinician workflow and "delivered at the time and location of decision making" while others add automatically generated versus user-initiated decision support to the list (Purcell, 2005). These characteristics of workflow suggest that automation should save time or at least require minimal time to use (Parasuraman, 1997a). It also shouldn't be difficult or time consuming to turn off the automated tool because this increase in workload could make the tool less attractive to users (Parasuraman, 1997a). Interestingly, such ideas facilitating the tool into users' workflow could, on the other hand, result in an undesired increase in automation-induced errors by making it, for example, easier to rely on the tool. Recently, Van Dongen et al. (2013) suggested that advice should be provided only after people have formed their own opinion, in order for them to bring more knowledge to the task. Such ideas are concerned with how the tool is implemented into the user's work, which is often dictated by policy-makers.

Philpotts et al. (2009) expand on these ideas by providing clear guidelines on how a CAD tool should be used. They stated that it is important to "never use CAD as an initial step in mammogram interpretation" because a user will likely miss many important features if they focus their attention primarily on CAD-marked areas. Another procedural restriction often used with CAD tools is requiring the user to reach a provisional decision (and sometimes

recording it) before seeing the tool's prompts (Food and Drug Administration Meeting Minutes, 2004). However as previously discussed, this may only work to mitigate errors due to complacency or reduced vigilance. Another restriction sometimes used with CAD tools is that the user can recall more cases after seeing the tool's advice but cannot go back on a case which their initial decision was to recall. This may help prevent users from using CAD as a diagnostic rather than detection tool, but on the other hand, may induce the homeostasis mechanism discussed in Section 1.5.8. Also, as Alberdi et al. (2010) propose, false negatives that a user catches without CAD and misses with CAD could be avoided by requiring users to mark all features they are going to consider in their decision, before showing the tool's prompts. In conclusion, the literature contains numerous examples of protocols used to mitigate automation-related errors, but as seen, a wider perspective of the causes of AB will likely lead to more effective protocols.

1.6.3.2 Providing Training to Mitigate Automation Bias

If automation use is determined based on the perceived reliability of the automated tool compared to the perceived reliability of one's own abilities, then it should be possible to alter automation reliance and help, as with other human biases, *debias* users by subjecting them to training (Bahner, 2008; Cuddeback, 2011). Such training is especially important for those who have low self-confidence or a bias towards automation (Parasuraman, 1997a).

What should this training include? Is providing the tool's reliability in terms of sensitivity and specificity, as in Petrick et al.'s study (2008), enough? Or should, for example, potential tool errors be illustrated with example cases as in Hartswood et al.'s study (1998)? Importantly, as noted in Bahner's study (2008), users may need to experience automation faults and not just be informed of them in order to reduce AB. The choice of cases, the frequency of errors and their temporal sequence in the training set are also likely to be important factors.

Based on discussions of training strategies in the literature, training aimed at mitigating automation bias should serve several purposes:

- Alter the idea that automated tools are perfect (Dzindolet, 1999), or useless
- Ensure that users understand when an aid is likely to make a mistake (Dzindolet, 1999) and its limitations (Philpotts, 2009)
- Expose users to automation errors (Bahner, 2008)
- Explain, to some extent how the system works (Hartswood, 1998; Parasuraman, 1997a; Kawamoto, 2005; Alberdi, 2010; de Vries, 2003)
- Be continuous (Parasuraman, 1997a)

    - For example by allowing users to update the training set with cases from their renewed experience (Hartswood, 1998a). Or by planting fictitious cases with incorrect tool advice in the user's normal workload, frequently enough to refresh the user's memory of possible tool faults (Alberdi, 2010).
    - This is also corroborated by the JAS literature which shows that decision accuracy increases when users are provided with feedback on their performance (Bonaccio, 2006). And as Van Dongen et al. (2013) mention, this feedback should provide information about the user's individual performance, that of the decision tool and that of the system as a whole.
- Provide information on attention strategies (Parasuraman, 2010)
    - For example, the variable-priority training proposed by Gopher (to devote more attention to one task compared to other tasks in one trial and then reverse the priorities in the following trial) led to better multitasking performance, and reduced complacency versus training that instructed users to give equal and constant attention to all tasks (cited by Parasuraman, 2010).

1.7 Aspects to Take into Account When Designing an Experiment in Automation Bias

As mentioned in the discussion of possible explanations of AB in Section 1.5, it is important when designing an experiment measuring the effects of AB, to carefully design the experiment in such a manner that will allow researchers to distinguish the causes of the errors. Some important factors to consider in the design and analysis phase of future experiments in AB are discussed in this section.

1.7.1 Experimental Outcomes Used to Measure Automation Bias

Wang et al. (2008) conveniently summarize different variables that have been used to measure reliance on automation. In the list below, we use their division of four types of indicators: consistency indicators, performance indicators, behavioural indicators and response bias to explain and show examples of each.

- *Consistency Indicators* measure the consistency between the automated advice and the users' final decision
- *Performance Indicators* measure the difference between system performance when receiving correct versus incorrect advice.
- *Behavioural indicators* encompass measurements of behavioural patterns.
- *Response bias* measures user reliance on 'signal' and 'noise' feedback in binary systems

Table 1-2 lists examples of each type of indicator.

Table 1-2: Examples of Different Indicators Used to Measure AB	
Type of Indicator	**Examples**
Consistency Indicators	• Percentage of times user accepts automation feedback • Correlation between automation feedback and users' decisions (Wang, 2008)

Performance Indicators	• Comparison between user performance when receiving correct tool advice to their performance when receiving incorrect tool advice (Wang, 2008)
	• Probability of user detecting automation failures (Wang, 2008)
	• Comparison of pre-advice and post-advice user decisions (Madhavan, 2007)
	  ○ Although we argue against this based on results from Chapter 2
	• Rate of negative consultations (Goddard, 2012a)
	  ○ Especially because users may have an incorrect decision before DSS advice
	  ○ Also important is how often these negative consultations were due to erroneous tool advice (Alberdi, 2010)

Behavioural Indicators	• Scanning of cued targets prior to non-cued targets • Attention allocation in sampling rate • Cross verification rate • Return to manual performance • Proportion of times a tool is used by the user (Wang, 2008)
Response Bias	• Response bias shift between two situations (Wang, 2008) • The differential adjustment of the response criteria to the states of the tool (Meyer, 2001)

1.7.1.1 New Indirect Measures of Behavioural Indicators

Parasuraman et al. (2010) note that when studying behavioural indicators in AB experiments, it would be beneficial to measure the relative importance of attentional effects versus other effects as discussed in Section 1.5.5. To achieve this, researchers need to measure indicators of attention allocation by using methods such as eye-tracking (Dreiseitl, 2012) or information sampling (Bahner, 2008). In the latter, to claim a user was complacent, it is important to compare their sampling rate to the optimal rate of that specific system and not just to the user's rate without the DSS (Parasuraman, 2010).

Most eye-tracking research has been performed in the field of radiology, and can be especially useful in distinguishing detection and classification errors (Dreiseitl, 2012). As Hartswood et al. note in their report "Computer-Aided Mammography: A Case Study of Error Management in a Skilled Decision-Making Task", search errors are due to lesions that do not enter a user's

useful field of view, detection errors occur when the visual dwell time falls below an empirically determined threshold (often difficult to determine), and classification errors are due to the visual dwell time exceeding this threshold. Use of eye-tracking is built on evidence of a close link between eye movements and attention. This is generally true; however, sometimes the eye may be fixed at a location while attention may be elsewhere - the inattentional blindness effect previously described.

1.7.2 Distribution of Tool Errors in an Experiment

Numerous studies have investigated the effects of variable reliability on system performance by manipulating tool reliability. However, only a small subset of these studies mention the actual distribution of tool errors across the experimental session (Goddard, 2012b). Two tools with the same overall reliability can have different temporal sequences of failures and different times between failures, which can influence results (Parasuraman, 2010).

For example, Meyer et al. (2001) noticed, when studying users' reactions to binary warning signals, that experiment block made a difference when advice was imperfect. They also noticed that participants initially relied on tool warnings, then apparently learned to respond more cautiously to them. In another example, Sanchez et al. found that participants exposed to automation faults at the beginning or end of the experiment showed more reliance than those exposed to errors consistently throughout the experiment (cited by Goddard, 2012b). Sanchez et al. concluded that automation that behaves reliably for an extended period of time causes increased reliance compared to automation that fails frequently and randomly (cited by Goddard, 2012b).

Another important concept related to the temporal sequence of faults in an experiment is the sequence of targets in an experiment. Users have a good perception of the work they have done and the reasonableness of their findings of targets (Hartswood, 2003). Thus, as Hartswood et al. (2003) quote one of the radiologists from their ethnographic study, "Paranoia can set in if I have a large number of films that have passed as normal – might think 'what have I missed?'". Another radiologist in the same study is quoted: "If you get to the end of a session, the end of a pile of reporting and you haven't recalled anything, then you think 'this is ... maybe I've missed something' then in the next bunch you find that you will recall every other one" (Hartswood, 2003). This is a complex effect incorporating time into the analysis of user behaviour. In psychology, it is termed the gambler's fallacy**,** and it arises when a user believes that "chance processes are self-correcting" (Detmer, 1978). The gambler's fallacy has been reported in clinical settings.

The studies cited in this section suggest that when designing an experiment it is important to think carefully about the temporal distribution of tool errors, the time between them, and the number and distribution of targets in the experimental session as they can all have a significant effect on experimental results.

1.7.3 The Hawthorne Effect

The Hawthorne or laboratory effect, where "users change their behaviour due to the presence of researchers and the knowledge that they are participating in a study, rather than due to the effect of experimental manipulations" (Goddard, 2012b), is another effect to be aware of when designing experiments (Philpotts, 2009). For example, when investigating the use of a decision-support tool for the diagnosis of pigmented skin lesions, researchers found significantly lower system performance in the real-world setting compared to the laboratory setting (Dreiseitl, 2005). In CDSS-evaluation studies, this effect may be observed because experimental sessions are enriched with targets, participants are aware that their interpretations do not impact patients, and there is no pressure to maintain a low recall rate (Philpotts, 2009). This effect may explain why prospective and more clinically relevant studies often report more moderate results (Philpotts, 2009).

1.7.4 The "Second Look" Effect

Goddard (2012) describes another important effect that is especially relevant in experiments where advice is presented sequentially. The "second look" effect posits that decisions may be different the second time a user attempts a problem, not because of advice from the DSS, but because users have had more time to think about the case and recall relevant information. One way this has been mitigated in the literature is by preventing users from looking at the raw data after tool advice is provided (Goddard, 2014). However, this is not feasible in all experiments and may not represent a real-world setting where users are able to refer back to the raw data. Another way second look has been distinguished in the analysis stage is by checking whether a user's change in decision matched the tool's advice (Ramnarayan, 2006). This method at least distinguishes changes due to second look effect which were not in the direction of tool advice.

1.7.5 Long-Term Effects of Automated Tool Use

Given that readers learn about the tool's behaviour and complacency and automation bias result from an *adaptive* process that develops over time (Hartswood, 1998), which is clearly depicted in

the positive and negative feedback loops of Parasuraman et al.'s attention model (2010), it is thus important to consider how use of the tool changes over time. Despite its importance, little attention has been paid to the long-term effects of using automated tools (Parasuraman, 1997a).

The concept that user interaction with the tool changes over time raises issues regarding evaluation studies of automated tools. Many of the benefits of these tools are unlikely to be apparent for a considerable time after their introduction, and usually only immediate benefits or drawbacks are evaluated (Hartswood, 2003). For example, a study cited by Philpotts et al. (2009) demonstrated an initial increase of eight cancer detections per year with the use of CAD, which dropped to one or two in subsequent years. Further research is needed to understand the causes of such trends (Philpotts, 2009). Research in this area would also benefit from manipulations of the positive and negative feedback loops affecting complacency potential in Parasuraman et al.'s (2010) attention model.

1.7.5.1 The User's Learning Curve

Based on interactions with the tool over time, users tend to develop biographies for the tool, describing what it might be good and bad at (Hartswood, 2003). Some users also learn to ignore frequent false prompts, a phenomenon that is referred to as the *cry-wolf syndrome* (Meyer, 2001). Others learn to economise on the effort needed to dismiss a prompt by anticipating where prompts might occur (Hartswood, 1998) and then dismissing prompts of a certain pattern or frequency.

This learning process is, actually, not necessarily a long one. One study claims radiologists were able to predict where prompts would occur after only 20 cases (Hartswood, 1998). And the consistency in Zheng et al.'s (2001) results also suggests that users quickly adapted to different cuing rates.

It is a quick and possibly dangerous process that has not been well studied. For example, dealing with false prompts based on a user's biography of the tool may cause premature closure of a case, may make a user miss a target surrounded by false prompts, or may cause a user to miss a true prompt in an area habitually known to have false prompts (Hartswood, 1998).

1.7.6 Stratification of Results in the Analysis Stage

While the previous subsections of Section 1.7 are concerned with important factors that should be considered primarily during the design phase of an experiment, Alberdi et al. (2010) highlight a key consideration that should be implemented during the analysis phase. As seen in several studies in Section 1.1, looking only at the average effects of performance with and without the tool may mask important effects (Tsai, 2003; Berner, 2003; Goddard, 2014; Madhavan, 2007). It is important to dive deeper into the data by measuring positive and negative consultations. Furthermore, as in Alberdi et al.'s (2014) analysis, stratifying the results, for example by identifying "strata" of readers and cases within which tool effects vary less dramatically can be important in uncovering some of these hidden effects.

Alberdi et al. (2008) also highlighted another important factor to take into account during the analysis stage of AB experiments. Tools such as CAD are often evaluated in terms of users' recalls of cases. By looking instead at the individual features in the mammogram rather than the overall cases, Alberdi et al. were able to more closely identify the effects of individual prompts on radiologists' decisions. In fact, using this fine-grained analysis revealed that 13.5% of recall cases were recalls of cancer cases without correct identification of the cancerous features in the mammograms (Alberdi, 2008). Such "false recalls" may, after a biopsy in the wrong area of the breast, result in a false negative although treated in many studies, looking only at the overall recall rate, as true positives (Alberdi, 2008).

1.7.7 Reliability Modelling

It would be ideal to be able to predict how users will respond to a tool's behaviour or to predict overall system performance. One field where a similar process is implemented is "reliability modelling" in engineering "concerned with predicting probabilities of future failures" in man-made systems (Alberdi, 2003). In reliability modelling, engineers often develop models based on mathematical specifications to estimate the probability that a system fails as a function of the failure of its components (Alberdi, 2003). These models are helpful because they can also provide insight about which component is most critical for the dependability of the system or which component should receive the most improvement efforts (Alberdi, 2003).

Strigini et al. (2003) applied reliability modelling in an unprecedented case study to gain useful insights on some of the complex interactions in human + automated tool systems. For example, their reliability modelling showed how much one can improve overall system performance by only

improving the tool's performance. Their model also showed how reliability modelling can be used to choose classes of cases to target for improvement in a tool's algorithm. For example, assume designers are presented with two alternatives: "a reduction by 10 of the failure probability for the 'easy' (and frequent) cases, or a similar reduction limited to the 'difficult' (and rarer) cases" (Strigini, 2003). Such modelling as presented in Strigini et al.'s paper (2003) can help inform such complex decisions and thus help designers spend effort where it will be most beneficial.

1.8 Areas Identified in the Literature that Require Future Work

From the literature review presented in Sections 1.1 to 1.7, we identify several gaps in the topic of automation bias that warrant further research. These gaps motivate the work we present in Chapters 2 to 4 and are briefly discussed in this section. Some of these gaps represent ideas which have not been previously explored such as effects of the mere presence of tool advice, the homeostasis mechanism, and tailoring tool settings for users of varying ability and cases of varying difficulty. Other ideas are extensions of concepts discussed in other areas such as medicine or psychology and which we argue would be beneficial to extend or give more attention to when studying automation bias. Such ideas include the effect of the sequence of tool advice on automation bias and the benefits of diversity between tool and user capabilities on system performance.

The literature review we presented in Sections 1.1 to 1.7 covered three main areas: factors affecting automation use, explanations for automation bias, and mitigation strategies proposed to reduce the rate of AB. The gaps we identify and plan to tackle in our work address areas in each of these three main topics. From explanations of automation bias, we identify a need to consider causal mechanisms beyond complacency, overtrust, overreliance and attention deficits, specifically the homeostasis mechanism. Related to this is the effect of the mere *presence* of tool advice. These two topics are discussed in Section 1.8.1. From factors affecting automation use, we identify effects of personality traits and sequence of advice on automation use as two factors warranting further research. These are discussed in Section 1.8.2. Finally, on mitigation strategies proposed to reduce the rate of AB we identify a lack of strategies personalized to individual user needs and two such strategies are proposed in Section 1.8.3.

1.8.1 Gaps in the Literature Explaining Causes of AB

In Section 1.5, we presented several explanations for automation-induced errors including overreliance, heuristic processing, complacency and attentional effects. We argue that these explanations are incomplete and we identify a need to explore other causes for automation bias.

1.8.1.1 Exploring Explanations for AB Beyond Overreliance and Complacency

Conventional explanations for automation bias include overreliance (Lyell 2016; Goddard, 2012), overtrust, complacency (Bahner, 2008) and attentional effects (Parasuraman, 2010). However, as researchers highlight, these factors contribute to some but not all forms of automation bias (Parasuraman, 2010; Ho, 2005; Wickens, 2007). For example, they do not explain instances of automation bias that occur despite users' attentive behaviour or sufficient monitoring (Alberdi, 2008; Ho, 2005). Even where there is evidence in an experiment that AB is caused by one of the conventional explanations, as in Bahner et al.'s experiment linking AB to complacency, researchers still find that not all the errors in the study can be attributed to this single cause (for example, Bahner et al. reported that 20% of participants followed incorrect tool advice despite sampling sufficiently). In fact, Parasuraman et al. explicitly suggest a need for future research on the relative importance of attentional effects versus other explanations for automation bias (Parasuraman, 2010).

Moving towards an explanation that doesn't place a value judgment on the user, is a concept introduced by researchers at City University London as the *homeostasis mechanism*, and which we previously introduced in Section 1.5.8. This mechanism suggests that automation-induced errors may occur when users reduce the threshold with which they respond to cues (Alberdi, 2009). This idea is not well adopted in the literature as an explanation for automation bias and we believe it warrants further research, especially as it could account for some of the errors unexplained by conventional explanations for AB. Our focus on the homeostasis mechanism does not negate the validity of the other explanations, but adds to them. Future research would benefit from the relative importance of the different explanations, especially in relation to different contexts and mediators. A holistic understanding of cognitive mechanisms involved when using decision-support tools is important, especially when making design and protocol decisions aimed at effectively mitigating automation bias. *We explore the homeostasis mechanism as a potential cause of automation bias in Chapter 2 through an experiment in glaucoma treatment.*

1.8.1.2 Effect of the Mere *Presence* of Tool Advice

Closely related to the homeostasis mechanism is the effect of the mere *presence* of tool advice. This concept is not well studied in the analysis of decision support systems. As Alberdi et al. (2009) mention in their work on causes of false negative errors, presence of the tool can cause, among other effects, increased time pressure, increased cognitive load, abdicating responsibility

to the tool which in turn can lead to reduced search, explicit diagnostic misuse, and a raised diagnostic threshold. Research has repeatedly shown that automation does not merely supplant human activity but rather alters it – sometimes in unexpected ways (Azar, 1998). Although previous research acknowledges these changes, it seems to assume that these changes occur only when the user is presented with advice, while changes caused by the mere presence of the tool are mostly ignored. Sometimes changes due to the DSS may not be manifested in an actual change in the user's decision between their pre-advice/without advice and post-advice/with advice decision as measured in positive and negative consultations. Instead, the change may be more subtle in the user's behaviour in the pre-advice/without advice decision because of their knowledge that advice will be provided.

For DSSs to be truly effective, these subtle changes need to be better understood (Parasuraman, 1997b). As Parasuraman et al. (1997b) mention in the driver collision system detection context, a system may be less effective than anticipated because drivers with the tool may drive differently than without it. *Such studies looking at the effects of the mere presence of tool advice are scarce in the literature and we address this gap in Chapter 2 through an experiment in glaucoma treatment.*

1.8.2 Gaps in the Literature on Factors Affecting Automation Use

As seen in Section 1.4, factors affecting automation use represent the bulk of studies in the review. Sometimes results show discrepancies in results such as in effects of trust and experience. Thus, we attempt to study these factors again in our own work. However, two factors which we note did not receive much attention in previous studies are: the sequence of tool advice and personality traits.

1.8.2.1 Effects of the Sequence of Advice

In Section 1.4.3.5, we presented a discussion on the effects of the sequence of tool advice. Although there are a number of studies analysing this effect both in the medical literature and the JAS literature, fewer studies in the automation bias literature discuss the importance of sequence of advice, and in the latest reviews on automation bias, effects of sequence of advice were not mentioned (McBride, 2014; Goddard, 2012; Parasuraman, 2010).

Many of the studies in the medical and JAS literature studying these effects report only overall changes in system performance and those that consider individual details of the types of errors

committed in the two modes do not mention the significance or discuss the causes/cognitive mechanisms that could have caused these differences. Importantly, these studies do not take into account users of varying experience or cases of varying difficulty in the design of the experiment or the analysis of the results. Moreover, many of these studies do not discriminate between cases of correct versus incorrect tool advice. *We address some of these gaps in Chapter 2 through an experiment in glaucoma treatment.*

1.8.2.2 Effects of Personality Traits on Automation Bias

As discussed in Section 1.4.1.2, personality traits may contribute to individual differences in automation use. Ho et al. (2005) cited older users' risk-aversion character as a factor in determining reliance in their own abilities versus the tool's advice and Prinzel et al. found a negative correlation between automation complacency and self-efficacy (cited by Parasuraman, 2010). Overall, however, strong associations between personality traits and automation complacency have not been found, especially that the current database does not warrant decisive conclusions (Parasuraman, 2010). *In Chapter 2 we add to this small corpus of literature on personality traits and their effect on automation use by looking at the specific maximizer/satisficer personality trait and its association with automation use.*

1.8.3 Gaps in the Literature on Strategies Used to Mitigate AB

In Section 1.6, we presented numerous strategies used to reduce the rate of automation-induced errors such as providing training, creating effective protocols, and providing explanations of tool behaviour. However, we note a lack of strategies individualized to user needs and we address this gap by proposing two such strategies: tailoring tool settings for individual users and increasing diversity between tool and user capabilities.

1.8.3.1 Personalized Tool Settings

Future research in human factors and ergonomics aims to use knowledge of user differences to improve design and training efforts (Oron-Gilad, 2005)**.** Other studies aimed at tailoring the way information is presented to individual users according to their cognitive styles show enhanced performance and a promising future (Goddard, 2012b). As explicitly stated by Parasuraman et al. (2010), similar research in the area of automation complacency and bias would appear worthwhile.

Humans vary in experience, in criteria, and in characteristics – so how can we incorporate these variations into the design of a given DSS (Oron-Gilad, 2005)? This is a challenging human factors question (Parasuraman, 1997b). What is needed is adaptive automation where the tool adapts to the needs of the human *not* the human adapts to the ways of the tool (Van Dongen, 2013). For example, a tool that takes control when the user's response is low (Oron-Gilad, 2005). Or rear-end collision detection alarms that are individualized to respond to likely accidents for the pertinent driver age group – a novice driver would likely benefit from more false alarms, but a trained driver would likely prefer a tool with almost no false alarms (Parasuraman, 1997b).

Such mitigation strategies tailored to the specific needs of each user have not been thoroughly considered in the automation bias literature. *In Chapter 4 we propose one such strategy, tailoring tool setting (sensitivity/specificity combination) to user ability and case difficulty, and test its effectiveness using a spell checker tool designed to aid participants edit passages of text.*

1.8.3.2 Diversity Between Tool and User Capabilities to Mitigate Automation Bias

As discussed in Section 1.6.2.2, Alberdi et al. (2010) raise an important issue: in some systems, it is not the tool's overall reliability that may be most important, but the diversity between the tool and the user. Diversity is used to describe an intuitive notion of how focussed the automated tool is on the true needs of the users (Alberdi, 2010), or in other words, how different the human and tool errors are. Currently, many tools seem not to be very diverse from their users, suggesting that this, rather than the tool's stand-alone accuracy, is an area for improvement. True, some tool designers have already begun incorporating concepts of diversity in their designs. However, we argue that diversity, an accepted principle in many critical applications (Salako, 2014), deserves more attention in the automation bias literature as an effective mitigation strategy, especially as recent reviews in the automation bias literature do not address its potential benefits. *We present, in Chapter 3, a series of simulations aimed at revealing benefits that can be gained from applying diversity to human + human systems and how this can be extended to human + tool systems.*

Chapters 2 to 4 present the design, results, and discussion of three unique experiments aimed at addressing the gaps discussed in Section 1.8.

Chapter 2: Effects of the Presence and Sequence of Decision Support Advice in a Glaucoma Risk Calculator Environment

Glaucoma is a disease characterized by retinal ganglion cell death and subsequent optic nerve damage, and is the second leading cause of blindness worldwide (Takwoingi, 2014). A recent study demonstrated that treatment is effective in delaying or preventing the onset of glaucoma in ocular hypertensive patients (those with elevated eye pressure but no visual field loss) (Kass, 2005). However, identifying high risk patients that would benefit from early treatment is a difficult task due to the different patient variables involved. The glaucoma risk calculator is one tool designed to help address this difficulty by offering clinicians the probability that a patient will develop glaucoma within the next five years. Given the recent use of this tool in clinical practice, we designed a study to examine the effects of DSSs in this particular context with a special focus on the general cognitive mechanisms involved in such decision-making environments.

Motivation

Established explanations for automation-induced errors suggest "complacency", "overtrust" and "overreliance". We argue that these explanations are incomplete, and thus risk leading to ineffective remedies (changes to CDSS designs and protocols). We set up this experiment to test whether other cognitive mechanisms, specifically the homeostasis mechanism, could also be the cause of such errors. We previously introduced this mechanism in Section 1.5.8 where we described it as users reducing the threshold with which they respond to cues (Alberdi, 2009). This experiment also considers the effect of the sole *presence* of a decision support tool - even before presentation of tool advice, and the effect of varying the sequence of CDSS advice.

According to our knowledge, only two studies in the literature have specifically looked at the glaucoma risk calculator's effects on physicians' decisions (Boland, 2008; Boland, 2011). Thus, besides being an experiment to test our hypotheses in human-computer interaction, this experiment is also designed to yield interesting results from a medical viewpoint. We use more informative patient cases compared to previous publications in the field, which allow us to consider cases where the tool may be misleading.

Methods

We use a matched pairs experimental design that exposes participants to two of three of the following conditions on a set of similarly matched cases: the *Simultaneous* condition where participants see the

case and the risk estimate at the same time, the *Sequential-Know* condition where participants see the case and only see the risk estimate after giving their initial management decision, and the *Sequential-Not Know* condition where participants see all cases before being surprised at the end of the experiment with the additional information of the risk estimate for each case and a chance to change their initial decisions.

Participants are only asked for a management decision on each presented case. From this variable, we calculate other outputs such as diagnostic risk threshold, inconsistency, confidence, case difficulty, accuracy etc. We manipulate the presence of advice, the order of advice and knowledge that advice will be given. Other variables which we expect will show variation among participants (such as trust, specialty, experience, etc.), we only measure, but not manipulate.

Results

52 ophthalmologists and optometrists completed our online experiment. Results provided evidence of: (1) the effect of the presence of tool advice on user decisions, even before its presentation, (2) the newly introduced homeostasis mechanism which may account for some automation bias errors previously observed but unexplained in the literature, (3) the role (although weak) of uncertainty and direction of advice on changes in user decisions after the presentation of tool advice, and (4) differences between self-reported trust and perceived tool reliability and differences between objective measures of difficulty defined in terms of variance and subjective measures of difficulty defined in terms of uncertainty. Study results showed no significant differences in user decisions based on the sequence of presentation of advice, even when analyses incorporated error type. Study results also did not reveal differences in the amount of change in user decisions after being presented with tool advice based on the maximizer-satisficer personality trait.

From a medical viewpoint, the study provided evidence of: (1) strong disagreement with National Institute for Health and Care Excellence (NICE) guidelines even among UK professionals and even after use of the calculator, (2) a significant effect of the calculator on user decisions, increased user confidence, and increased user consistency, (3) differences in treatment recommendations based on location of practice but not according to specialty or profession, and (4) a lack of consideration for patient age and pattern standard deviation in clinicians' treatment decisions.

Discussion

These results have implications for future experiments analysing the efficacy of new decision support tools. Many of these studies compare the decision of a user before and after advice is presented. Our

results showed that the mere knowledge that advice will be presented can alter the cognitive mechanisms of the user and thus, a user's post-advice decision should be compared to a control group that does not receive advice nor knowledge of the presence of advice. Evidence of a different cognitive mechanism, the homeostasis mechanism, involved when using decision support tools also has important implications for managers and designers establishing suitable strategies to mitigate automation-induced errors as some of these protocols can serve to exacerbate the mechanism. Finally, highlighting user groups that are more susceptible to tool advice and showing how integration of a DSS into user workflow can affect decisions are also important in creating effective protocols for the use of DSSs. For the medical community, few studies have analysed the effect of the glaucoma risk calculator and its usefulness and this study adds to this small cohort.

Chapter Organization

Although this is a summary of the overall chapter, more details can be found in the pages that follow. Sections 2.1-2.5 provide a detailed literature review of the glaucoma risk calculator. Although much of the detail may seem quite specific, a thorough understanding of the medical research in this field was necessary, especially when selecting patient cases of varying difficulty, selecting cases which appear to contradict advice from the calculator, or comparing participants' responses to a specific standard of treatment. The literature review starts with an introduction of the main study in this field: the Ocular Hypertension Study (Section 2.1), then an introduction of the glaucoma risk calculator (Section 2.2) and its advantages and limitations (Section 2.3). Following this is a presentation of current treatment guidelines in the UK and internationally (Section 2.4), and finally a discussion of publications specifically looking at the effect of the glaucoma risk calculator on physicians' decisions (2.5). Starting from Section 2.6 is a documentation of the new experiment we designed in this context. Section 2.6 presents the experiment hypotheses. Section 2.7 reviews the experiment design in terms of procedure, variables, and resources. Section 2.8 presents the results. Finally, Section 2.9 highlights the experiment's contribution to existing knowledge.

2.1 The Ocular Hypertension Treatment Study

Glaucoma is a disease characterized by retinal ganglion cell death and subsequent optic nerve damage. In some patients, this optic nerve damage can cause severe visual dysfunction and even blindness. In fact, glaucoma is the second leading cause of blindness worldwide and is predicted to affect more than 79.6 million people by 2020 (Takwoingi, 2014). Unlike other ocular diseases, the decisions of when to diagnose a patient with glaucoma and when, or even whether, to begin treatment are not clear-cut decisions. Furthermore, the treatment itself is not a cure for glaucoma, but only a method to arrest, delay, or limit the disease's progression (Weinreb, 2004).

One of the most important studies in the 21st century contributing to the understanding of glaucoma prevention is the Ocular Hypertension Treatment Study (OHTS) where researchers evaluated the efficacy of topical ocular hypertensive medication in delaying or preventing the onset of glaucoma in patients with ocular hypertension (Kass, 2005). Ocular hypertension (OHT) is defined by medical professionals as elevated eye pressure in the absence of visual field loss. Approximately 3-5% of adults over 40 years have OHT, which is equivalent to an estimated one million people in England alone (Diagnosis and Management, 2009). Importantly, OHT is an established and leading risk factor for the development of POAG, and is presently the only modifiable risk factor (Kass, 2010; Takwoingi, 2014).

In the OHTS, researchers followed 1636 patients for five years and showed that the probability of developing POAG was reduced by 60% in patients who received treatment compared to those who were only monitored (Kass, 2005). The OHTS also showed that older age, higher intraocular pressure (IOP), greater pattern standard deviation, thinner central cornea, and larger vertical cup-to-disc ratio all had a significant predictive association with the risk of developing glaucoma (Kass, 2005).

2.1.1 Which Ocular Hypertensive Patients Should be Treated?

The OHTS demonstrated that treatment is effective in delaying or preventing the onset of POAG in ocular hypertensive patients. Based on this finding, clinicians may choose to treat all OHT patients. From a cost perspective, this would be too expensive (Stewart, 2008). Besides the high material cost, the potential benefit of treatment would have to outweigh the low conversion rate to POAG as well as the inconvenience and potential harmful effects of medication (Kass, 2005). Glaucoma treatment, once initiated, usually continues for the remainder of a patient's life and has been associated with increased risk of cataract (Kass, 2010; For the Screening, 2010), falls, mortality (Kass, 2010), hyperaemia, and ocular surface disease (Boland, 2008).

On the other extreme, clinicians may choose to withhold treatment for all OHT patients until early glaucomatous damage is detected. However, not all patients return for follow-up appointments and so the onset of the disease may be missed. More importantly, delayed treatment may allow the start of irreversible optic nerve deterioration (Kass, 2005).

If treating all OHT patients is not cost-effective (Stewart, 2008) and presents unnecessary risks to some patients while awaiting clear signs of the disease involves accepting some irreversible damage (Weinreb, 2004), then treating a selective subset of OHT patients could be a reasonable compromise. To decide to treat only moderate to high risk OHT patients, would require researchers to first understand whether there is a penalty associated with delaying treatment for low risk patients. This question was addressed in the second phase of the OHTS where researchers compared the efficacy of earlier versus later treatment in preventing POAG in OHT patients (Kass, 2010).

This second phase revealed that early medical treatment was most effective in high-risk patients, but showed little absolute benefit in low risk individuals (Kass, 2010). Furthermore, after initiating medication in the original observation group, the incidence of POAG rapidly decreased and approached the incidence in the medication group (Kass, 2010). These results raised a new question: how can clinicians distinguish high risk patients who are the ones most likely to benefit from early treatment?

2.2 The Glaucoma Risk Calculator

If a clinician is aware of the five risk factors identified by the OHTS and is convinced that treating only patients with the highest risk is most effective, the decision to treat or not treat a patient is still very difficult. As Mansberger (2006) explains, assuming that one divides the continuous variables of age, IOP, corneal thickness, and pattern standard deviation into thirds and uses nine different combinations for cup-to-disc ratio, there are 1458 different combinations describing a potential ocular hypertension patient. To further complicate the decision, the majority of ocular hypertensive patients have combinations of variables that increase or decrease their risk of developing glaucoma (Mansberger, 2006).

Subsequent work after the OHTS resulted in the development of a Scoring Tool for Assessing Risk (STAR) calculator, which is supported by Pfizer Inc. (Woloshin, 2003). Later, the OHTS-EGPS equation was developed using the pooled analysis from the OHTS and the European Glaucoma Prevention Study (EGPS, 2007) groups and has now superseded the OHTS equation (Takwoingi,

2014). The risk calculator which is based on this OHTS-EGPS model is available at http://ohts.wustl.edu/risk.4/. For a given patient, physicians input measurements for the five risk factors (age, corneal thickness, intraocular pressure, pattern standard deviation, and cup-to-disc ratio) and the calculator returns a risk estimate of this patient developing glaucoma within the next five years. A snapshot of the online calculator is included below for better visualization of the tool.

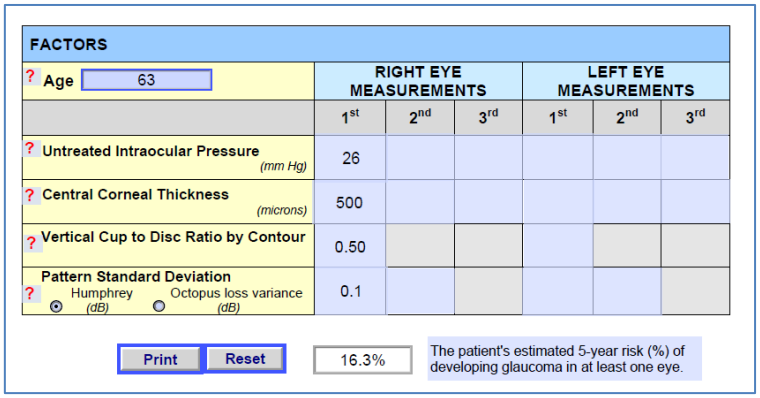


Figure 2-1: Snapshot of the OHTS-EGPS Calculator
(Available at http://ohts.wustl.edu/risk)

As Goni (2007) puts it, "the appearance of the glaucoma risk calculator is one of the most significant recent contributions to clinical ophthalmological research". It suggests a shift in the view of glaucoma treatment from simply lowering the IOP to a risk reduction process. Although new to ophthalmology, risk calculators are not new to medicine and have been previously used in deciding the risk of breast cancer, Down syndrome in babies, coronary artery disease, and osteoporosis (Mansberger, 2004; Medeiros, 2009).

2.3 Advantages and Disadvantages of the Glaucoma Risk Calculator

Like any decision support tool, the glaucoma risk calculator has advantages and disadvantages, and an understanding of these trade-offs gives researchers insight on scenarios where users may be more/less inclined to rely on the calculator's estimate.

Several advantages of the risk calculator have been noted in the literature. As stated by Mansberger (2004), the calculator simplifies the complex OHTS results and encourages the integration of evidence-based medicine into clinical practice. The ability of the calculator to give patient-tailored risk information based on personal factors is also important (Woloshin, 2003), especially that clinicians' ineffectiveness in estimating risk has been shown in previous publications in glaucoma as well as other areas of medicine (Mansberger, 2006; cited by Boland, 2011). As Health Behaviour research suggests, patient compliance increases when patients are involved in the decision and given a

clear expectation of risk, rather than a vague statement of "high" or "low" risk (Mansberger, 2008). This is especially important in a disease like glaucoma which is a lifetime disorder that could require lifetime treatment.

On a more specific level, clinicians might benefit from continuous values used by the calculator for difficult indices like pattern standard deviation (PSD). Clinicians mostly rely not on the absolute value of PSD, but on pre-set thresholds to determine whether it is abnormal. The risk estimator, on the other hand, calculates a higher risk for larger PSD values, including those that do not reach a pre-set value for abnormality (Boland, 2011).

However, these benefits do not come without drawbacks. Firstly, the calculator cannot be used to assess the risk of glaucoma *progression* or the risk of blindness (Woloshin, 2003). This is important because in the end, the ultimate goal in glaucoma care is to prevent loss of vision (Mithal), and only a small subset of patients who develop glaucoma end up losing functional vision during their lifetime (Weinreb, 2004). Most risk calculators also do not include confidence intervals (which are often quite large) and thus the calculators "can give a false impression of reliability in terms of prediction" (For the Screening, 2010).

Some physicians also express concern that use of the risk calculator could lead to rote treatment, as "it runs the risk of being used like risk calculators for systemic hypertension and coronary artery disease—namely, if the patient falls in the 'treat' category, then that's the standard of care" (Karmel). According to such opponents, the calculator poses a risk of "overtreatment".

2.3.1 Limitations Due to Important Factors in the Decision to Treat which are not Included in the Risk Estimate

Another important drawback of the risk calculator is that there are other important factors, some outside the discipline of eye care, which are not included in the risk calculator's estimate but which may predict risk or influence the decision to treat (Cheng, 2009). Social history, patient's willingness to treatment, fear of blindness (Mithal), risk aversion (Karmel), dislike of medication (Mithal), the impact of visual function on daily living and activities (POAG, 2013) (such as a patient who earns a living through driving), and the effect of the daily administration of medications on quality of life (Mansberger, 2004) (such as a patient who cannot easily administer drops) are all important factors to consider when examining a glaucoma suspect.

During examination, a detailed medical history is also important. As suggested by the American Academy of Ophthalmology Guidelines, physicians need to collect information regarding ocular

history, systemic history, and family history (POAG, 2013). For example, learning of any medications taken by the patient can affect the decision because some drugs, such as steroids, may cause IOP to increase, and others drugs, such as beta blockers, may lower IOP.

Older age is associated with greater risk of glaucoma progression. However, all other factors being equal, the shorter life expectancy of older patients also needs to be considered, and this is not factored into the calculator's risk estimate (Goni, 2007). In other words, the decision to treat an OHT patient should consider the possibility that an older patient may die before developing any glaucoma symptoms (Fingeret), while a younger patient will be exposed to the risk for much longer (Boland, 2011).

Safety of the medication, efficacy, and cost are all factors that also impact the management decision (Weinreb, 2004). These factors as well as the social, personal, medical, and life expectancy factors discussed in this section are not included in the calculator's risk estimate, and are important in the final decision, especially when the calculator determines a borderline risk (Mansberger, 2004).

2.3.2 Limitations Based on the OHTS-EGPS Sample

Another limitation of the calculator lies not in the absence of important factors from the risk estimate, but in the data used to develop the model it is based on. All predictive models are based on a certain population of patients selected using strict inclusion criteria and which may not be representative of the larger patient population. Thus, use of these models should be restricted to patients similar to those selected to be part of the study used to develop or validate the model (Medeiros, 2009), and it is vital to note that the OHTS sample has been cited as a "convenience-based", not "population-based", sample (Kass, 2010).

For the combined OHTS and EGPS analysis, patients met the following criteria: age between 30 to 80 years, IOP between 21 mm Hg and 32 mm Hg, best corrected visual acuity of 20/40 or better, normal automated visual fields, normal optic disks, and open angles on gonioscopy (Goni, 2007). The OHTS and EGPS excluded patients with a history of previous intraocular disease/surgery, patients with life-threatening or debilitating disease, patients with diabetes who had evidence of diabetic retinopathy, patients on medication, and pregnant or nursing women (Kass, 2005).

Thus, a risk calculator can be imprecise if used to calculate the risk of a patient who has conditions from the excluded persons' categories. It can also be inaccurate when a patient has a rare combination of characteristics included in the study population such as a cup-to-disc ratio less than

0.2, an IOP above 29mm Hg, or an age above 70 years, because the estimates in such cases would be based on a smaller sample size and thus have a larger confidence interval (Mansberger, 2004). Also, despite evidence that shows their association with increased risk of developing glaucoma, any IOP secondary to any disorder (pseudo-exfoliation, pigmentary dispersion, inflammation, traumatism, primary or secondary angle closure) was not included in the model (Goni, 2007).

Any predictive model derived from a particular sample is not guaranteed to work on a different sample. However, the model which the risk calculator is based on has been validated on several different populations. The European sample from the EGPS provided the strongest evidence of the generalizability of the OHTS results (OHTS & EGPS, 2007). Not only did analyses of the EGPS data identify the same predictors that increased the risk of POAG, but also the risk coefficients for each of these factors were not significantly different from those in the OHTS (OHTS & EGPS, 2007). The OHTS-EGPS prediction equation was also validated on an independent population of ocular hypertensive patients at the University of California San Diego (Medeiros, 2009) and four cohorts in the UK (Takwoingi, 2014).

These validation studies showed some variation in the performance of the OHTS equation in different patient samples, but that the calculator was still overall useful in identifying high risk patients and that future research "may seek to update the equation" (Takwoingi, 2014) to reflect the role of other risk factors and that further validation, especially in non-Caucasians is needed (Burr, 2012).

2.3.3 Limitations Due to Important Risk Factors Not Confirmed by the OHTS-EGPS Analysis

The exclusion criteria of patients poses limitations on the model, but also risk factors for ocular hypertension described in past studies have not all been consistent with the OHTS findings. Most importantly: family history, race, diabetes, gender and cardiovascular disease were identified as risk factors in previous studies but not in the OHTS (Kass, 2005; Stewart, 2008; For the Screening, 2010). Researchers attribute this to the power of the study and sometimes to methodological weaknesses in the study.

In conclusion, as presented in the various subsections of Section 2.3, a risk calculator based only the OHTS-EGPS results fails to include risk factors outside the discipline of eye care but influencing the decision to treat (Section 2.3.1), may exhibit limitations in populations different from that of the OHTS-EGPS studies (Section 2.3.2), and fails to incorporate the increased risk of other factors

demonstrated in previous studies (Section 2.3.3). The advantages and disadvantages of the glaucoma risk calculator presented in all subsections of Section 2.3 are summarized in Table 2-1 below.

Table 2-1: Advantages and Disadvantages of the Glaucoma Risk Calculator

Advantages	Disadvantages
Simplifies complex results	Does not assess risk of glaucoma progression or the risk of blindness
Integrates evidence-based medicine into clinical practice	Does not include confidence intervals so gives false impression of the reliability of the risk estimate
Generates patient-tailored risk	Could lead to rote treatment and thus overtreatment
Uses continuous values, rather than ranges, of risk factors to estimate risk	Does not include life expectancy, social, personal, and other medical factors that may affect the decision to treat (Section 2.3.1)
Can strengthen the physician-patient relationship and enhance compliance	Is based on a restrictive model that may not represent all patients (Sections 2.3.2 and 2.3.3)

2.4 Standards in Initiating Treatment for Glaucoma Suspects

In the American Academy of Ophthalmology guidelines on glaucoma suspects, the decision to begin treatment in a glaucoma suspect is defined as "complex and depend[ing] on ocular, systemic, medical, and psychosocial factors" (Cheng, 2009). In such a complex decision that affects a large group of individuals and exhibits considerable variations in management, it is useful to consider the stance of different healthcare guidelines. One of the main challenges we were faced with at the beginning of our experiment design was the absence of a golden truth for the treatment of ocular hypertensive patients. This led us to formulate many of our hypotheses and subsequent analyses independent of a golden truth. However, for a few hypotheses where it was difficult to avoid having a golden truth, it was essential to consider the stance of different healthcare guidelines in order to use these guidelines as a basis for the comparison of participants' decisions.

2.4.1 NICE Guidelines Followed in the UK

In the UK, healthcare professionals in the NHS are expected to follow the National Institute for Health and Care Excellence (NICE) clinical guidelines (Takwoingi, 2014). NICE Guidelines on the management of OHT patients cover what measurements and tests should be ordered at each visit and monitoring intervals for patients depending on their stage and risk. For example, "in the UK patients considered to have a moderate or high risk are often kept in secondary care" while those considered to have a low risk are monitored in primary care (optometry) (Takwoingi, 2014). Of most interest to us are the NICE guidelines on the *treatment* of OHT patients. Interestingly,

NICE guidelines suggest that patients with OHT should be offered treatment based on estimated risk of conversion to glaucoma using only IOP, CCT and age (Diagnosis and Management, 2009). Of note is the word "offered" likely because treatment and care must always take into account people's needs and preferences. Also of note is the use of only three of the five risk factors identified in the OHTS. This could be due to the same reasons previously discussed about the inclusion of PSD and vertical cup-to-disc ratio in the model despite some experts' argument that these factors are actually early indicators of glaucoma. Below is a copy of the NICE treatment guidelines based on IOP, CCT, and age.

Table 2-2: NICE Guidelines on the Treatment of OHT Patients
(Diagnosis and Management, 2009)

CCT	**More than 590 micrometres**		**555–590 micrometres**		**Less than 555 micrometres**		**Any**
Untreated IOP (mmHg)	> 21 to 25	> 25 to 32	> 21 to 25	> 25 to 32	> 21 to 25	> 25 to 32	> 32
Age (years)[a]	Any	Any	Any	Treat until 60	Treat until 65	Treat until 80	Any
Treatment	No treat-ment	No treat-ment	No treat-ment	BB[b]	PGA	PGA	PGA

The guidelines also mention that "treatment should not be routinely offered to people over the age threshold unless there are likely to be benefits from the treatment over an appropriate timescale" and that "once a person being treated for OHT reaches the age threshold for stopping treatment but has not developed glaucoma, healthcare professionals should discuss the option of stopping treatment" (Diagnosis and Management, 2009).

2.4.2 Global Standards on the Treatment of Glaucoma Suspects

Such guidelines are useful but exhibit differences across borders regarding the management of OHT patients. On a more global scale, a group of experts was asked the question: "At what risk should therapy be initiated for OHT?" The consensus was as follows (Karmel):

- Low risk is defined as less than 5%
    - Monitoring the patient is sufficient
- Moderate risk is defined as 5-15%
    - Treatment should be considered
- High risk is defined as over 15%
    - Recommendation for treatment

These cut-offs were subsequently used in other studies to identify patients for treatment (Mithal). However, Steven Kymes, member of the OHTS study group, suggested that 10% is the most cost-effective treatment threshold (Karmel).

Despite some of these thresholds, most experts agree that "there is no consensus on the appropriate level of risk for which to initiate treatment" (Cheng, 2009). The OHTS-EGPS calculator does not suggest a risk level at which to initiate treatment as developers "specifically tried to avoid telling people when to treat because there are so many exceptions" (Karmel). In fact, it is likely that each physician has his/her own treatment threshold, and some consider these attempts of a global threshold to be arbitrary.

An important and well-designed study that addressed the same need for a global standard is by Cheng et al. (2009). In this study, authors aimed to produce a tool for guiding clinicians when deciding whether to treat a given OHT patient. An 11-member diverse expert panel was appointed to rate the appropriateness of initiating treatment for 1800 cases through a two-round modified Delphi method. Factors used to define the cases were limited to: age, life expectancy, intraocular pressure, central corneal thickness, cup-to-disc ratio, disc size, and family history. Permutations of values of these variables according to numbers that reflect patients typically seen in clinical practice created 1800 unique cases, and the final ratings for cases were categorized as appropriate, uncertain, or inappropriate to treat after collapsing the initial 1-9 scale (Cheng, 2009).

Based on the panel's responses, authors produced a simple criterion for appropriateness of treatment: "IOP >26 mm Hg and <90% 5-year risk of mortality", and a simple criterion for inappropriateness of treatment: "IOP ≤23 mm Hg and cup/disc ratio <0.9". These criteria correctly distinguish most of the cases judged as appropriate or inappropriate by the panel, but authors also developed a point-system, which although more complex to apply, more closely conforms to the panel's decisions than these simple heuristics (Cheng, 2009).

2.5 The Effect of the Risk Calculator on Physicians' Decision-Making

Although risk calculators have been in use in medicine for many years, their impact on physician decision-making and on improving patient outcomes is still unclear. A recent study found minimal or no changes in physicians' management of patients with cardiovascular disease when they were provided with a Framingham-derived risk estimate (Weinreb, 2004). Another recent study found that doctors in an intensive care unit provided with a prediction from a risk model that suggested a patient would die within a week, did not change their treatment decision (cited by Mansberger, 2006).

Use of a risk assessment would likely differ between a critical care setting and an ophthalmology clinic, but these results show that simply providing a risk estimate does not guarantee it will be used by physicians (Mansberger, 2006). As Boland et al. (2010) state, few studies in the literature have looked at the glaucoma risk calculator and its effect on physicians' decisions. These studies, most relevant to the work we are proposing, are individually summarized in this section.

Cheng, Eric M., et al. "For which glaucoma suspects is it appropriate to initiate treatment?." *Ophthalmology* 116.4 (2009): 710-6.

The study by Cheng et al. (2009) was previously covered in Section 2.4.2 where standards for initiating treatment were discussed, but it also represents a role in the discussion of risk calculator effects. In this study (Cheng, 2009), a risk estimate was offered in 1/30 of the cases in the first discussion round and dropped in the second round of the Delphi discussion model. Cheng et al. mentioned that the calculator's risk estimate was dropped in the second round as they did not believe it had a strong influence on ratings in the first round. They also mentioned that during the discussion at the face-to-face meeting, they noticed that not all panellists used the calculator. Third, even among those who used the calculator, there was no consensus on the treatment risk threshold.

Authors also noticed contradictions between the panel's decisions and the calculator's risk estimate when IOP was set at the highest and lowest values. Two such examples are given:

- Patient of Age=62 years, IOP=22 mm Hg, C/D=0.3, CCT=500 μm, and PSD= 1.9 dB, and calculated risk=22.8% was rated by the panel as inappropriate to treat.
    - The panel stated that although lowering the IOP is the only available treatment (even for those without high levels of IOP), "the efficacy of lowering IOP in this range [which is lower than the OHTS-EGPS inclusion criteria] for preventing glaucoma development has not been established". The panel also stated that they may have misjudged the impact of the patient's low CCT.
- Patient of Age=50 years, IOP=30 mm Hg, C/D=0.3, CCT=600 μm, PSD=1.9dB, and calculated risk=6.5% was rated by the panel as appropriate to treat.
    - The high IOP value likely accounted for the panel's decision to treat as they stated that because IOP is the only modifiable risk factor, it is weighted heavily in the decision. In addition, the patient's younger age and thus prolonged exposure to the risk swayed their decision.

Mansberger, Steven L., and George A. Cioffi. "The probability of glaucoma from ocular hypertension determined by ophthalmologists in comparison to a risk calculator." *Journal of glaucoma* 15.5 (2006): 426-431.

The study by Mansberger et al. (2006) presented the OHTS results to ophthalmologists at an education meeting. Fifty-one ophthalmologists in the audience gave their estimate of the probability of developing glaucoma over a 5-year period for 4 hypothetical OHT patients and whether or not they would treat each patient.

For all four cases, there was a significant difference between the mean risk estimate by the ophthalmologists and the risk estimate from the calculator. More specifically, ophthalmologists tended to give lower risk estimates compared to the calculator, sometimes by a difference of 40%. The participants' estimates also showed high variability, with a minimum range of 0% to 20% in one patient, and a maximum range of 1% to 100% in another patient. Finally, even when ophthalmologists had similar risk estimates, they sometimes differed in their treatment decisions. This probably reflects differences in the individual ophthalmologists in terms of their preferences, training, and confidence in the OHTS study results (Mansberger, 2006).

These results should be interpreted in light of the fact that participants estimated the risk immediately after the presentation and with the aid of the hand-outs. This would decrease their chance of forgetting the study results and their predictive ability in a clinical setting would likely be much lower (Mansberger, 2006).

Boland, Michael V., Harry A. Quigley, and Harold P. Lehmann. "The impact of risk calculation on treatment recommendations made by glaucoma specialists in cases of ocular hypertension." *Journal of glaucoma* 17.8 (2008): 631-638.

Boland et al.'s (2008) goal was to determine whether a risk calculator changes treatment recommendations made by glaucoma specialists in cases of OHT. Participants were asked to review 100 cases and to make treatment recommendations on a 7-point scale. Half of the cases were presented with the calculator's risk estimate and the other matched half were presented without. Each case included information about the five risk factors as well as gender and diabetes status.

The results of this important study are summarized in bullet form below:

- 50% of participants reported using a risk calculator at least sometimes.

- The calculator's risk estimate was a statistically significant predictor of the treatment decision for 91% of physicians, and carried a higher normalized regression coefficient than all other variables combined (physician sex, race, length of practice, monthly patient volume, proportion of glaucoma patients in the practice, and current use of a risk calculator).
- Male physicians and those reporting an increased patient volume were less likely to recommend treatment.
- Asian, Black, or ''other'' race physicians and those reporting occasional use of the calculator were more likely to recommend treatment, while glaucoma specialists were less likely to recommend treatment compared to non-glaucoma specialists.
- In a third of the cases, physicians made a different decision with the calculator.
- The agreement between the decisions with and without advice "was only fair as measured by an average k statistic of 0.37" between the risk-matched pairs.
- Physicians were slightly more likely to recommend treatment with the calculator than without it. More specifically, the average physician treatment threshold was 23% without the calculator and decreased significantly to 17% with the calculator.
- Confidence (measured as percentage of sure responses) increased from 58% without the calculator to 66% with the calculator.
- Inconsistency (sum of the variances of risk in the Yes, No, and Unsure decisions for each physician) decreased from 0.04 without the calculator to 0.02 with the calculator.
- Of the five risk factors, the most significant in shaping physicians' treatment decisions were IOP (79% of physician models), cup-to-disc ratio (CDR) (71%), and corneal thickness (70%), but IOP and CDR were more heavily weighted than the OHTS would suggest.
- On the other hand, age and PSD were not significant in shaping physicians' decisions, despite their importance in the OHTS results.
- Physicians were less likely to treat older patients, although older age is associated with increased risk of glaucoma in the OHTS results, but this trend could be due to life expectancy considerations.
- When given the risk estimate, risk factors were incorporated in a manner more consistent (but not the same as) the OHTS results.

Boland et al. (2008) discussed that further research should evaluate the impact of risk calculators in clinical settings and on patient outcomes. They also stated that whether the recommendations are more ''correct'' and lead to better patient outcomes is an issue not addressed in the study. The study was based on fictitious cases, did not take into account patient preferences, and did not require that physicians assess the optic disc or interpret visual field test results – all factors that could affect treatment recommendations (Boland, 2008).

In a later study by Boland et al. (2011), authors explored whether glaucoma subspecialty training and formal risk estimation have an impact on physician treatment recommendations. Researchers found that non-glaucoma specialists were overall more likely to treat and that similar to their previous study in 2008, the risk threshold was lower when the risk calculation was made explicit - for both speciality groups. The effects of both calculator use and speciality were statistically significantly related to the calculated physician risk threshold. Results of these two studies helped us: predict some of the effects we might expect in our own study, understand how different participants may react, and have a useful starting point for some of our analyses.

To conclude, although previous studies report use, at least sometimes, of the glaucoma risk calculator by clinicians in their everyday practices, its impact on decision-making and patient outcomes is still unclear (Boland, 2011). Cheng et al. (2009) did not believe the calculator had a strong influence on user decisions; however, evidence from Boland et al.'s studies suggests that the calculator is capable of changing user's treatment decisions in almost a third of decisions. More specifically, their results, most relevant to our work, show that users with the calculator are: more likely to recommend treatment, more confident, more consistent with the OHTS results and more likely to treat higher risk patients. Researchers do not agree on the appropriate level of risk for which to initiate treatment, likely because of differences in user preferences, specialty, training, and confidence in the OHTS study results. However, researchers agree that further research would benefit from evaluating the impact of risk calculators in clinical rather than experimental settings, and taking into account factors not previously considered such as patient preferences, assessment of the optic disc or interpretation of visual field test results. Most importantly, future research would benefit from considering whether recommendations produced using the calculator result in better patient outcomes.

2.6 Experiment Hypotheses

Based on the previous work done in assessing the effects of the glaucoma risk calculator on physicians' decisions, we set up a matched pair experiment in this context to study effects of this decision support tool, especially when the sequence and presence of tool advice are varied. In order to present the specific hypotheses we aimed to address in this study, we need to start by defining our three experimental condition: the *Simultaneous (SIM)* condition where participants see the case and the risk estimate at the same time, the *Sequential-Know(SEQK)* condition where participants see the case and only see the risk estimate after giving their initial management decision, and the *Sequential-Not Know (SEQNK)* condition where participants see all cases before being surprised at the end of the experiment with the additional information of the risk estimate for each case and a chance to change their initial decisions. A more detailed description of these conditions is presented in Section 2.7.2. Each participant was exposed to two of the three conditions on a set of similarly matched cases. Also

important to note is that in the presentation of the hypotheses, *pre-advice* decisions refer to user decisions before seeing the calculator's risk estimate while *post-advice* decisions refer to user decisions after seeing the tool's advice.

The experiment is designed to answer two sets of hypotheses. The first set, presented in Section 2.6.1, is centred around human-computer interaction, and is a set of hypotheses we hope to be extendable to other contexts, in and outside the medical field. This set of hypotheses is mostly concerned with examining the psychological mechanisms at work when the sequence of DSS advice and knowledge of the presence of DSS advice are varied. The second set of hypotheses, presented in Section 2.6.2, is concerned with the medical aspect of the experiment and the effectiveness of the glaucoma risk calculator as a specific CDSS. We delay discussion of all further details of the experimental design to Section 2.7 after presentation of the hypotheses.

2.6.1 Human-Computer Interaction Hypotheses Addressed in the Study

This experiment is motivated by five main hypotheses, which we present in this section along with a short explanation of the reasoning and implications, where applicable, of each hypothesis.

Hypothesis 1 - Pre-Advice Sequential-Know Decision vs. Pre-Advice Sequential-Not Know Decision: Participants will be less likely to order treatments in their pre-advice decisions when they know they will be given a risk estimate (SEQK) than when they do not know (SEQNK). Thus, the diagnostic risk threshold of participants when they know they will be given support will be higher than their threshold when they do not know they will be receiving further support.

Reasoning: This hypothesis looks at whether user behaviour changes before seeing the actual tool advice and is based on the proposed homeostasis mechanism explained in Section 1.5.8. As a recap of that explanation, researchers at the Centre for Software Reliability at City University London describe cases where a false negative can occur without complacency or reduced vigilance, without trust in the tool, even without the user relying on the tool, and despite cautious protocols and training (Alberdi, 2009). Instead, they attribute the error to a simple change in the user's decision threshold. More specifically, users may reduce the threshold with which they respond to cues; for example, by not recalling a borderline case before seeing the tool's advice and committing to a decision (Alberdi, 2009). With the homeostasis mechanism, a tool could have an effect on a users' decision but this would not be observed by positive or negative consultations (where the correct diagnosis is not there before using the DSS but is there after and vice versa), but by differences in the pre-advice decisions between groups that know they have access to a tool compared to those who do not know.

Implications: The ultimate goal in studying automation bias is to mitigate, or ideally eliminate it. This requires researchers firstly to understand its causes. Researchers at the Centre for Software Reliability argue that established explanations for automation-induced errors in the literature suggesting "complacency", "overtrust" and "overreliance" are incomplete as they do not explain all instances of automation bias-induced errors. Thus, this hypothesis may offer evidence of a different cognitive mechanism causing automation bias.

Evidence of a different cognitive mechanism causing automation bias would better equip researchers attempting to design suitable mitigation strategies. As noted by Alberdi et al. (2009), the homeostasis mechanism is likely affected by both the need to keep the alarm rate below a certain limit and further exacerbated by protocols that do not allow users to change their pre-advice decision from target to no target. Thus, with evidence of the homeostasis mechanism, such protocols could be altered to help effectively mitigate automation bias effects caused by this specific mechanism.

Hypothesis 2 - Post-Advice Sequential-Know Decision vs. Post-Advice Sequential-Not Know Decision: The post-advice decisions in the sequential know and sequential don't know conditions will be significantly different.

Reasoning: We hypothesize that users who know they will receive advice will process data differently (for example, by changing the threshold with which they respond to cues) in the pre-advice decisions or react to advice from the tool differently (for example, give it a different weight) compared to those who do not know they will receive advice. This may or may not result in different management decisions in the pre-advice decision between the two conditions, which we check for in Hypothesis 1.

This second hypothesis addresses the possibility that differences between the two sequential conditions may not surface in the pre-advice management decisions. Then, showing a difference between the post-advice decisions between the two conditions could be an indication of differences in users' pre-advice processing of information. This is because if pre-advice decisions were the same, then we would expect that when presented with the same advice (the risk estimate), post-advice decisions would also be the same. An analogy that can help explain this reasoning is to consider two beakers filled with unknown liquids in a chemistry lab. Preliminary tests may suggest that the two beakers contain the same liquid (analogous to the checks suggested by Hypothesis 1). However, if we add the same substance (analogous to presenting tool advice) to both beakers and notice different reactions so that now the liquid in each beaker is, for example, a

different colour (analogous to different post-advice decision), then based on these different results, we could conclude that the initial liquid in the two beakers was in fact different.

Implications: This hypothesis is important because it shows that the presence of a decision support tool can affect user decisions even before its use. This is critical in the design of studies that measure the efficacy of a DSS based on a comparison of pre and post advice decisions in the sequential-know set up, when in fact comparison between a sequential-not know group and the post-advice decision would give more reliable results.

Hypothesis 3 - Changes in Sequential-Know vs. Sequential-Not Know Decisions: There will be fewer changes in users' management decisions after seeing the risk estimate in the group that doesn't know they will receive support compared to the group that knows they will receive support. Thus, more positive and negative consultations are expected in the latter condition.

Reasoning: Again, this hypothesis focuses on how the mere presence of a decision support tool can change the pre-processing of a user. The rationale is that if, as suggested by previous researchers, participants may exhibit overtrust, complacency, and overreliance when using DSSs, then the dominance of these effects would result in those aware of the presence of the tool being more willing to change their decision, less anchored to their pre-advice decision, and more easily swayed by the tool's advice. A result in the opposite direction may suggest that user behaviour contradicts these explanations and that other mechanisms may in fact be in effect.

Hypothesis 4 - Factors Affecting the Magnitude of Changes in Users' Decisions: The probability that a user will change his/her pre-advice decision increases:

a. When the pre-advice decision is uncertain
b. If change is towards the "safer" side (change from "not treating" to "treating")
c. For more difficult cases
d. For users that report more trust in the calculator
e. For less experienced users
f. For non-glaucoma specialists
g. For maximizers more than satisficers

Reasoning: Much of the reasoning in this hypothesis is based on common sense and on results from previous publications which showed that users are more likely to accept DSS advice when they exhibit uncertainty (Hartswood, 1998, Bonaccio, 2006, Dreiseitl, 2005; Goddard, 2014), when they are less experienced (Dreiseitl, 2005), for more difficult tasks (Goddard, 2012b), and when they trust the tool more (de Vries, 2003).

However, we note that reports on the effects of experience (Balleyguier, 2005) and trust (Parasuraman, 2010) show discrepancies between study results, thus warranting further exploration. Furthermore, Driestel et al. (2005) found no tendency to follow tool suggestions more willingly if they led to a "safer" decision (recalling a skin cancer suspect for further investigation) - an unexpected result, again warranting further exploration. We note that a "safer" decision does not always represent a correct decision, but that false positive errors in such a context generally represent less drastic penalties compared to false negative errors. Hence, we expect that healthcare providers would be more inclined to accept advice if it is towards this "safer" side of treating.

We also expect that users with maximizer personalities will be more likely to change their decisions, as by definition, these are people who are constantly in search of more information to improve their decisions, which they are never entirely satisfied with. This personality trait's association with automation use has not been previously studied in the literature.

Implications: With this hypothesis we are aiming to highlight groups of users that are more susceptible to tool advice and thus automation bias. This information is especially important for managers and designers when writing protocols on DSS use. Furthermore, strong associations between personality traits and automation bias have not been found (Parasuraman, 2010), which would make such a result interesting in the literature.

Hypothesis 5 - Simultaneous Decision vs. Post-Advice Sequential Decision when calculator is misleading: When the risk calculator gives "misleading advice", users are more likely to err if the risk estimate is presented simultaneously (SIM) rather than sequentially (SEQK and SEQNK). Thus, automation bias is more pronounced in the simultaneous mode.

Reasoning: The rationale behind this hypothesis is that when tool advice is presented simultaneously, it becomes a part of the evidence that the participant uses to form their decision and thus we expect it to carry a higher weight in the decision than when presented after a provisional decision in which case it becomes new evidence that must challenge the already existing decision the user has come to after reasoning through the data him/herself.

Implications: Such a result would prove useful when setting up protocols involving proper use of a DSS as it would show that the sequence in which advice is presented affects the rate of automation bias-induced errors. Thus, this may encourage designers and managers to encourage sequential rather than simultaneous presentation of tool advice where appropriate.

2.6.2 Context-Specific and Medically-Related Explorations Addressed in the Study

In this section is a list of the questions we plan to address in our analysis of the experimental results, but which are specific to the glaucoma risk calculator as a specific CDSS and most relevant to researchers in the medical community. These questions are more exploratory than hypothesis-based, thus we define them as "explorations". Also, some of them are inspired from discussions with our two co-investigators Dr. Waleed Tantawy, an ophthalmology consultant at West Wales Hospital, and Professor John Lawrenson, a Professor at the School of Optometry at City University London, which we gratefully thank for their invaluable input, their role in our understanding of the medical information and in ensuring the ecological validity of the experiment.

1. What do participants more strongly conform to: NICE guidelines or the expert panel's decisions (Cheng, 2009)?

2. Does use of the calculator make doctors better or worse where better or worse is defined according to:
    a. More closely conform to NICE guidelines
    b. More closely conform to expert panel's decisions
    c. More consistent in ordering treatment for higher risk patients

    This question attempts to takes Boland et al.'s (2008) results one step further by exploring, as they described in their future work, "whether the recommendations are more 'correct' and lead to better patient outcomes". However, as we previously discussed in Section 2.4, one of the challenges of this context is that there is no "golden truth" or unanimously correct decision for each patient case. This is why we avoid the need for such a "golden truth" in most of our analyses, but in certain explorations such as this where it is unavoidable, we consider several reasonable definitions of the "truth".

3. Does use of the calculator:
    a. increase confidence
    b. decrease intra-participant inconsistency
    c. increase the likelihood to recommend treatment

This is a replication of Boland et al.'s (2008, 2011) questions with a new set of cases. Unlike Boland et al., cases now include information regarding family history, life expectancy, and disc size in an effort to create a more realistic clinical setting. This also allows us to observe effects of offering users more information than the calculator, and behaviour in instances where calculator advice could be considered misleading.

4. Are glaucoma specialists less likely to recommend treatment compared to non-glaucoma specialists? How do management decisions differ between optometrists and ophthalmologists - if in fact there is a difference?

5. Does the calculator truly encourage physicians to consider all five patient variables as claimed, or is this a step they already perform in their clinics?

   One of the stated benefits of the glaucoma risk calculator is that it helps integrate evidence-based medicine into clinical practice as the estimate incorporates the risk from five factors found, through research, to affect the onset of glaucoma. We are interested in comparing whether clinicians already consider these factors in their decision

2.7 Experimental Design

In this section, we provide a detailed description of the experimental design we used to test the hypotheses presented in Section 2.6. We start with an overall view of the experimental procedure (Section 2.7.1), experimental conditions, (Section 2.7.2), variables and outputs (Section 2.7.3), participant and case resources (Section 2.7.4), and finally design issues (Section 2.7.5).

This design received final ethical approval from the Proportionate Ethics Review Committee in the Division of Optometry at City University London in November 2015.

2.7.1 Experimental Procedure

The view of the experiment from a participant's perspective is summarized in the user journey depicted below. Rectangles are actions taken by the user while eclipses are actions taken by the system. Only normal flow through the system is represented while exceptions are excluded. Each of these steps are explained, in chronological order, in the remainder of this section.

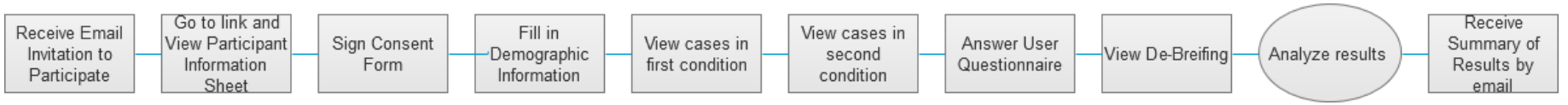


Figure 2-2: User Journey Representing Normal Flow Through the Glaucoma Experiment

Potential participants are first sent an invitation via email to participate in the study. At the end of the email are the Participant Information Sheet and a link that interested participants can click to participate (see Appendix 2A for a copy of the invitation email and Participant Information Sheet).

Participants who click the participation link are led directly to the welcome screen of the experiment which is hosted on the Qualtrics platform. This screen is followed by a consent form (see Appendix 2B for a copy of the Consent Form).

Participants that provide their consent are first asked some demographic questions to collect information about their:

- Email address
    - Inputted twice in case of errors
- Gender
- Profession
    - Ophthalmology or optometry
- Primary area of specialty
    - Cornea, paediatrics, cataract, glaucoma, retina, oculoplasty, general ophthalmology, or other
- Number of years in practice, Average number of patients (in total) they currently see per week, and Number of those patients who are glaucoma patients or glaucoma suspects
    - We define a glaucoma suspect to participants as "a patient at risk of developing glaucoma, but who currently does not exhibit glaucomatous optic neuropathy and visual field loss".
- Country of practice
    - Important when comparing conformance with NICE guidelines
- Factors they generally consider before offering treatment to a glaucoma suspect
    - Age, intraocular pressure, central corneal thickness, cup-to-disc ratio, pattern standard deviation, family history, life expectancy, or other
- The risk estimate at which they consider it appropriate to initiate treatment for a given patient

Next, participants are briefed about the cases, the calculator, and the management scale used in the experiment. This briefing is carefully worded to be as precise, as clear, and as uninfluential as possible, in terms of calculator effects, while still providing enough information about the calculator, especially for users who may have never heard of or used it before. The exact wording of this briefing is copied below:

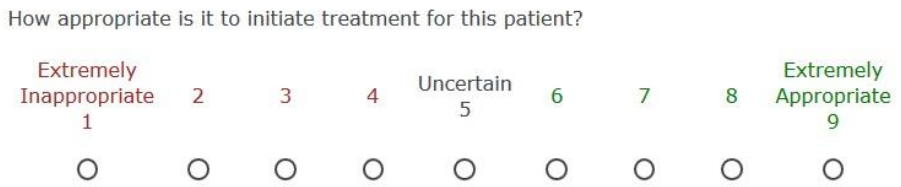
> You will now be shown ten cases in one condition followed by another ten cases in another condition. Cases are characterized by patient information which may include:
>
> 1. Age
> 2. Intraocular pressure (IOP)
> 3. Central corneal thickness (CCT)
> 4. Cup-to-disc ratio (CDR)
> 5. Pattern standard deviation (PSD) with Humphrey measurement

6. Family history
7. Disc size described as small, medium or large
8. Life expectancy categorized as normal or >90% chance of death within 5 years

Cases may also include a patient's estimated risk of developing glaucoma in the next five years.

- This risk estimate is obtained from the glaucoma five year risk estimator available at http://ohts.wustl.edu/risk/calculator.html.
- The estimator uses age, IOP, PSD, CCT, and CDR to calculate the risk of developing glaucoma.
- The risk estimate is derived from results of the Ocular Hypertension Treatment Study (OHTS) and the European Glaucoma Prevention Study (EGPS).
- Participants in these studies were aged between 30 to 80 years; IOP 20-32 mm Hg; CDR 0-0.8; CCT 465-650 microns. Thus, estimates are more likely to be accurate for patients who are similar to these participants. Any value outside these ranges causes the estimator to issue a warning and to round the value outside the range down to the nearest appropriate value.
- "There is no guarantee that the estimated risk is accurate for individual patients" and it is "designed to aid but not replace clinical judgment".

For each case, we ask you to rate the appropriateness of initiating patient treatment on a 9-point scale.



We define “appropriate” to mean that the expected health benefits of treatment (reduction in risk of developing glaucoma, reduction in risk of developing complications from glaucoma, reduction in anxiety while untreated, etc.) exceed the expected negative consequences (adverse events, increase in anxiety while under treatment, difficulty in administering eye drops, etc.). A rating of "uncertain" means that risks and benefits are evenly balanced, or there is not enough information to judge.

Following this briefing, participants are randomly directed to one of the three experimental conditions: Simultaneous, Sequential-Know, and Sequential-Not Know.

After seeing cases in one condition, each participant then sees a similarly matched set of cases in one of the other two conditions. Thus, each participant is randomly assigned to one of the following groups. In all groups, order of cases, order of conditions and sets of cases are randomized to avoid ordering effects. Randomization is discussed in more detail in Section 2.7.5.6.

- *Group 1:* Participant is exposed to the Simultaneous and Sequential-Know conditions
- *Group 2:* Participant is exposed to the Simultaneous and Sequential-Not Know conditions
- *Group 3:* Participant is exposed to the Sequential-Not Know and Sequential-Know conditions

Each condition contains an instruction screen complete with a visual flowchart of the instructions to ease understanding, followed by ten randomized patient cases.

Finally, at the end of the experiment, there is a short questionnaire designed to measure factors such as trust and knowledge that may affect users' decisions or their use of the tool. Using a 5-point Likert-scale (where appropriate), we ask participants about:

- Their familiarity with:
    - the Ocular Hypertension Study
    - NICE guidelines on the management of glaucoma suspects
    - the rationale of the risk estimate provided by the glaucoma risk estimator
- Their agreement with the following statements:
    - The glaucoma risk estimator is a reliable tool
    - I trust the glaucoma risk estimator
- The percentage of glaucoma suspects they choose to initiate treatment for in their clinic
- How often they use a glaucoma risk estimator in their practice
- At what risk estimate they consider it appropriate to initiate treatment for a given patient
    - This same question is also asked at the beginning of the study to determine whether this threshold changes due to learning effects
- Whether they are satisficers or maximizers (i.e., do they generally make a decision once an acceptability threshold is met versus they attempt to find the best decision available so can only make a decision once they've examined all information)
    - We determine this tendency using a standardised questionnaire from the psychology literature. The 6-item maximization scale (Nenkov, 2008) we used is a reduced and validated version of the original 13-item maximization scale suggested by Schwartz

et al. (2002) and was tested on 1700 participants. Based on users' average response on a 7-point Likert scale to six questions, maximizers are categorised as having an average response > 4 and satisficers as < 4. The questions in the 6-item scale we used are included in Appendix 2C for reference.

Finally, participants are thanked and de-briefed. After the analysis is complete, and as an incentive for their participation, participants are sent a report outlining how their decisions compared to their peers' decisions and how their decisions compared to NICE guidelines and expert opinion.

2.7.2 Experimental Conditions

Each participant is exposed to two of the three conditions described in Table 2-3.

Table 2-3: Experimental Conditions

Experimental Condition		Description of Condition
Condition SIM	Simultaneous	For each case, the patient information and the calculator's risk estimate are presented to the participant in the same screen and the participant is asked for their management decision.
Condition SEQK	Sequential-Know	For each case, the participant sees the patient information and a button to show the calculator's risk estimate. This button is only activated after the participant inputs their management decision based on the given patient information. When the management decision is given and the button is pressed, the calculator's risk estimate is shown (along with the patient information and the participant's pre-advice decision) and the participant is asked for their management decision using this new information. Their new management decision may be the same as their previous decision, or different in either direction.
Condition SEQNK	Sequential-Not Know	Each participant goes through all the cases and gives his/her management decision for each case using only the patient information presented in each case, and without knowledge that any further information will be given. At the end of all the cases, the participant is shown each case again with the calculator's risk estimate included (along with the patient information and the participant's pre-advice decision) and allowed to change his/her decision in either direction.

2.7.3 Experiment Variables and Outputs

In this experiment, there are a number of variables that are likely to affect the results. We manipulate only three of these variables:

- Presence of advice
- Sequence of advice
- Knowledge that advice will be given

In the experiment, we collect a number of key outputs. The primary output is the management decision (pre and post for the sequential conditions) for each case. Other secondary outputs, mostly from the questionnaires at the beginning and end of the experiment, are listed below. These are factors which we expect could affect management decisions and which we only measure, but not manipulate.

- Gender
- Profession
- Primary area of specialty
- Experience based on: number of years in practice, average number of patients (in total) they currently see per week, number of those patients who are glaucoma patients or glaucoma suspects
- Country of practice
- Factors they generally consider before offering treatment to a glaucoma suspect
- The risk estimate at which they consider it appropriate to initiate treatment for a given patient (both at the beginning and end of the study)
- Time elapsed before a participant makes each management decision
- Awareness of the calculator's rationale, NICE guidelines, and the OHTS study
- Self-reported trust and perceived reliability of the glaucoma risk estimator
- Percentage of glaucoma suspects they choose to initiate treatment for in their clinic
- Previous use of the calculator
- Score on the maximizer/satisficer personality test

Using the above outputs, we calculate/statistically infer a number of interesting measures. Where applicable, each of these measures are calculated for the five different decisions: the Simultaneous condition and the pre and post-advice decisions in the Sequential-Know and Sequential-Not Know conditions:

- Diagnostic risk threshold
    - Using ordinal regression and the method proposed by Boland et al. (2010), we derive the risk threshold at which participants are equally likely to make a Treat or Don't Treat recommendation (i.e., are maximally unsure).
- Confidence
    - Percentage of sure (Treat/Don't Treat) responses

- Inconsistency
    - As previously measured by Boland et al. (2008) as the sum of variances of the calculator's risk estimate of cases in each participant's three treatment groups (Treat/Unsure/Don't Treat).
- Case difficulty:
    - Subjective: from the uncertainty of the responses for that case
    - Objective: from the variance of the responses for that case
- Experience:
    - Years of experience combined with clinical volume and proportion of glaucoma patients in the practice
- Difference in treatment decisions on the Likert scale between the pre-advice and post-advice decisions in all sequential conditions
- Accuracy (Sensitivity/Specificity) of decisions where accuracy is measured according to two possible standards:
    - NICE treatment guidelines
    - Expert panel decisions (Cheng, 2009)
- Percentage of times participants switched their treatment decisions after seeing the risk calculator's estimate
- Weight of each of the patient factors in shaping the treatment decision
    - Using ordinal regression

2.7.4 Experimental Resources: Participants and Cases

For this experiment two main sets of resources are required: cases and participants.

2.7.4.1 Experiment Participants: Inclusion Criteria and Recruiting

Participants are limited to practising optometrists or ophthalmologists. They do not necessarily have to be NHS staff and may be community optometrists or practising outside the UK. Participants do not have to be glaucoma specialists although these differences in the participant population are factored into some analyses.

The Division of Optometry and Visual Science at City University has previously worked with community and hospital optometrists throughout England and has strong research links with Moorfields Eye Hospital. Through these contacts, we distributed an email invitation to potential participants. Dr. Tantawy, an ophthalmologist in West Wales Hospital, and who is

also involved in the project, also recruited many ophthalmologists by sending email invitations to his contacts/mailing lists. We are thankful to both our co-investigators, Dr. Waleed Tantawy and Professor John Lawrenson, for their invaluable role in helping us recruit participants.

Participants were not provided with any monetary incentives. However, after the end of the experiment and when the bulk of the analysis is complete, they will be sent a summary of their treatment decisions for each unique case alongside an anonymized distribution of other participants' decisions and the decision based on NICE guidelines as well as expert opinion (Cheng, 2009). Speaking with potential participants has confirmed our belief that this report could be viewed as a useful incentive for participants as it provides them with an opportunity to calibrate and compare their decisions as well as reflect on the overall treatment of OHT patients.

2.7.4.2 Choice of Patient Cases Used in the Experiment

Previous publications in this area have all relied on fictitious cases (Mansberger, 2004; Boland, 2008; Cheng, 2009; Boland, 2011). We too decided to use fictitious cases as real cases complicate ethical approval and do not significantly enrich the results. We primarily relied on Cheng et al.'s study (2009). Cheng's cases contain variables identified by their international panel of experts as being associated with the risk of developing glaucoma, and cover, for each of these variables, all such permutations of the values that, according to the panel, reflect patients typically seen in practice. Each of the 1800 cases published in the appendix of this publication includes a value for: age, life expectancy, intraocular pressure, central corneal thickness, cup-to-disc ratio, disc size, and family history. The publication also includes a decision of the panel on each case including the distribution of panel decisions and their mean deviation from the median rating. The decision of the expert panel provides a useful baseline with which to compare treatment decisions to, especially for participants outside the UK where NICE guidelines may not be relevant.

Our aim was to choose a careful subset of these cases which would best address our hypotheses. The choice of cases was done in collaboration with project members whose medical experience was insightful to the decision: Dr. Tantawy and also Professor Lawrenson. We reduced the 1800 cases to ten cases using the following method:

1. Decisions of Cheng's panel are either: Appropriate to Treat, Inappropriate to Treat, or Uncertain. To start with, all uncertain cases, where the panel did not have a majority decision, were filtered out. Although these uncertain cases were where we could expect

the most changes due to tool advice, they did not contain an international standard to compare participants' decisions to (both in the analysis and in the incentive report provided to participants). Instead, we focussed on cases with the greatest uncertainty (measured as the highest mean deviation from the median rating of panel decisions on each case), but which contained a majority decision by the panel.

2. NICE Guidelines are available for all cases with IOP between 21 and 32. A fifth of Cheng's cases have IOP=19. These too were filtered out as there were no NICE guidelines for their treatment.
3. The risk calculator does not give an explicit recommendation to treat/not treat. However, several publications cite risks <5% as low and not warranting treatment, and >15% as high and warranting treatment (some use a cut-off of 10% but we chose the greater cut-off). Thus, all cases whose risk was between 5-15% were omitted as the calculator's advice was considered indecisive.

The remaining cases then represented the subset of cases where a clear decision for all three references (expert panel, NICE, and the calculator) existed. However, they represented a range of possible types summarized in Table 2-4.

Table 2-4: Types of Cases				
NICE & Expert Panel	**Calculator's Risk**	**Case Difficulty**	**Priority of Including Cases in the Experiment**	**Considerations**
Agree	Contradicts NICE and expert panel	Easy	Easy cases where calculator's advice contradicts both standards. Participants may ignore the tool, change their decision and/or adjust their trust in the tool	40 cases All 85 years Chose high risk cases to create the most contradiction
Agree	Contradicts NICE and expert panel	Difficult	Difficult cases where advice contradicts both standards are where we expect to see the most changes in decisions due to tool advice	87 cases All 85 years
Agree	Agrees with NICE and expert panel	Easy	Not important cases because we don't expect many changes in easy cases where advice agrees with standards	--
Agree	Agrees with NICE and expert panel	Difficult	May cause some people who were confused at the start due to the difficulty of the case to more closely conform to the standards	148 cases Chose cases with very low risk (4%) as these were the only examples of "No Treat" by the calculator in our sample
Disagree	Agrees with expert panel	Easy	Good for challenging NICE guidelines, especially cases where NICE may have overlooked the effects of important variables outside its guidelines of 3 variables	99 cases Chose high risk cases with family history and a large CDR (two patient variables affecting risk, but not considered in NICE's guidelines)
Disagree	Agrees with expert panel	Difficult	Again good for challenging NICE guidelines. Also, interesting to see if people who share the opinion of NICE guidelines are swayed by the contradictory advice (such instances are more likely to be with difficult cases)	138 cases Chose high risk cases with family history and a large CDR (two patient variables affecting risk, but not considered in NICE's guidelines)
Disagree	Agrees with NICE	Easy	Interesting to see if people who share the expert panel's decision are swayed by advice to better conform to NICE guidelines	11 cases All 62 years Chose high risk cases with varying variables
Disagree	Agrees with NICE	Difficult	Interesting to see if people who share the expert panel's decision are swayed by advice to better conform to NICE guidelines	48 cases Chose high risk cases for greatest effect

We objectively judged the difficulty of each case based on the mean absolute deviation from the median rating of the expert panel for that case. This deviation increases with greater variance in the expert panel's decisions. We also subjectively judged difficulty by discussing the cases with Dr.Waleed Tantawy and using his judgement. To get a good mix of cases that address our hypotheses, we chose at least one case from each of the rows above (besides the third row) while taking into account the insights in the "Considerations" column.

We were sure to include at least one case from each possible value of the six variables and to get a balanced representation of each of the ranges. For example, according to the expert panel's view of values typically seen in clinical practice, CCT in the cases defined by Cheng et al. (2009) could have a value of 500, 550, or 600 μm. We included at least one case from each of these three categories, and similarly included cases for each of the possible values of the other five variables. The final set of cases was seen by both Dr. Waleed Tantawy and Professor John Lawrenson who both commented on the difficulty and variety of the cases from a medical viewpoint. For reference, the cases are included in Appendix 2D while the averages and standard deviations are included in the table below.

Table 2-5: Statistics Summarizing Cases Chosen

Patient Factor	Mean ± SD (N=10)
Age (years)	68 ± 15
IOP (mm Hg)	26 ± 3
CCT (μm)	535 ± 45
CDR	0.6 ± 0.1
PSD (dB)	1.8 ± 0.4
Risk	40 ± 27
Family History (% with)	50%
Disc Size (% with small, medium, large)	30%, 20%, 50%
Life Expectancy (% with abnormal life expectancy)	30%

As the experiment is based on a matched pairs design, we needed, in order to compare user performance in the different conditions, to present subjects with a set of cases in the second condition which match the cases in the first condition. As subjects were performing the experiment in a single setting, these cases could not be exactly the same to avoid learning effects. Instead, for each case, we manipulated each risk factor based on its marginal contribution to the risk equation while maintaining the overall risk of the case. Boland et al. (2008, 2011) achieved this by increasing or decreasing the CDR by 0.1. If the CDR was increased, then age was decreased by 5 years, IOP was reduced by 1 mm Hg, PSD was decreased by 0.1 dB, and CCT was decreased by 4mm. If CDR was decreased then the changes

were reversed. We followed Boland et al.'s lead while maintaining life-expectancy, family history, and disc size between the two sets of cases.

Unfortunately, the cases published by Cheng et al. did not include a value for PSD, but PSD is one of the risk factors used by the calculator to obtain a risk estimate. Thus, we had to include PSD in order to compare decisions with and without the calculator, or else there would be a flaw in the experiment design in that more information would be provided to participants in conditions including a risk estimate compared to those without a risk estimate. In Cheng et al.'s first round of Delphi models, when a risk estimate was included, authors used the average PSD value from the OHTS. We decided against this because besides the implications of providing the same PSD for each case, we were also concerned that this would push borderline cases towards a certain decision. Instead, to resolve this issue, we used the average and standard deviation PSD values from the OHTS but used a PSD equal to (average – SD) for cases where the decision by the expert panel was to not treat and (average + SD) for cases where the decision by the expert panel was to treat. In this way, we attempted to resolve the issue but by providing varying PSD values that we aimed would help decisions, especially for borderline cases, converge to those of the expert panel.

2.7.5 Experimental Design Issues

In designing the glaucoma experiment, there were several challenging design questions which resulted in much discussion. We use this section to present these issues and capture some of the discussion, experience, and reasoning behind our resulting decisions, which many times were a difficult balance between the different trade-offs.

2.7.5.1 Choice of Likert Scale

The size of the Likert scale is a common concern in many similar studies and each size has its advantages and disadvantages as summarized in Table 2-6. We chose the 9-point scale mostly: (1) because we expect changes due to our manipulative variables such as sequence and presence of tool advice to be very small changes and we do not want to mask these important effects by choosing a coarser scale, and (2) to match Cheng et al.'s publication as we use it as a reference to compare users' decisions to. Furthermore, we keep the neutral option in the Likert scale again to conform to previous publications in the field, but more importantly because the unsure option allows us to calculate the empiric risk threshold and confidence level in user recommendations – two important measures in the analysis.

Table 2-6: Comparison of Options for the Likert Scale	
Scale	**Advantages and Disadvantages**
5 point scale	Clear differences between the categories
7 point scale	Matches Boland et al.'s study Produces better distribution (normality) of the data compared to smaller 5-point scale, especially if we plan to use it as a continuous scale Allows us to detect smaller changes, but categories become more difficult to distinguish compared to a 5-point scale.
9 point scale	Matches Cheng et al.'s study, especially that this will be used as a reference guideline Produces better distribution (normality) of the data compared to smaller scales, especially if we plan to use it as a continuous scale Allows us to detect smaller changes, but categories become more difficult to distinguish compared to a 7-point scale
Remove the neutral option	Mimics a clinical setting by forcing participants to make a decision But removing this option does not match previous studies and affects our calculation of the empiric risk threshold and confidence level as both these measures rely on the presence of uncertainty in recommendations

2.7.5.2 Choice Whether to Ask Users for their Risk Estimate for Each Case

In our early design discussions, we thought it might be beneficial to ask users for their risk estimate of each case. The main reason was that this estimate would give us a clear measure of the weight of advice in the different conditions. For example, using a formula such as that suggested by Yaniv (2004) in the Judge Advisory System literature where weight of advice was measured as:

weight of advice = abs(user's final estimate – user's initial estimate) /
abs(advisor recommendation – user's initial estimate)

In fact, such a formula would give us a measure of advice utilization for each case rather than a series of cases.

However, we decided against this set up for several reasons. Firstly, asking for a risk estimate for each case is likely to have a strong influence on users' decisions and their stance with the calculator that provides a directly comparable numeric measure. Secondly, it affects participant decisions to treat in a way that is unnatural to the clinical setting. Furthermore, it significantly increases the time and effort of the participants which is an important consideration as these are professionals and are not being offered monetary incentive. Finally, the ability of participants to assign a numeric risk estimate is less important than how their treatment recommendation reflects the underlying risk (Boland, 2008) and research shows that most physicians are uncomfortable in providing numerical values, are poor at estimating probabilities, and are better at stating whether a given patient is above or below a certain risk (Boland, 2010).

Instead, we chose another approach which removes this need to ask participants for their risk estimate. Since the risk associated with each patient can be objectively obtained from the glaucoma risk calculator, participants can be asked for their treatment decision only and then, using regression, we can estimate physician treatment thresholds (Boland, 2010). Besides removing the need for participants to explicitly estimate patient risk, this method also uses the same risk (from the calculator) for each case instead of the many intuitive estimates provided by clinicians which allows us in turn to compare treatment thresholds across participants (Boland, 2010). Furthermore, as put by Boland et al. (2010), this method allows researchers to compare the impact before and after "interventions like risk calculators [and] clinical decision support tools" (Boland, 2010). We describe details of this method in Section 2.8.3.1.

2.7.5.3 Choice of Patient Factors Used to Characterise Each Case

When choosing the patient characteristics that would define each patient case we had several choices that either matched previous publications or extended their work. We chose to define cases based on the five risk factors included in the calculator's estimate (age, IOP, CDR, CCT, and PSD), as well as family history, life expectancy and disc size. This allowed us to extend some of the previous work by Boland et al. as we now had cases where the calculator could be considered misleading as it did not take into account the additional factors (family history, life expectancy, and disc size) that could affect the patient's risk. However, at the same time, this option did not sacrifice the ability to compare to the international reference provided by Cheng et al (2009).

Table 2-7: Comparison of Options for Patient Factors that Can Be Used to Characterize Experiment Cases

Patient Factors	Advantages and Disadvantages
Only provide the five calculator risk factors (age, IOP, CDR, CCT, and PSD)	Cannot use the expert panel as a reference Doesn't effectively reflect clinical practice Limits our exploration of hypotheses where the calculator is misleading
Also provide: life expectancy, disc size, family history	Allows us to compare to expert panel decisions Is significantly different and more realistic than Boland's study Allows analysis of hypotheses that consider user reactions when the calculator is misleading
Also provide: race, gender, occupation, patient preferences, etc.	Cannot use expert panel decisions as a reference Allows us to specify realistic patient cases Allows analysis of hypotheses that consider user reactions when the calculator is misleading
Use a different set of factors for different cases	Allows us to win the advantages in each of the above rows, but may confuse participants and affect their depiction of a "typical" patient when certain information is absent

2.7.5.4 Choice of Exact Wording of the Treatment Question

In Boland et al.'s studies (2008, 2011), for each patient, physicians were asked how likely they would be to recommend treatment to prevent glaucoma in this patient. Responses were recorded along a 7-point scale that included: Definitely No (no treatment), Probably No, Possibly No, Unsure, Possibly Yes, Probably Yes, and Definitely Yes.

In Cheng et al.'s study (2009), panellists were asked to "rate the appropriateness of initiating treatment on a 1- to 9-point scale" where 9=extremely appropriate (benefits greatly exceed risks), 5=uncertain (benefits and risks about equal), and 1=extremely inappropriate (risks greatly exceed benefits).

Although we, especially the medical members of our team, preferred the wording by Boland et al., we chose to use the wording by Cheng et al. to maintain comparability with their results. We contacted the authors for the exact wording and applied it in the study as we felt that changing the wording from appropriateness to probability to treat might create a fine difference in some responses. Because of the lack of a ground truth in the treatment of OHT patients previously discussed in Section 2.4 and because our expectation that many of our participants would be from outside the UK and thus could not be held accountable to NICE guidelines, it was important to maintain comparability with Cheng et al.'s study which provided an international guideline and incentive to our participants.

2.7.5.5 Choice of Method to Distribute Participants Across Conditions

We had discussions to stratify participants when assigning them to the different conditions. We considered stratifying them according to specialty as Boland et al. (2011) showed that glaucoma specialists were less likely to recommend treatment, according to profession as Professor Lawrenson expected differences between optometrists and ophthalmologists' decisions, or according to location of practice where we expected training and guidelines would have an effect on treatment recommendations. To be truly effective, this stratification would likely need to take into account a combination of the three characteristics as there may be, for example, an optometrist in a hospital-glaucoma clinic who sees more OHT patients than an ophthalmologist not specialized in glaucoma.

In the end, we assigned participants to the three conditions randomly so that any given participant was equally likely to see the sequential-know or sequential not-know condition, but

only two-thirds as likely to see the simultaneous condition. Our rationale for this was that to stratify participants we would need to divide their participation into two parts: one where we ask about their specific characteristics, and one where they actually took part in the study according to the condition they were assigned to. Due to implementation constraints, the two steps could not be combined into a single step. The trade-off was to simplify participation to a single setting in order to encourage the greatest number of participants possible while trusting that participants of varying characteristics would eventually become equally distributed across conditions.

The reason that the simultaneous condition was weighted slightly less than the sequential conditions (2:3:3 ratio of being assigned to SIM, SEQNK and SEQNK, respectively) was that the sequential conditions address 4 out of 5 of our main hypotheses and thus increasing the probability that a participant is exposed to the sequential rather than simultaneous condition gives more power to these hypotheses by increasing their sample size. This increase is especially important because we expect the effect size to be very small.

2.7.5.6 Randomization of Order and Cases

Participants in the experiment were randomly assigned to one of the following three experiment URLs. Sets A and B refer to the ten cases chosen and their ten equivalently matched cases, respectively (described in Section 2.7.4.2):

- Link 1:
  - SEQK (set A)
  - SIM or SEQNK (set B)
- Link 2:
  - SEQNK (set A)
  - SIM or SEQK (set B)
- Link 3:
  - SIM (set A)
  - SEQK or SEQNK (set B)

In each link, the order between the two conditions was randomized so that no condition was always first or last. Also, the case sets were varied between the conditions to control for any variability due to the set itself. Thus, no condition always contained the same set of ten cases. Finally, the ten cases in each condition were always presented in a random order so that no case was always seen first or last.

2.7.5.7 Concerns in the Sequential Conditions

We acknowledge an important limitation in the design of the experiment which is that the second look effect may be a factor in some people's decisions in the SEQNK and SEQK conditions and would be interpreted as a result of the advice. The "second look" effect posits that decisions may be different the second time the user attempts the problem, not because of advice from the DSS, but because users have had more time to think about the case and recall relevant information (previously described in Chapter 1). Distinguishing between the two effects is challenging although certainly desirable. Our unsuccessful attempt to distinguish these two effects is discussed again in Section 2.8.1.

Another concern about responses in the SEQNK condition was that after going through the ten cases and then being presented with advice from the tool, users may become less energetic and keep their responses the same due to fatigue or boredom. This concern is difficult to avoid. However, we include, in the briefing, statements about the importance of the research and of participants' responses in informing this research in an attempt to remind users of the importance of maintaining their attention, as much as possible, across conditions. Furthermore, in the implementation of the interface, we ensure participants need to re-enter their management decision in the post-advice condition, even if it is the same as their pre-advice decision in order to equalize the work in both situations.

2.8 Results

This section presents results obtained from analysing the data. All data was exported from Qualtrics and analysed using Excel and SPSS.

2.8.1 Results of the Pilot Study

In December 2015, we recruited three professionals for the pilot study. No logistic/technical issues were found. Besides the standard study questions, participants were also presented with a list of pilot questions. In the following, these questions are listed along with the responses we received:

1. How did you find the study instructions and wording of questions (in terms of clarity, length, etc.)?
    - Very clear and concise
    - Very appropriate

- Easy to understand and clear for the task

2. Are there things we can do to improve the interface (in terms of user-friendliness, overall appearance, etc.)?
   - Not really – I liked it
   - You did well
   - Looked good, not overcrowded or busy

3. Please comment on the length of the study (was it within the predicted 30 min slot, was it reasonable, did you lose interest towards the end?)
   - Appropriate
   - On time, reasonable, and no I didn't
   - Was interesting and quick

4. How did you find the invitation email (was it encouraging to sign up, was it clear that the study needs to be completed once you click the link, etc.)?
   - Fine
   - Yes

5. Do you have any more comments or suggestions to help us improve the study?
   - I enjoyed the study – nicely done
   - You need to consider the clinical experience of the glaucoma specialist, diurnal curve of IOP reading, refractive errors, visual needs, psychic build-up of the patient, compliance to medications and follow-up, economic status of the patient, and availability of drops or laser treatment

Although there did not seem to be any logistic issues, we observed two key concerns:

1. Sometimes a participant's treatment decision moved towards "Inappropriate to Initiate Treatment" along the management scale after seeing the risk estimate despite the risk being above 15% (which we consider as high and corresponding to advice to treat based on previous publications). This suggests that people interpret the risk in different ways and sometimes this interpretation results in a change in the direction opposite to what we expect.

   To resolve this first concern, we considered:
   - Asking participants, "Do you see this as a suggestion to treat/not treat?" after each risk estimate. However, this is tedious and does not represent a real-world setting.

- Stating at the beginning of the study that "experts recommend treating above 15%". However, this puts users in direct confrontation with the calculator's advice, and is almost asking users to diminish their own judgement.
- Asking participants, "At what level do you treat?". We chose this last option to determine each person's self-reported diagnostic threshold.
    - Furthermore, we asked this question at the beginning and end of the study in case of learning effects.
    - We deliberately made the response choices to this question: 5, 10, 15, 20, 25, 30%, and a text field for "Over 30%". By including only these values rather than the entire 0-100% scale, we attempted, in case a user's threshold was high, to keep user thresholds as close as possible to the suggested level of 10-15% (Karmel).

2. In the simultaneous condition in the pilot study, we repeated one of the ten cases twice, once towards the beginning of the set and once towards the end. We did this to see if we would notice any differences in decisions between the two repeated cases and attribute them to second-look effects.

   For some participants, we noticed a maximum 1-point difference between treatment decisions on this same case. However, after much discussion, we decided that this difference could not truly be attributed to the second-look effect as the participants were not explicitly informed that this is the same case they had seen. Instead, this type of difference is better defined as expected noise in participants' decisions, and the use of statistical significance tests should be sufficient in preventing us from attributing such noise to effects of the factors of interest (within the defined confidence levels). Finally, as discussed in Section 2.7.5.7, we were unable to design other methods to factor out second-look effects and thus we acknowledge, like many previous researchers, the limitation that any second-look effects are not distinguishable from other effects we are studying in the experiment.

2.8.2 Demographic Data

We ran the experiment from May 6, 2016 to June 7, 2016. In this timeframe, and including the three participants from the pilot study, a total of 52 ophthalmologists and optometrists completed the study. Four participants started the study and did not complete it. Table 2-8 summarizes the demographics of the participants, where numbers in parentheses are the actual number of participants in each category.

Table 2-8: Demographics of Study Participants (N=52)	
Characteristic	**Mean ± SD**
Gender (male)	88.5% (46)
Profession (ophthalmologists)	88.5% (46)
Location of practice United Kingdom Saudi Arabia Egypt United Arab Emirates, Russia, Mexico, Maldives, Jordan	 28.8% (15) 28.8% (15) 25% (13) 13.5% (7)
Specialty (glaucoma)	44.2% (23) (Also visually depicted in Figure 2-4)
Average length of practice (years)	16.5 ± 7
Average number of patients/week	121 ± 78
Percentage of glaucoma/glaucoma suspect patients	24% ± 24%
Use of glaucoma calculator Never Rarely About half the time Frequently	 40.4% (21) 32.7% (17) 15.4% (8) 11.5% (6)
Self-reported diagnostic risk threshold At the beginning of the study At the end of the study	 21% ± 9% 25% ± 11%
Percentage of glaucoma suspects treated in clinic	32% ± 19% (Also visually depicted in Figure 2-3)
Familiarity with OHTS (familiar)	84.6% (44)
Familiarity with NICE guidelines Familiar Familiar among the UK sample	 40.4% (21) 33.3% (5)
Familiarity with rationale of risk calculator (familiar)	50% (26)

Figure 2-3 depicts the distribution of percentage of glaucoma suspects treated in each participant's practice, while Figure 2-4 shows a bar graph representing the different specialties of the 52 participants.

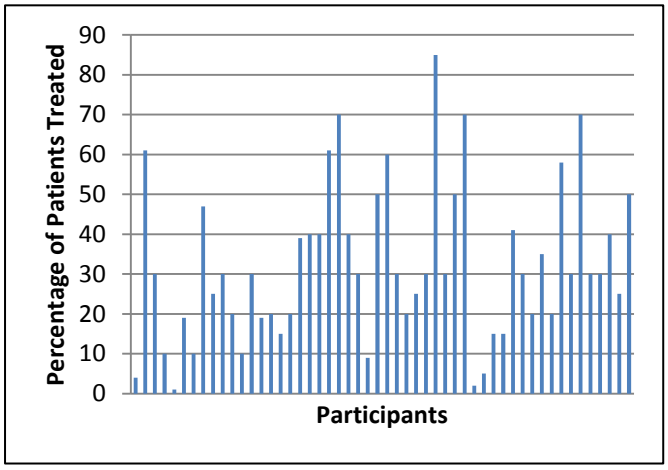


Figure 2-3: Distribution of Percentage of Glaucoma Suspects Treated in Each Practice

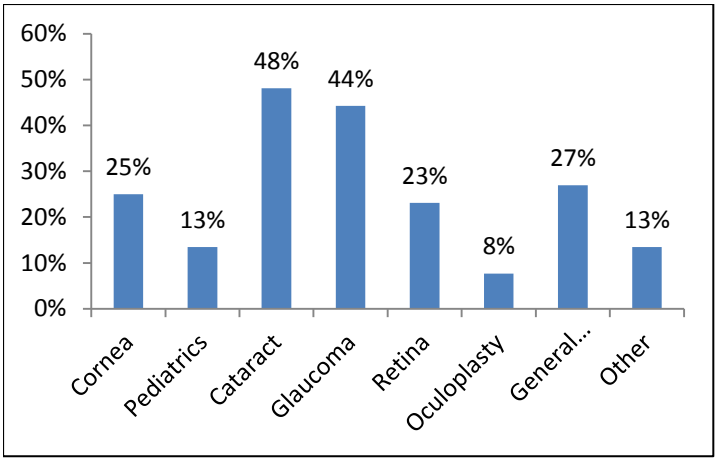


Figure 2-4: Distribution of Participant Specialties

We note that the average self-reported risk estimate that participants considered treatment at was 21% ± 9% which is slightly higher than the recommended 15% (Karmel). Some participants cited risk as low as 5% and others cited as high as 50% (and this maximum may be higher for some participants as several of those who chose "Over 30%" as their answer did not then write a risk threshold in the text field). By the end of the study, this average had risen slightly to 25% ± 11%. Again, some participants cited risk as low as 5% and others cited as high as 50%. Of the 44 participants who did not leave an empty threshold in their answer at either the beginning or end of the study, 24 did not change their threshold (54.5%). Most changes (75%) were an increase in treatment threshold. Absolute changes ranged between 5% to 25% on the risk scale.

2.8.3 Results of the Human-Computer Interaction Hypotheses

In this section we present results for each of the five hypotheses described in Section 2.6. Each subsection addresses a single hypothesis by: (1) restating the hypothesis, (2) describing the method and statistical tests used to test the hypothesis, (3) presenting the results of these tests, and (4) providing a brief discussion explaining the results and their implications. As a shorthand notation, we sometimes refer to pre-advice and post-advice decisions in the SEQK condition as SEQKA and SEQKB, respectively. Similarly, pre-advice and post-advice decisions in the SEQNK condition are sometimes abbreviated as SEQNKA and SEQNKB, respectively.

2.8.3.1 Hypothesis 1: Pre-Advice Sequential-Know Decision vs. Pre-Advice Sequential-Not Know Decision

Participants will be less likely to suggest treatments in their pre-advice decisions when they know they will be given a risk estimate (SEQK) than when they do not know (SEQNK).

Method of Analysis:

For this hypothesis, our analysis is limited to the 23 participants who experienced both the SEQK and SEQNK conditions. We start exploring differences between the pre-advice decisions in the SEQK and SEQNK conditions by measuring the rank-order association between the treatment recommendations of the same participants in the two conditions using Kendall's tau-b, which is a suitable correlation measure for ordinal variables with many ties (Wiredu, 2014). We compare the frequency of decisions to treat between the pre-advice recommendations in the SEQK and SEQNK conditions using McNemar's test as our two groups are related (Wiredu, 2014). Furthermore, because the data violates Shapiro Wilk test for normality, we apply instead the non-parametric Wilcoxon signed rank test (Wiredu, 2014) to compare mean pre-advice decisions in the SEQK and SEQNK conditions.

Next, we use the method suggest by Boland et al. (2010) which derives treatment thresholds from actual physician treatment recommendations. Using ordinal regression, and using only empiric, uncertain treatment recommendations, we derive the risk threshold at which participants are equally likely to make a Yes or No recommendation (i.e., are maximally unsure). As defined by Boland et al. (2010), this threshold, *r,* is:

$$r = -\frac{\alpha_{yes} + \alpha_{unsure}}{2\beta}$$

where α is the intercept coefficient for that level of recommendation on the ordinal scale and β is the regression coefficient for risk. Advantages of this method have been previously discussed in Section 2.7.5.2.

For this important hypothesis, we also calculate the empiric risk treatment threshold for each participant in each condition separately. This allows us to see individual changes between participant's thresholds. Unfortunately, many of the participants who saw both the SEQK and SEQNK conditions used only No and Yes in their recommendations and thus Boland's method, which is based on uncertainty in treatment decisions (some "Unsure" decisions), could not be used to calculate their empiric risk threshold. This reduced this specific matched pair analysis of individual changes in risk threshold to 7 participants instead of 23.

Results:

As we expected, pre-advice recommendations in the SEQK and SEQNK conditions showed only moderate association (Kendall's tau-b = 0.458; $p<0.0005$). This significant difference was further supported by a Wilcoxon signed-rank test which revealed a significant difference between mean treatments in the pre-advice decisions between the two sequential conditions, $Z=-3.075$, $p=0.002$.

For the same 23 participants that experienced both the SEQK and SEQNK conditions, 49.6% of decisions were to treat when participants knew they would receive advice and 58.3% were to treat when they did not know they will receive advice. McNemar's test determined that these proportions were significantly different, $p=.017$.

We analysed the direction of this difference by calculating the average risk treatment threshold across all cases and for all participants in the pre-advice decisions in the two conditions. This treatment threshold was 28.98% in the SEQK condition and 16.58% in the SEQNK condition.

Six of the seven participants who saw both sequential conditions and who expressed uncertainty in their recommendations had a higher treatment threshold when they knew they would be presented with advice. Again this matched our expectation and is visually depicted in the graph in Figure 2-5.

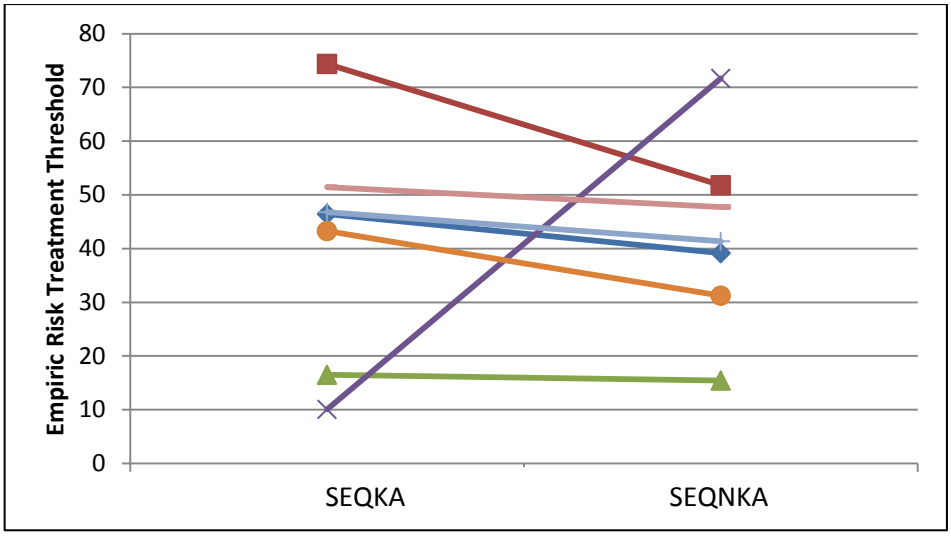


Figure 2-5: Comparison of Risk Treatment Thresholds in the Pre-Advice Decisions of Participants who Experienced the SEQK and SEQNK Conditions

Each line represents the performance of a single participant where one end is the participant's risk treatment threshold in the Sequential-Know condition and the other end is their threshold in the Sequential-Not Know condition. As depicted, 6 of the 7 participants had a higher treatment threshold when they knew they would be presented with advice

Discussion:

These results support the hypothesis that the mere knowledge of the presence of a decision support tool can alter users' decision thresholds. These results also explain the direction of this difference: that participants are less likely to suggest treatment in their pre-advice decision when they know they will be given advice. This could be evidence of the proposed homeostasis mechanism previously described in Section 1.5.8.

This result is especially important for studies that measure the effect of a decision support tool by comparing participants' pre-advice decision when they know they will be receiving advice with their post advice decision because the result suggests that this pre-advice decision is already influenced by the mere knowledge of the presence of the tool.

This result is also important when considering suitable mitigation strategies for automation bias. For example, one protocol used to combat automation bias in mammography does not allow users to change their pre-advice decision from "target" to "no target". Although such a protocol may help prevent some human omission errors caused by tool false negatives, it would, on the other hand, exacerbate effects of the homeostasis mechanism and thus possibly introduce a new set of unaccounted errors. Thus, to avoid such conflicting benefits, a complete

understanding of all possible cognitive mechanisms that may affect users supported by DSSs is necessary.

2.8.3.2 Hypothesis 2: Post-Advice Sequential-Know Decision vs. Post-Advice Sequential-Not Know Decision

The post-advice decisions in the sequential know and sequential don't know conditions will be significantly different.

Method of Analysis:

We measure the rank-order association between the treatment recommendations in different conditions using Kendall's tau-b, which is a suitable correlation measure for ordinal variables with many ties (Wiredu, 2014). Furthermore, because the data violates Shapiro Wilk test for normality, we apply instead the non-parametric Wilcoxon signed rank test (Wiredu, 2014) to compare post-advice decisions in the SEQK and SEQNK conditions. All comparisons are between matched pairs that experienced the two conditions being compared.

Results:

Table 2-9: Association Between Treatment Recommendations in the Different Conditions

	First Condition	Second Condition	Kendall's tau-b	Kendall's tau-b Significance
Effect of Knowledge of the Presence of a Tool	SEQKA	SEQNKA	0.458	$p < 0.0005$
	SEQKB	SEQNKB	0.709	$p < 0.0005$
Effect of Advice	SEQKA	SEQKB	0.699	$p < 0.0005$
	SEQNKA	SEQNKB	0.535	$p < 0.0005$
Effect of Sequence of Advice	SEQKB	SIM	0.641	$p < 0.0005$
	SEQNKB	SIM	0.605	$p < 0.0005$

Effect of Knowledge of the Presence of a Tool: As we expected and as discussed in the results from Hypothesis 1, pre-advice recommendations in the SEQK and SEQNK conditions showed low levels of association, suggesting that knowledge of the presence of a tool was sufficient to change some decisions. This low association between pre-advice decisions in the SEQK and SEQNK conditions rose to 0.709 in the post-advice decisions, suggesting that tool advice helped

differences between these two conditions converge. Based on Wilcoxon signed rank test, no significant difference between post advice decisions in the two sequential conditions was found; Z=-1.42, p=.155.

Effect of Advice: The moderate correlations (0.535 and 0.699) between pre-advice and post-advice decisions in the two sequential conditions suggest that although decisions were different before and after advice, they were positively related, as would be expected. Wilcoxon signed rank test between pre and post-advice decisions reveal non-significant differences in both conditions; Z=-1.46, p=0.143 for SEQK and Z=-1.44, p=.149 for SEQNK. Interestingly, correlation between participants' recommendations before and after they received advice was higher in the SEQK condition compared to the SEQNK condition. This suggests that participants changed their recommendations more often/more drastically in the SEQNK condition and is contrary to our expectation as will be discussed further in Hypothesis 3.

Effect of Sequence of Advice: Correlation between decisions in the simultaneous and sequential conditions was moderate, suggesting that the sequence in which advice was presented affected the treatment outcome. However, these differences were not supported by a Wilcoxon signed rank test between SIM and SEQNKB: Z=-.89, p=.373 or between SIM and SEQKB: Z=-1.47, p=.142.

Discussion:

The comparison of association levels between treatment recommendations in the different conditions suggests that all three factors have an effect on treatment decisions: (1) advice from the tool, (2) knowledge of the presence of tool advice, and (3) sequence of advice. However, at the 95% confidence level, only differences due to the knowledge of the presence of the tool are supported by the Wilcoxon signed rank test. Contrary to Cheng et al. (2009) who concluded that the calculator did not have a strong influence on users' decisions, we found that users changed their decisions after the presentation of advice and that 8-15% of these changes were complete reversals (as will be discussed further in Hypothesis 3).

We note that association levels between the pre-advice decisions in SEQK and SEQNK, based on Kendall's tau-b, increased between post-advice decisions in the two conditions. This suggests strong differences between pre-advice decisions, but differences that converge to a common association when presented with advice. These results may suggest different levels of pre-processing of the information by users depending on the availability of advice.

Similar to findings in Hypothesis 1, this finding is important for studies that measure the effect of a decision-making tool based on pre and post advice decisions where users have been informed of the presence of a tool that will be provided to aid them in their decision. Results suggest that a more reliable comparison is based on a control group that neither sees advice from a tool nor knows about the presence of the tool.

2.8.3.3 Hypothesis 3: Changes in Sequential-Know vs. Sequential-Not Know Decisions

There will be fewer changes in users' management decisions after seeing the risk estimate in the group that doesn't know they will receive support compared to the group that knows they will receive support.

Method of Analysis:

We measure changes in decisions using two variables: (1) the number of complete reversals from Treat to Don't Treat or vice versa and (2) the absolute magnitude of change along the 9-point treatment scale between the pre and post advice decisions for all cases and all matched pairs in the two conditions. This absolute magnitude of change can take on values between 0 and 8. As our two samples are related, McNemar's test is used to explore if there are significant differences between the number of complete reversals in participant decisions in the SEQK and SEQNK conditions (Wiredu, 2014). As the subjects in the two conditions are the same and because the data violates the Shapiro Wilk test for normality, we use the non-parametric Wilcoxon signed rank test to determine if there are significant differences between the mean absolute magnitude of change between the two conditions (Wiredu, 2014).

Results:

Contrary to our expectations, changes were greater when people did *not* know they will receive advice, both in terms of the average absolute change on the treatment scale and in terms of the percentage of complete reversals.

The percentage of complete reversals was 8.37% in the SEQK condition and 15% in the SEQNK condition, which are similar to the proportion reported by Boland et al. (10%). Using McNemar's test, we found a statistical significant difference between the proportion of complete reversals in the two conditions at the 90% confidence interval; p=0.052.

Based on the Wilcoxon signed rank test, there was also a significant difference between the mean absolute magnitude of change in the SEQK (M=1.07, SD=1.42) and the SEQNK (M=1.50, SD=1.80) conditions; Z=-3.02, p = 0.003.

Furthermore, as we saw in the analysis of Hypothesis 2 in Table 2-9, association levels between pre and post-advice decisions were greater in the SEQK compared to the SEQNK conditions. Again this suggests that participants changed their decisions more often and/or more drastically when they did *not* know they will receive advice.

However, there seemed to be slightly more fence sitting in the preliminary decision when participants knew they will be receiving advice compared to when they did not know they will be receiving advice (10.9% in SEQK condition and 9.69% in the SEQNK condition). And although this conforms to our expectations, McNemar's test (p-value=0.349) suggests that this difference between the number of uncertain decisions in the SEQK and SEQNK conditions was non-significant.

Discussion:

Contrary to our expectations, users changed their decisions more often when they did not know they will receive advice. We predicted that if conventional explanations (overtrust, overreliance, complacency) of AB were dominant, then users not knowing they will receive advice would become less susceptible to advice, more confident in and more anchored to their original decision.

Importantly, this result may be an indication that users are not being overreliant or complacent as suggested by conventional explanations of AB. Interestingly, this is not the first time that results in the literature point against conventional explanations. For example, Bahner et al. (2008) reported that 20% of participants in their study sampled the raw data sufficiently. Another example is the study by Sniezek at al. (1995b) where judges outperformed unadvised decision-makers but did not differ significantly from them in terms of information search or effort; thus, the idea that the availability of advice is taken as an indication by users that they can exert less effort than unadvised decision-makers (i.e., ideas of complacency, social loafing, or reduced vigilance) was not supported. Such results, emphasize the importance of exploring other possible cognitive mechanisms, such as the homeostasis mechanism.

However, we also acknowledge that the experimental setup may have had an effect on the result of Hypothesis 3. In the SEQK condition participants saw the case with advice directly after seeing it without, while in the SEQNK condition they saw the case again only after seeing and giving a

decision on the full set of ten cases. There was no way around this set up as there needed to be an element of surprise in the SEQNK condition which could only come after all ten cases were presented. However, perhaps this led to a learning effect, a stronger second look effect on these cases, or a reduced anchoring effect that may explain the results.

The fact that there were slightly fewer unsure decisions in the SEQNK condition compared to the SEQK condition may suggest that some users delayed their decision until after they saw the tool's estimate. This delay likely occurred for the more difficult cases, and if these cases were also difficult for the tool due to the lack of diversity that exists between many DSS tools and their users, then the tool may err on such cases. The problem is that such errors could push these increased number of unsure decisions in the wrong direction due to the false assurance relayed to the user who waited to see the tool's advice. This may also suggest that if future experimenters manipulate the frequency of such borderline cases in the data set, then it may be possible to obtain significant differences between the sequential post-advice decisions, unlike the non-significant differences we found in Hypothesis 2.

2.8.3.4 Hypothesis 4: Factors Affecting Changes in Users' Decisions

The probability that a user will change his/her pre-advice decision increases:

a. When the pre-advice decision is uncertain
b. If change is towards the "safer" side of treating
c. For more difficult cases
d. For users that report more trust in the calculator
e. For less experienced users
f. For non-glaucoma specialists
g. For maximizers more than satisficers

Method of Analysis:

Our dependent variable measures the magnitude of change on the treatment scale between each user's pre and post-advice decision. We also include a sign to indicate how this change compares to the direction of advice (positive if in the direction of advice and negative otherwise). To incorporate differences in how users interpret the calculator's risk estimate, we determine the direction of advice by comparing the case risk with the average of the self-reported thresholds of each user at the beginning and end of the experiment.

Our independent variables are defined as follows:

a. Uncertainty of pre-advice

Create a variable which ranks the uncertainty in a participant's decision as follows:

Table 2-10: Uncertainty Scores

Participant's Decision	Uncertainty Score
1 or 9	1 (low level of uncertainty)
2 or 8	2
3 or 7	3
4 or 6	4
5	5 (highest level of uncertainty)

b. Participant's knowledge that he/she will be receiving advice

Create a nominal variable which codes decisions from SEQK as 1 and those from SEQNK as 0.

c. Change towards the "safer" side of treating

Create a variable which measures the sign of the change between a participant's post and pre advice decisions. Set the variable to 1 if the change is towards treatment, -1 if it's towards no treatment, and 0 if there is no change.

d. Case difficulty

Previous publications (Alberdi, 2010) have measured the difficulty of a cancer case based on the fraction of readers that miss the cancer in the unprompted condition. A similar measure is difficult in our context because of the lack of a clear ground truth for each case.

Instead, we use two measures of difficulty: (1) the variance in participants' pre-advice decisions in the SEQNK condition for each case and (2) the average level of uncertainty in participant's pre-advice decisions in the SEQNK condition for each case. We choose to focus only on decisions in the pre-advice SEQNK condition as they resemble the unprompted condition in other publications and measure responses in their most uninfluenced condition – without even knowledge of the presence of advice.

e. User trust in the calculator

Measured on a Likert scale in the final questionnaire. As a closely related measure, we also input the user's perceived tool reliability, which was also measured on a Likert scale in the final questionnaire.

f. User experience
   Measured using two variables: the number of years of experience and the user's patient volume.

g. Glaucoma speciality
   Define a nominal variable which is 1 for glaucoma specialists and 0 for non-glaucoma specialists. We also use percentage of glaucoma patients seen in the user's clinic as another measure of glaucoma specialty.

h. Maximizer/Satisficer Tendencies
   Measured as the score of participants in the standard decision-making survey at the end of the experiment.

We consider all participants and cases in both sequential conditions. As both our dependent variables, absolute magnitude of change and signed change, violate the Shapiro Wilko test for normality, we consider non-parametric measures of correlation. Although the non-parametric Spearman and Kendall tau-b measures of correlation yield similar results, we report only Kendall's tau-b results as some of our data contain ties (Wiredu, 2014).

Results

Table 2-11 presents results of Kendall's tau-b measure of correlation along with significance values for each of our ordinal/continuous independent variables with our two measures of change: absolute magnitude of change and signed change.

Table 2-11: Correlation Between Change in User Decisions and Different Factors				
	Signed Change		Absolute Value of Change	
Variable	**Kendall's tau**	**p-value**	**Kendall's tau**	**p-value**
Uncertainty*	0.235	$p < 0.0005$	0.205	$p < 0.0005$
Towards "safer" side of treating*	0.229	$p < 0.0005$	0.009	$p = 0.773$
Difficulty By variance* By uncertainty*	 0.059 0.070	 $p = 0.042$ $p = 0.017$	 0.086 0.062	 $p = 0.003$ $p = 0.031$
Self-reported Trust*	0.029	$p = 0.390$	0.093	$p = 0.004$
Perceived Tool Reliability*	0.061	$p = 0.070$	0.109	$p = 0.001$
Experience In Years By Patient Volume	 0.008 -0.002	 $p = 0.788$ $p = 0.938$	 -0.02 -0.008	 $p = 0.487$ $p = 0.788$
Percentage of glaucoma patients*	0.04	$p = 0.16$	-0.078	$p = 0.006$
Maximizer/Satisficer Tendencies	0.041	$p = 0.159$	0.03	$p = 0.30$

The results in Table 2-11 suggest that in this dataset, none of our independent variables show a strong association with the amount of change in a user's decision after presentation of advice. Some of these low association levels may be due to chance as seen by their high p-values. However, uncertainty in a user's pre-advice decision, change towards the "safer" side of treatment, case difficulty, trust, perceived tool reliability and percentage of glaucoma patients seen in a user's practice show correlation values which, at the 95% confidence level, cannot be attributed to chance. Of these relationships, only uncertainty in a user's pre-advice decision and change towards the safer side of treatment reveal a positive relationship - although weak (0.235 and 0.229, respectively). The other correlations (case difficulty, self-reported trust, perceived tool reliability and percentage of glaucoma patients), although significant, reveal no apparent relationship with the amount of change in a user's decision after presentation of advice.

To measure effects of the nominal glaucoma specialty variable, we define an average amount of change for each user based on the amount of change in their decisions across all cases in the sequential conditions, and then run a Mann Whitney U Test to determine differences between glaucoma specialists and non-glaucoma specialists. Results are non-significant; $U=251.5$, $p=0.693$ for signed change and $U=205.5$ and $p=0.169$ for absolute change.

As studies examining associations between personality traits and tool use are limited (Parasuraman, 2010), we run further tests to determine whether the maximizer/satisficer personality test score affects changes in user decisions. We determine each user's average amount of change across all cases and run a Mann Whitney U Test to determine whether this amount of

change is significantly different between maximizers (score > 4) and satisficers (score < 4). No significant differences are found; U=209.5, p=0.843 for signed change and U=215, p=0.951 for absolute change. Furthermore, even when we look at extreme maximizers (defined by Schwartz, 2004 as having an average score >5.5) and extreme satisficers (defined by Schwartz, 2004 as having an average score <2.5), again no significant differences are found using the Mann Whitney U Test, U=2, p=0.510 for signed change and U=1, p=0.272 for absolute change. However, we note that this latter sample was very small and contained only a single extreme satisficer.

Discussion:

Contrary to our expectations, results showed no relationship between the amount of change in a user's decision after being presented with tool advice and most of the variables we considered and only a weak relationship with (1) increased uncertainty in the pre-advice decision, and (2) direction of change towards the "safer" side of treating. Future studies may consider the effects of these variables in a different context or using a larger sample. Future work may also consider alternative definitions of change in a user's decision, such as focussing only on complete reversals.

Interestingly, results draw attention to two interesting differences: (1) the difference between trust and perceived reliability and (2) the difference between difficulty defined in terms of variance or uncertainty.

Since data violated the Shapiro Wilko test for normality, we used the Wilcoxon signed rank test to test differences between the means of users' self-reported trust and perceived tool reliability. Results showed a significant difference; Z=-3.128, p=0.002. For almost all participants who gave different ratings to the two measures, a higher rating was given to perceived tool reliability compared to trust. This could indicate that users may perceive a tool as reliable but still choose not to trust it, perhaps because confidence in their own ability exceeds confidence in the tool's reliability. This result could also suggest that the two terms may be measuring different concepts where perceived reliability measures the person's opinion of the tool and trust measures the likelihood they will use its advice.

Although often used interchangeably, it is important also to distinguish between uncertainty and difficulty which may not always represent the same phenomena. Measuring difficulty based on uncertainty relies on users acknowledging the difficulty of the case in the level of uncertainty they express in their decision, while measuring based on variance in user responses is a more objective measure across all participants. We notice that in our dataset, the two measures of difficulty were

only moderately related (Spearman's correlation=0.612, p<0.0005 and Kendall's tau-b=0.333, p<0.0005) and were not always consistent as can be seen in Figure 2-6.

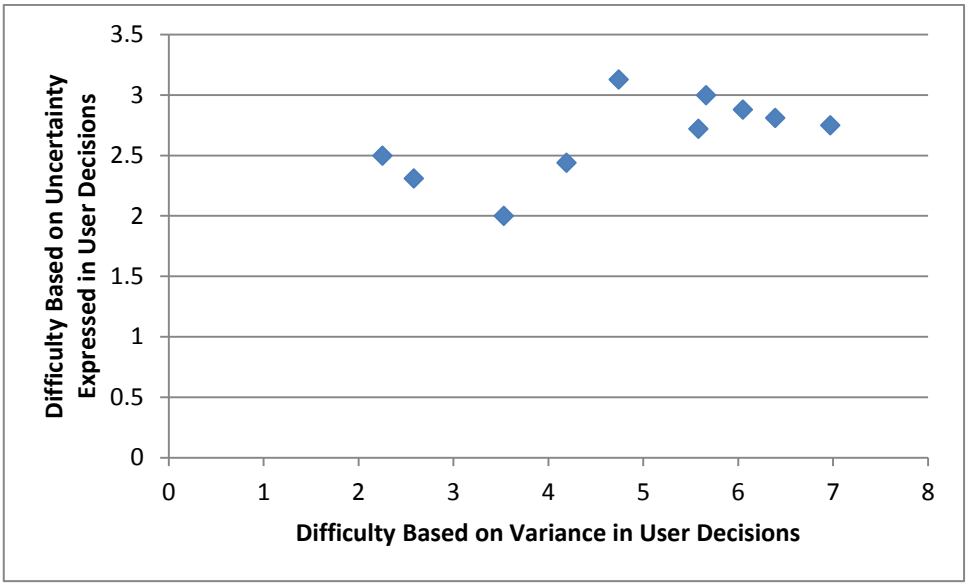


Figure 2-6: Relationship Between Different Measures of Case Difficulty

Finally, similar to the few studies cited by Parasuraman et al. (2010) which have looked at different personality traits and their effect on automation use, our results show no association between change in user decisions due to tool advice and the specific maximizer/satisficer personality trait, not previously studied in the literature.

2.8.3.5 Hypothesis 5: Simultaneous Decision vs. Post-Advice Sequential Decision when calculator is misleading

When the risk calculator gives "misleading advice", users are more likely to err when the risk estimate is presented simultaneously (SIM) rather than sequentially (SEQK and SEQNK). Thus, automation bias is more pronounced in the simultaneous mode.

Method of Analysis:

This hypothesis relies on the existence of a "golden truth" for each case. However, as previously explained in Section 2.4, such a unanimous standard on the treatment of glaucoma suspects does not exist. Instead, for our analysis of this hypothesis, we choose NICE guidelines as the basis for comparison as it is the standard that healthcare providers in the UK are expected to abide by.

As this hypothesis considers cases of "misleading advice", we thus limit our analysis to the five of our ten cases where advice from the calculator contradicts NICE guidelines. In these cases, NICE guidelines suggest no treatment while the calculator provides a high risk suggesting treatment (24, 50, 52, 77 and 87%). We test whether more participants choose to treat these patients in the simultaneous condition compared to the sequential conditions, and we explore this using McNemar's hypothesis testing as our comparison is based on the matched pairs in the two conditions (Wiredu, 2014).

Another way to measure whether the "weight" of advice is greater in the simultaneous versus sequential modes is to measure the number of agreements/disagreements with the calculator in the two conditions and check, using McNemar's hypothesis testing, whether these proportions are significantly different in the two conditions. For this analysis we consider all ten cases.

Finally, we explore whether sequence of advice has an effect on the type of errors committed by users (false negatives and false positives). We do this using McNemar's hypothesis testing while using NICE guidelines as the reference for "correct" decisions and again while considering all ten cases.

Results:

Using McNemar's test, we found that the proportion of decisions to treat the five patients of interest (where the calculator's risk estimate contradicted NICE guidelines) was not significantly different to the same proportion in the simultaneous condition; p-value=.388.

Also, the percentage of agreements with the calculator's advice for all cases was approximately equal between the matched pairs in the simultaneous mode (83.4%) and the sequential modes (84.1%). Based on McNemar's test, this small difference in percentage of agreements was non-significant; p-value=1.0.

We also compared, using NICE guidelines as the reference, the number of true positives/false negatives and the number of true negatives/false positives in all the matched pairs between the simultaneous and sequential conditions. Based on McNemar's test, no significant differences were found between the number of true positives/false negatives in the two conditions; p=0.332. Also, no significant differences were found between the number of true negatives/false positives in the two conditions; p=0.791.

Discussion

Results do not support the idea that advice carries a higher weight when presented simultaneously versus sequentially. This is in contrast to several studies which reported that the simultaneous mode reduced observer vigilance (Van Dongen, 2013; Taylor, 2008b) reduced sensitivity (Taylor, 2008b) and reduced performance (Zheng, 2004). Furthermore, we extended the work of previous studies by analyzing differences between the two sequences while incorporating error type. However, even with this finer analysis, the two conditions still did not show significant differences.

Related to how a DSS is used and especially how it is implemented into a user's workflow is the sequence in which the tool's advice is presented. Despite literature and protocols which assume the effect of the sequence of advice from an automation bias perspective, few studies have empirically examined its influence. We hypothesized that the sequence of advice would affect cognitive mechanisms and user's internal processing of information, but results did not reveal such differences in this context. Future work may explore the difference between the two sequences while taking into account case difficulty or correctness of advice. Future work may also look at such differences in different contexts as results are likely to be affected by the specific DSS used.

2.8.4 Results of the Medical Explorations

In this section we present results for each of the five medically related questions described in Section 2.6.2, which are specific to the glaucoma risk calculator as a specific CDSS and most relevant to researchers in the medical community. Similar to subsections in Section 2.8.3, each subsection here also addresses a single exploration by: (1) restating the question, (2) describing the method and statistical tests used to explore it, (3) presenting the results of these tests, and (4) providing a brief discussion explaining the results and their implications.

2.8.4.1 Exploration 1: Conformance to NICE and Expert-Panel Decisions

What do participants more strongly conform to: NICE guidelines or the expert panel decisions (Cheng, 2009)?

Method of Analysis:

To measure correlation between participants' decisions and NICE guidelines or the expert panel's decisions we used Kendall's tau-b, a suitable correlation for ordinal variables with

many ties (Wiredu, 2014). To match the information we have from NICE and the expert panel, we first collapsed the 9-point scale into three levels: Treat, Unsure and Don't Treat and then removed all Unsure decisions from the analysis. The comparisons were done using only pre-advice decisions in the sequential-not know condition as this condition best represents participants' decisions without any support or knowledge of the presence of support.

Results:

Overall, combined results from all cases in the SEQNK pre-advice condition, show that participants more often agreed with expert opinion than with NICE guidelines. In fact, correlation with NICE guidelines was extremely low (Kendall's tau-b=0.113, p=0.046). Correlation with expert opinion was moderate (Kendall's tau-b=0.496, p<0.0005).

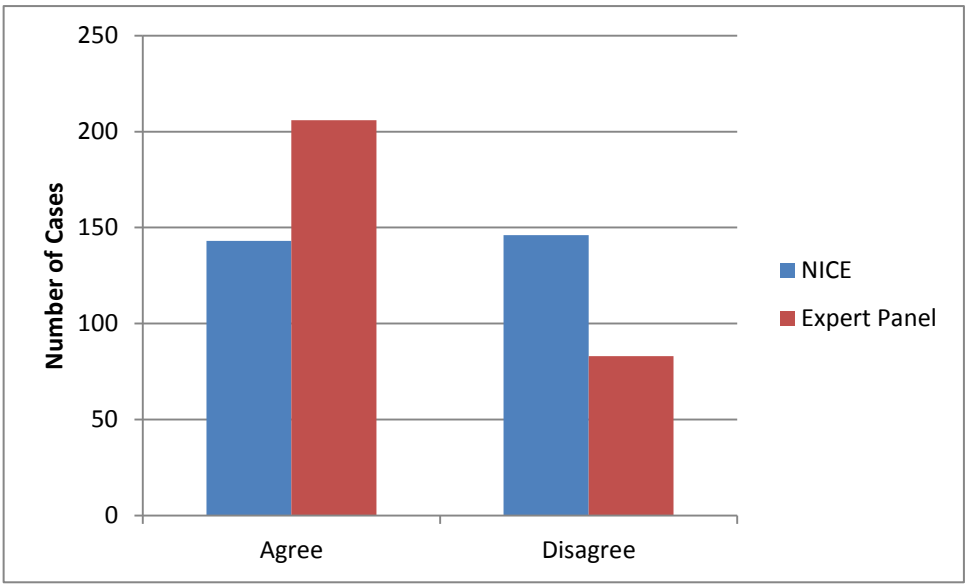


Figure 2-7: Comparison of Agreement Between Participant Decisions and NICE Guidelines and Expert Panel Decisions

It is also important to focus on the subset of participants from the UK and their conformance with NICE guidelines as they, unlike participants from outside the UK, are expected to know and abide by NICE guidelines. For this subset of our sample, results indicated genuine disagreement with the guidelines apparent in the very low, negative correlation (Kendall's tau-b=-0.072, p=0.466). There was, however, a lack of significance; perhaps because the subset of UK participants was smaller than our overall sample. Furthermore, when users were asked about their familiarity with NICE guidelines on the management of glaucoma suspects, only 33.3% of UK participants (5 out of 15) stated that they were familiar.

Discussion:

It is interesting to note that there was very strong disagreement with NICE guidelines, even among UK participants. However, these results should be observed with caution because half of the hand-picked cases we selected were chosen especially for their stark contradiction between NICE guidelines and expert opinion. These are cases where variables not considered by NICE guidelines, such as family history and cup-to-disc ratio, contribute to elevated levels of risk. Thus, agreement with NICE guidelines may be higher for a different set of chosen cases. Nonetheless, these results suggest that either ophthalmologists and optometrists in the UK need to become more aware of the guidelines, or that current guidelines need to be revisited to reflect modern advances in research and current treatment trends.

2.8.4.2 Exploration 2: Confidence, Inconsistency and Likelihood to Treat

Use of the calculator: (1) increases confidence, (2) decreases inconsistency and (3) increases the likelihood to recommend treatment.

Method of Analysis:

We explore differences in treatment recommendations between decisions in the SEQNK pre-advice and SIM conditions. We choose pre-advice SEQNK to measure decisions without tool support as this condition best represents participants' decisions without any support or knowledge of the presence of support. We choose SIM to measure decisions with support as this condition most closely represents how the calculator is used in a real clinical setting. Although combining post-advice decisions from SEQK and SEQNK with SIM decisions would increase the power of the results, we would be mixing decisions which we previously showed in our analysis are different.

We determine the empiric risk treatment threshold in these two conditions using all participants and cases in the SEQNKA and SIM conditions. All other analyses are done using the 9 matched pairs which participated in both these conditions. For the matched pairs, statistical differences are measured using the paired t-test and McNemar's test, depending on the normality of the data (Wiredu, 2014).

Confidence:

We measure confidence using two variables: the number of unsure responses in each condition and an average uncertainty score based on the coding in Table 2-10, which reflects the amount of uncertainty in each response.

Consistency:

We measure consistency in the same way as defined by Boland et al. as the sum of the variances of risk in each of the three decisions (Treat, Unsure, and No Treat) for each participant in each condition. This measure is highest for a participant who would be equally likely to assign a patient to any of the three decisions and lowest for a participant who consistently treats patients with a higher risk and withholds treatment for patients with a lower risk while using firm boundaries between recommendation levels. However, we use this measure with caution as it is highly dependent on the risks of the selected cases and also because it measures consistency based on the five patient characteristics included in the risk measure, which exclude family history, life expectancy and disc size.

Tendency to Treat:

We measure the tendency to treat in the two conditions both based on a count of treatment recommendations as well as the empiric risk treatment threshold in these two conditions calculated using the method presented by Boland et al. (2010) and previously described in Section 2.8.3.1.

Results:

The empiric risk threshold of participants with the calculator was 24% compared to 16.6% without the calculator. This suggests that users were less likely to treat with the calculator.

When we focus our analysis only on the nine matched pairs, results are in the direction expected, but are all non-significant. This may be due to the small size of this reduced sample. We notice that participants made a different treatment recommendation (Treat versus Don't Treat) between the SIM and SEQNKA conditions in 24% of responses. Overall, however, participants made a decision to treat 54.4% of patients in both conditions. Participants were more confident with the calculator based on their mean uncertainty scores, but according to the Wilcoxon Signed Rank test (as distributions violated assumptions of normality), this difference was nonsignificant; $Z=-0.645$, $p=0.519$. There were three times more unsure responses without the calculator, but based on McNemar's test, this difference was non-significant, $p=0.109$. Finally, participants were more consistent in treating higher risk patients when using the

calculator but a paired t-test found that this difference was again nonsignificant; t(6)=0.646, p=0.542.

Discussion:

Results are similar to those found by Boland et al. (2010). When presented with a risk estimate, participants' treatment recommendations: (1) changed in 24% of cases (compared to 10% in Boland et al.'s findings), (2) became more confident, and (3) became less inconsistent. However, unlike Boland et al. we found that the average treatment threshold increased, rather than decreased, with the calculator. This contradiction in findings suggests that the exact effect of the calculator remains undetermined and warrants further research. Naturally, this effect depends on the cases chosen and the participant sample. Finally, we note that results are non-significant, which may be due to the small size of this reduced sample, and thus again warrant further exploration.

2.8.4.3 Exploration 3: Improvement in User Decisions

Does use of the calculator make doctors better or worse where better or worse is defined according to:

a. More closely conform to NICE guidelines
b. More closely conform to expert-point system results
c. More consistent in ordering treatment for higher risk patients

Method of Analysis:

We compare the number of times participant decisions agree with NICE guidelines and the expert panel, again in the pre-advice SEQNK and SIM conditions, and explore significance using McNemar's hypothesis testing.

Results:

We found no significant differences between agreements with NICE guidelines in the pre-advice SEQNK and SIM conditions; p=.359. Similarly, no significant differences were found when using the expert panel's decisions as the reference for comparison; p=.648. However, we did find that there was a large difference in the magnitude of inconsistency in the two conditions suggesting that the calculator helped participants consistently treat patients with a higher risk threshold.

Discussion:

This exploration attempts to takes Boland et al.'s results one step further by exploring, as stated by the authors themselves, "whether the recommendations are more ''correct'' and lead to better patient outcomes". The difficulty in measuring this "correctness" is the absence of a clear golden truth for each case.

Undoubtedly, consistency, or the ability to make repeatable decisions given similar data is an important measure of the quality of medical care. We note, like Boland et al., that the calculator was effective in increasing intra-participant consistency of user decisions.

However, we find that the calculator failed to help participants better conform either to NICE guidelines or to expert opinion. This result is not entirely surprising. The calculator's estimate is based on five risk factors, NICE guidelines are based on 3 risk factors, while expert opinion is based on 6 risk factors. Since the three references base their recommendations on different criteria they naturally do not always agree, and use of the calculator will thus not always aid user decisions converge to one of the references. However, these discrepancies in the different references should serve as an eye opener to unify more of the decisions across references, which would make the job of medical professionals easier and the quality of care more consistent across hospitals.

2.8.4.4 Exploration 4: Effects of Specialty and Profession

We expect, based on Boland et al.'s results (2011), that glaucoma specialists will be less likely to recommend treatment compared to non-glaucoma specialists. However, we are unsure how optometrists' management decisions will differ from ophthalmologists' decisions, if in fact there is a difference.

Method of Analysis:

Again, all analyses were performed on decisions in the SEQNK pre-advice condition. A mean treatment response to the ten cases in this condition was obtained for each participant. When this distribution of mean participant responses in the SEQNKA condition violated the Shapiro Wilk test for normality for a certain factor, the non-parametric Mann-Whitney U Test was used instead. Otherwise, an independent t-test was used. We compared differences between the treatment recommendations of: (1) optometrists and ophthalmologists, (2) glaucoma specialists

and non-glaucoma specialists, (3) participants practising in the UK and those practising outside the UK.

Results:

No significant difference [U=34.5; p=0.220] was found between treatment recommendations of optometrists and ophthalmologists.

No significant difference [t(30)=-.459; p=0.650] was found between treatment recommendations of glaucoma specialists and non-glaucoma specialists.

However, a significant difference [U=34.5; p=0.001] was found between treatment recommendations of participants practising in the UK and those practising outside the UK, where participants outside the UK were more likely to offer treatment.

Discussion:

Compared to Boland et al. (2010), we found no difference in treatment recommendations depending on specialty. We also found no difference in recommendations based on profession. This is a constructive result as many glaucoma suspects in the UK are seen by optometrists and consistency in ophthalmologist and optometrist decisions is thus important. Finally, the finding that participants outside the UK were more likely to suggest treatment is likely the result of different training, guidelines, and economic considerations. Future work may consider a finer analysis of decisions rather than the coarser mean treatment recommendation.

2.8.4.5 Exploration 5: Incorporation of All Risk Factors in Clinicians' Decisions

Does the calculator truly encourage physicians to consider all five patient variables as claimed, or is this a step they already perform in their clinics?

Method of Analysis:

For this question, we inspected the variables users self-reported at the beginning of the experiment as factors they generally consider before offering treatment to a glaucoma suspect, and we compared these variables to the factors incorporated in the calculator's estimated patient risk.

Results:

All the participants cited that they consider IOP before offering treatment to a glaucoma suspect. CDR, CCT, and family history were also considered by over 80% of participants while life expectancy appeared to be the least important factor. However, age and PSD, both included in the calculator's risk estimate, were not considered by 35% and 40% of participants respectively.

Other factors mentioned by participants were: OCT, visual field, angle, vascular risk factors, race, IOP phasing, fellow eye, myopia, reliability to follow-up, previous ocular history, corneal diameter, refractive error, general health, and systemic diseases such as diabetes and hypertension.

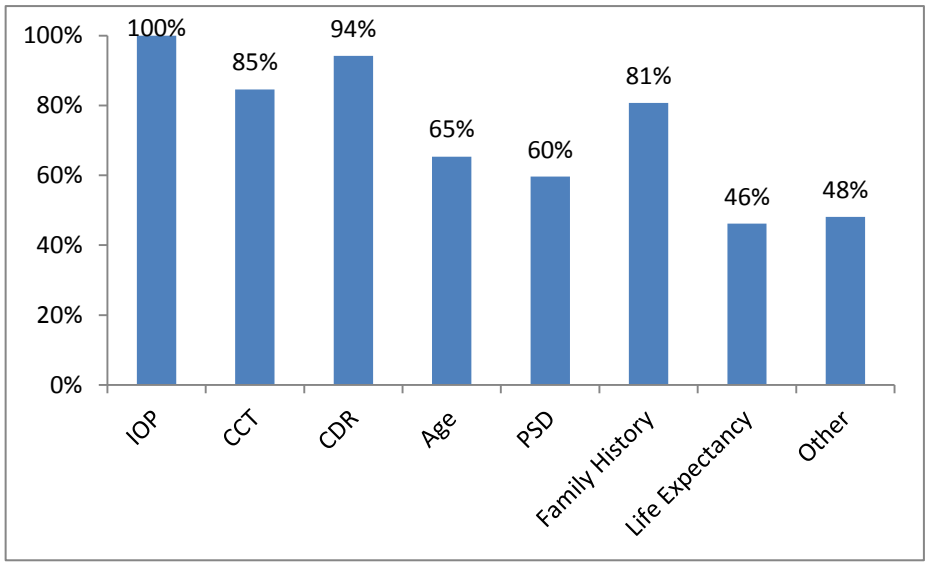


Figure 2-8: Factors Participants Cited They Consider Before Initiating Treatment for a Glaucoma Suspect

Discussion:

Results show that for about 40% of participants, use of the calculator would be beneficial in helping them incorporate risk due to patient age and PSD into their decision, while about 15% would benefit from incorporating CCT into their results. Future work may incorporate analysis of the weight of the different patient factors on participants' treatment decisions based on regression analysis rather than self-reported answers. This would be advantageous as self-reported results are likely to be exaggerated and also because the latter analysis would show direction of treatment trends (i.e. whether each factor increased/decreased a participant's likelihood to treat), which Boland et al. found were sometimes opposite to expectations.

It is interesting to note that age and PSD are the same two factors that Boland et al. found, based on regression analysis of decision recommendations, were not significant in shaping physician decisions - despite their importance in the OHTS results. This may suggest that use of the calculator can help clinicians better incorporate these factors into their decisions.

2.9 Overall Discussion and Contribution to Existing Knowledge

This study provided evidence of: (1) the effect of the presence of a decision support tool on user decisions, even before its presentation, (2) the newly introduced homeostasis mechanism which may account for some automation bias errors previously observed but unexplained in the literature, (3) the role (although weak) of uncertainty and direction of advice on changes in user decisions after the presentation of tool advice, and (4) differences between self-reported trust and perceived tool reliability and differences between objective measures of difficulty defined in terms of variance and subjective measures of difficulty defined in terms of uncertainty. Study results showed no significant differences in user decisions depending on the sequence of presentation of advice, even when analyses incorporated error type. Study results also did not reveal differences in the amount of change in user decisions after being presented with tool advice based on the maximizer-satisficer personality trait.

From a medical viewpoint, the study provided evidence of: (1) strong disagreement with NICE guidelines even among UK professionals and even after use of the calculator, (2) a significant effect of the calculator on user decisions, increased user confidence, and increased intra-user consistency in decisions, (3) differences in treatment recommendations based on where a participant practices, and (4) lack of consideration for patient age and PSD in clinicians' treatment decisions. Results showed no significant differences in treatment recommendations based on specialty or profession.

This summary of the overall contributions of the study from a human-computer and medical viewpoint are elaborated on in the discussion that follows.

According to our knowledge of the literature, this experiment is the first to examine the effect of the sole **presence** of tool advice. We found evidence that the mere knowledge that tool advice will be provided is sufficient in significantly altering users' behaviour. This result may suggest different levels of pre-processing of information by users depending on the presence of a decision support tool.

Furthermore, results explained the direction of this difference: that users are less likely to suggest treatment in their pre-advice decisions when they know they will receive advice. We consider this to be direct evidence of the homeostasis mechanism, where users increase the threshold with which they respond to cues. This **cognitive mechanism**, suggested by researchers at the Centre for Software

Reliability, extends established explanations in the literature for automation-induced errors such as "complacency", "overtrust" and "overreliance".

Insight into the cognitive mechanisms involved when using decision support tools can help designers establish more effective **strategies to mitigate automation-induced errors**. For example, one protocol used to combat automation bias in mammography does not allow users to change their pre-advice decision from "target" to "no target". Although such a protocol may help prevent some human omission errors caused by tool false negatives, it would, on the other hand, exacerbate effects of the homeostasis mechanism and thus possibly introduce a new set of errors.

Assessing the effects of the presence of a tool before the actual presentation of its advice is also important in the **design of future experiments** analysing the efficacy of new decision support tools. Many of these studies compare the decisions of a user before and after advice is presented. We showed that the mere knowledge that advice will be presented can alter the cognitive mechanisms of the user and thus to effectively assess the effect of a decision support tool, a user's post-advice decision should be compared to a control group that does not receive tool advice nor knowledge of the presence of tool advice.

Changes after receiving advice were greater and more often when users did *not* know beforehand that they will receive advice. This too may indicate that users were not being complacent and overreliant on the advice of the tool; thus emphasizing further the need to explore cognitive mechanisms, such as the homeostasis mechanism, that go beyond conventional explanations for AB.

Results did not support the idea that advice carries a higher weight when presented simultaneously versus sequentially. Results also did not show strong differences in responses based on the **sequence** in which advice was presented, even when analysed according to error type. This is in contrast to several medical studies which reported that the simultaneous mode reduced observer vigilance (Van Dongen, 2013; Taylor, 2008b), sensitivity (Taylor, 2008b) and performance (Zheng, 2004). Future work may explore the difference between the two sequences while taking into account case difficulty and correctness of advice. Future work may also look at such differences in a different context as results are likely to be affected by the specific DSS used.

Besides showing a weak relationship between the amount of **change** in a user's decision when presented with tool advice and (1) increased uncertainty in the pre-advice decision, and (2) direction of change towards the "safer" side of treating, results also drew attention to two interesting differences: (1) the difference between trust and perceived tool reliability and (2) the difference

between objective measures of difficulty defined in terms of variance or subjective measures defined in terms of uncertainty.

For almost all the participants who gave different ratings to the two measures, a higher rating was given to perceived tool reliability compared to trust. This could indicate that users may perceive a tool as reliable but still choose not to trust it, perhaps because confidence in their own ability exceeds confidence in the tool's reliability. This result could also suggest that the two terms, **trust and perceived tool reliability,** may be measuring different concepts where perceived reliability measures a person's opinion of the tool and trust measures a person's likelihood that they will use its advice.

Although often used interchangeably, it is important also to distinguish between **uncertainty and difficulty** which may not always represent the same phenomena. Measuring difficulty based on uncertainty relies on users acknowledging the difficulty of a case in the level of the uncertainty they express in their decision, while measuring based on variance is a more objective measure across participants. We notice that in our dataset, the two measures were only moderately related and were not always consistent.

Also, this study attempted to add to the small corpus of literature on **personality traits** and their effect on automation use (Parasuraman, 2010). While Ho et al. (2005) cited risk-aversion character as a factor in determining reliance on tool's advice, Singh et al. found no relationship between automation complacency and differences in extraversion-introversion personalities (cited by Parasuraman, 2010) and Prinzel et al. also found no such relationship with boredom proneness or absentmindedness (cited by Parasuraman, 2010). Our study adds maximizer/satisficer tendencies to the list of personality traits apparently not affecting use of tool advice.

When presented with a risk estimate, participants' treatment recommendations (1) changed in 38.7% of cases, (2) became more confident, and (3) became less inconsistent (measured as the sum of the variances of risk in each participant's three decisions (Treat, Unsure, and Don't Treat)). Contrary to Cheng et al. (2009) who believed that the **calculator** did not have a strong **influence** on users' decisions, we found that users significantly changed their decisions after the presentation of advice. However, unlike Boland et al. (2008) we found that the average treatment threshold increased, rather than decreased, with the calculator. This contradiction in findings suggests that the exact effect of the glaucoma calculator remains undetermined. Naturally, this effect depends on the cases chosen and on the participant sample.

It is interesting to note that there was very strong disagreement with **NICE guidelines**, even among UK participants. However, these results should be observed with caution because half of the hand-

picked cases we selected were chosen especially for their stark contradiction between NICE guidelines and expert opinion. Nonetheless, these results suggest that either ophthalmologists and optometrists in the UK need to become more aware of the guidelines (especially that only a third self-reported that they are familiar with the guidelines), or that current guidelines need to be revisited to reflect modern advances in research and current treatment trends.

Our exploration attempted to takes previous results one step further by exploring as stated by Boland et al. (2008), whether "recommendations [were] more ''correct'' and [led] to better patient outcomes". The difficulty in measuring this "**correctness**" is the absence of a clear golden truth for each case. We found, like Boland et al., that the calculator was effective in increasing the intra-participant consistency of user decisions. Undoubtedly, consistency is an important measure of the quality of medical care. However, we also found that the calculator failed to help participants better conform to either NICE guidelines or expert opinion. This result is not entirely surprising as the three references base their recommendations on different criteria. However, these discrepancies should serve as an eye opener to unify guidelines on the treatment of glaucoma suspects, which would make the job of medical professionals easier and the quality of care more consistent across hospitals.

Results showed that for about 40% of participants, use of the calculator would be beneficial in helping incorporate the risk of patient **age and PSD** into their decision, while about 15% would benefit from incorporating CCT into their decisions. It is interesting to note that patient age and PSD are the same two factors that Boland et al. found, based on regression analysis of decision recommendations, did not have a significant effect on physician decisions, despite their importance in the OHTS results. Future work may incorporate analysis of the weight of the different patient factors on participants' treatment decisions based on regression analysis rather than self-reported answers.

Compared to Boland et al. (2010), we found no differences in treatment recommendations depending on **specialty**. We also found no differences in recommendations based on **profession**. This is a constructive result as many glaucoma suspects in the UK are seen by optometrists and consistency in ophthalmologist and optometrist decisions is thus important. Finally, the fact that participants **outside the UK** were more likely to suggest treatment is likely the result of different training, guidelines, and economic considerations.

This study did not require participants to assess the appearance of the optic disc or interpret the full results of a visual field test. Decisions may be affected by such additional information as well as other factors such as patient preferences, general health, and ocular history. Although these limitations also exist in previous studies and affect our medically-related analyses, they do not affect our study of the more general cognitive mechanisms involved. However, future work, especially from a medical

viewpoint, may benefit from such an enrichment of patient cases. Future work may also consider the extendibility of results, showing the effect of the presence of a decision support tool on user decision and the newly introduced homeostasis mechanism, to other contexts.

Chapter 3: The Use of Diversity to Improve Double Reading Benefits in Breast Cancer Screening Programs

In Chapter 1 we presented a review on automation bias and in Chapter 2 we presented an experiment we designed to explore causal mechanisms and factors affecting automation bias. However, the ultimate goal when studying automation bias is to reduce, or ideally eliminate it. We present two separate experiments exploring potential mitigation strategies: in Chapter 3, using benefits of diversity between system components and in Chapter 4, using personalized tool settings (sensitivity / specificity combinations).

Alberdi et al. (2010) raise an important issue: in some decision support systems, it is not the tool's overall reliability that may be most important, but the diversity between the tool and the user. Diversity is used to describe an intuitive notion of how focussed the automated tool is on the true needs of the users (Alberdi, 2010), or in other words, how different the human and tool errors are. For example, if the tool is improved such that it prompts targets that the user already detects on their own, then little overall benefit will be observed in system performance (Alberdi, 2005; Philpotts, 2009; Madhavan, 2007; Petrick, 2008).

We attempted to briefly study the effects of diversity on system performance. Unlike the other experiments presented in this thesis, the system in this context is composed of two humans rather than a human and a tool. The main reason behind this choice was our limited resources, as this study is based on a retrospective analysis of existing data. Nevertheless, similar to results from the judge-advisory system literature looking at human + human systems, and which have been successfully integrated into the automation bias literature (Goddard, 2012; Madhavan, 2007), we argue that similar benefits from diversity can be obtained by applying such ideas to decision-support systems, and hope that future studies will build on the results we obtained by extending them to systems composed of a human and automated tool.

The domain of this chapter is reading of X-ray images in cancer screening. The data we used came from a previous study (Taylor, 2004) where 50 radiologists, radiographers and breast clinicians (collectively often referred to as "readers" in this context), each analysed 180 mammograms - 60 with cancer and 120 normal. Each reader marked each of the cases on a scale of 1 to 4 where (1: recall, 2: discuss but probably recall, 3: discuss but probably not recall, 4: no recall), and results were collapsed to a binary scale (recall and not recall for analysis). "Recall" in this context refers to the decision of a reader to see the patient again for further tests (i.e., a biopsy or further imaging) to verify the reader's suspicion that the patient may have cancer.

We simulated double reading pairs using this data. Double reading (DR) is the interpretation of a single medical image by at least two pairs of eyes. In a double reading setting, each reader gives their decision to recall or not recall a given case, and the two decisions are combined to obtain a single decision for the case. To combine decisions, we used the simple OR Recall Rule where a patient is recalled if at least one of the two readers decides to recall the case. The simulations resulted in a rich set of double reading pairs and their varying performance, which we then used to retrospectively investigate whether double reading benefits could be increased by forming complementary reader pairs according to indicators of user ability. In other words, we explored whether diversity between the two system components (the two readers) could improve system performance (decision of the double reading pair).

Although our main goal was to explore the benefits of diversity on system performance, our simulations also have important implications to the medical community. Breast cancer screening programs exist in over 22 countries worldwide (Sharpiro, 1998), and at least 12 of these countries and 90% of UK hospitals implement double reading (Smith-Bindman, 2003). With such a large number of hospitals implementing double reading, there is naturally major discussion in the literature regarding its overall effectiveness compared to single reading and compared to single reading with a computer-aided tool. There is also considerable literature examining how the two reader decisions should be combined and how disagreements should be handled. There is less literature in the double-reading context looking at the independence and order of the two decisions. Furthermore, there are even fewer papers examining the effectiveness of double reading based on which reader is paired with whom. Our work helps address this latter gap to motivate managers to consider concepts of diversity when forming double-reading pairs and to assist them in deciding how best to form these pairs.

We used formal statistical tests, information theory, and regression analysis to analyse the data. Our results show that some forms of pairing by complementary ability levels, compared to random pairings, can significantly improve sensitivity, with an insignificant effect on specificity. Results also suggest that pairing readers according to convenience could be significantly less effective. Our work has taken the first step of showing the benefits of mechanically combining two diverse components, but future work may take this further by considering how interactions between the two system components may affect these benefits, especially when the second component is an automated tool rather than human.

In the rest of this chapter, this work on double-reading pairs is described in more detail. The chapter starts with a brief review of the double-reading literature and motivation for the research in Section 3.1. Section 3.2 defines our research questions while Section 3.3 presents a logical motivation for the work. Section 3.4 describes the data and methods used. Section 3.5 contains the results. Section 3.6

discusses study limitations. Finally, Section 3.7 presents the overall discussion and contribution of the work. In the thesis appendices, two abstracts summarizing the work from this chapter and which were presented at two international medical decision-making conferences, can be found.

3.1 Background

We begin this chapter with a short overview of the relevant literature. Section 3.1.1 presents a brief discussion of the importance of tool diversity in the automation bias literature while Section 3.1.2 presents related work from the medical double-reading literature.

3.1.1 Tool Diversity in the Automation Bias Literature

Alberdi et al. (2010) raise an important issue: in some decision support systems, it is not the tool's overall reliability that may be most important, but the diversity between the tool and the user. Diversity is used to describe an intuitive notion of how focussed the automated tool is on the true needs of the users (Alberdi, 2010), or in other words, how different the human and tool errors are. For example, if the tool is improved such that it prompts targets that the user already detects on their own, then little overall benefit will be observed in system performance (Alberdi, 2010).

As noted early on by Sorkin et al. (1985), optimizing a tool's performance alone or a human's training alone may be ineffective in optimizing overall system performance. Such a lack of improvement in overall system performance could, for example, be a result of optimizing the tool to improve its performance on a set of cases that the human is already successful at. In fact, researchers argue that a poor stand-alone tool performance is permissible as long as the tool prompts targets typically missed by the human (Strigini, 2003; Lawrence, 2010). Diversity can also help reduce system false prompts caused by automation. For example, Kobayashi et al. (1999) reported that it was not very difficult for users to disregard tool false positives because they were generally different from those of the users. Moreover, diversity could help decrease users' cognitive load, for example, by designing tools that only prompt features users have not already noticed in their initial decision (Alberdi, 2014). In fact, a study from the JAS literature cited by Bonnacio et al. (2006), suggests that users would prefer such advice as judges sought more advice from advisors with unique information compared to advisors with information redundant to that of the judges.

Currently, many tools seem not to be very diverse from their users, suggesting that this, rather than the tool's stand-alone accuracy, is an area for improvement. For example, Alberdi et al. (2005) reported, when analysing a CAD tool used to detect breast abnormalities, that the tool and user errors were strongly correlated. Philpotts et al. (2009) also stated that CAD does not do well with subtle lesions, although these are the lesions that radiologists need the most help with. Likewise, others report that cases that were hard for users were also hard for the DSS (Madhavan, 2007; Petrick, 2008).

Building tools that aid the user in cases where they need the most help is not necessarily easy, but "building the tools on principles different from replicating human behaviour" is an important starting point (Alberdi, 2005). Some tool designers have already begun incorporating concepts of diversity in their designs. For example, Balleyguier (2005) described a CAD tool designed to specifically identify lesions that are difficult to detect with conventional radiological procedures and which are smaller than 3 cm (does not prompt larger lesions as they are usually correctly diagnosed by radiologists). This limits the rate of tool false positives and addresses the needs of users by helping them detect the most difficult lesions - although it does limit the use of CAD alone. Such pursuit of diversity is promising and has been effective in other fields (Alberdi, 2005).

We attempted to briefly study the effects of diversity on system performance. Unlike the other experiments presented in this thesis, the system in this context is composed of two humans rather than a human and a tool. The main reason behind this choice was our limit in resources as this study is based on a retrospective analysis of existing data. Although double reading pairs are systems consisting of two human components, many of the concepts of diversity inherent in the analysis are applicable to systems consisting of a human and a tool, and we hope may be extended to such systems in future studies. Before describing the study, we present a short overview of relevant work from the double-reading literature.

3.1.2 Double-Reading Literature

National, regional, or pilot population-based breast cancer screening programs exist in at least 22 countries worldwide (Shapiro, 1998), and at least 12 of these countries and 90% of UK hospitals implement double reading (Smith-Bindman, 2003). Double reading (DR) is the interpretation of a single medical image by at least two pairs of eyes. In a double reading setting, each reader gives their decision to recall or not recall a given case, and the two decisions are combined to obtain a single decision for the case.

With such a large number of hospitals implementing double reading, there is naturally major discussion in the literature regarding its overall effectiveness compared to single reading (Taylor, 2008; Harvey, 2001). A systematic review of seventeen studies comparing double reading to single reading showed that double reading increases cancer detection (Taylor, 2008c), and Dinnes et al.'s (2001) review of ten studies reported an improvement of 3-11 per 10,000 women screened. However, both studies also reported an increase in recall rate. This rate is not insignificant. As Hartswood et al. (1996) report, only 10% of recalled cases represent true cancer patients. The unnecessary follow-ups and possible biopsies associated with such false alarms cause emotional, physical, and material costs that should not to be ignored (Roehrig, 1999). Other studies compare

double reading to single reading with a computer-aided tool (Marten, 2004; Gilbert, 2008). Gilbert et al.'s study showed that DR and single-reading + CAD are equally effective at cancer detection while Marten et al. (2004) showed similar results limited to CAD + experienced readers only.

3.1.2.1 Differences in How Double Reading Decisions are Combined

Besides comparing the efficacy of the double reading procedure to single-reading and single reading + CAD, there is also considerable literature examining how the two reader decisions should be combined and how disagreements should be handled. The simplest method for combining decisions is the OR recall rule where a patient is recalled if at least one of the two readers decides to recall the case. Harvey et al. (2003) evaluated the performance of this method and found that the second reader increased cancer detection by 6.3% and increased recall rate by 1.5%. Another popular method is to arbitrate when there is disagreement between the two readers. Double reading when combined with arbitration has repeatedly been shown to increase cancer detection rates but decrease recall rates (Mucci, 1999; Taylor, 2008c). A third method is dual reading which is defined as a pair of observers reading the mammogram simultaneously while interacting and discussing any differences in opinion. Quekel et al. (2001) compared double reading to dual reading and found that double reading increased sensitivity but decreased specificity while dual reading increased sensitivity while maintaining specificity. Matcham et al. (2004), proposed a novel recall strategy where all mammograms marked for recall were subjected to consensus opinion. Their results showed a significant decrease in recall rates and the consensus meetings were also described as a strong forum for teaching and an asset in reducing stress levels amongst readers.

3.1.2.2 Differences in the Order and Independence of Double Reading Decisions

As noted in the previous section, there are many different ways that the two reader decisions can be combined and disagreements resolved. As Elmore et al. (2007) state, "comparing double reading studies is less like comparing apples to oranges than like comparing apples to asparagus". Yet another important difference is not in how the decisions are combined, but in their independence and order. In many double-reading settings, the second reader is not blinded to the first reader's decision and naturally this may introduce a form of bias (Liston, 2003). In some cases, the result of the first reader may lead to "carelessness or prejudice" by the second reader (Quekel, 2001). Even if the second reader is blinded to the first reader's decision, just knowing that they are the second reader may introduce some bias in their decision. In an important study by Swensson et al. (1985), authors compared the detection rate when radiologists were asked to report all significant chest findings, versus when they were

asked to evaluate findings in particular film locations. This latter condition simulated readings where indications for an exam are passed on from a first reader. Swensson et al. (1985) found that readers performed better in the former scenario and that examining films where possible abnormalities have been indicated can lead to excessive false positives.

In cases where the decision is not mechanically combined, but is based on discussion and consensus between two or more readers, other important psychological factors may also come into play. For example, "there are problems with consensus reading in situations where one reader is the more dominant personality leading to overriding of the other readers views" (Cornford, 2005). In fact, in one study, consulting radiologists conformed to the majority opinion even when it was wrong in 48% of cases and behavioural studies attribute such behaviour to peer pressure (cited by Hillman, 1976). This type of pressure has led some researchers to suggest non-interactive decision making techniques such as those based on the Delphi model. Although its clinical efficiency is questionable, in one experiment, this model produced an additional 20% mean improvement compared to interactive readings (Hillman, 1976).

Although "it is recognized that the second reader is influenced by the first reader's decision" (Liston, 2003) and although some of these concepts have been discussed in the psychology literature, little work has been done to study these effects in a double-reading context or to consider their effects when deciding appropriate hospital protocols. Interesting questions that would be worth further research include:

- Should the second reader be blinded from the first reader's decision? If not, then how much information should be shared – the recall/no recall decision, or a complete highlighting of suspicious areas on the mammogram?
- In the case that decisions are not blinded, who should read first? For example, how would a junior radiologist react to reading a mammogram previously marked by a senior radiologist?
- Even if decisions are not blinded, is the second reader's knowledge that they are in fact the second reader sufficient in significantly affecting their decision or vigilance in searching the mammogram?

Although these are interesting areas for further research, they do not directly address our objectives, nor are they easily addressed using only the data already available at the Centre for Software Reliability. Instead, these questions would require a new set of experiments.

3.1.2.3 Effectiveness of Double Reading Based on Strategic Pairing of Readers

Unlike literature exploring how decisions should be combined and the order and independence of decisions, there are very few papers examining the effectiveness of double reading based on which reader is paired with whom. Perhaps the most informative article in this respect is by Beam et al. (1996) who explored the gains and losses from independent double reading and concluded that readers can form complementary or non-complementary reader pairs. Beam et al. used the Variability In Diagnostic Interpretation-Screening Mammography (VIDI-SM) study to simulate 131 unique double reading pairs using the OR recall rule.

Beam et al. were able to identify pairs where double reading increases readers' sensitivity with a small decrease in specificity, pairs where small gains in sensitivity come at the cost of large decreases in specificity, and other pairs where sensitivity and specificity hardly change. In numbers, sensitivity increases ranged from 4% to as much as 42% in their study, while false positive rates increased from 1% to as much as 47%. Interestingly, authors also noted that adding the reading from a more experienced radiologist does not always improve the sensitivity of a less experienced radiologist.

Based on these observations, Beam et al. concluded that the decision to implement double reading should be decided on an individual basis after knowledge of individual reader's "sensitivities and specificities as well as the degree of their diagnostic concordance".

In another paper, Marten et al. (2004) explored the effect of replacing one radiologist in a double reading pair with a CAD tool for the detection of pulmonary nodes. Although the primary goal of this study was not to explore the effect of reader variability on the performance of a given double reading pair, authors noted that the performance of an experienced radiologist was not substantially altered by dual reading with an inexperienced radiologist. Marten et al. also predicted that the very good inter-observer agreement among experienced radiologists, measured using Cohen's Kappa coefficient, suggested that a two-experienced reader-consensus would result in a roughly similar performance compared to a single experienced radiologist.

These two papers point at the importance of pairing readers strategically. However, the UK National Institute for Health and Clinical Excellence (NICE) do not provide protocols for this strategic pairing. The UK NHS do provide some suggestions as they state in their Quality Assurance Guidelines for 2011 that "inexperienced readers should be paired with experienced readers and, ideally, readers with high recall rates should be paired with readers who have below-average recall rates and low cancer miss rates." No empirical evidence or rationale for

these specific pairings are provided, nor is it clear that the population of existing record would allow implementation of these policies.

3.2 Research Questions

A summary of the different double-reading possibilities presented in the brief literature review in Section 3.1 are summarized in Figure 3-1. **How** decisions are **combined** represents the area most covered in the literature, although studies simultaneously comparing several different methods are scarce. **How** decisions are **ordered** represents a study of the psychological effects inherent in the decision-making process and are not well covered in the double-reading context. The questions of **Who**? and **Which cases**? both shaded in a darker colour in the figure are areas new to the double-reading literature which we proposed and analysed for their effects in improving the double-reading procedure as a system.

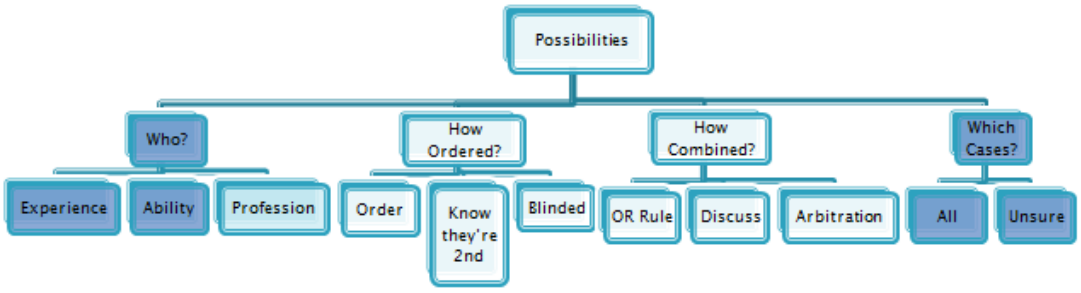


Figure 3-1: Summary of Double-Reading Possibilities Highlighting Our Contributions

The first question we addressed was which cases should be double read – all cases or only a subset? Our aim was to minimize the number of cases that needed to be double-read by identifying a subset of cases that would be as comparable in effectiveness to double reading all cases. Although this is an interesting research question, it is less directly related to the focus of this thesis. Our work on this was presented at the Society for Medical Decision Making Meeting in Antwerp, Belgium in June 2014. All discussion and results from this research question are presented only in Appendix 3A and lightly touched upon in Section 3.6. The rest of this chapter focuses on our work in answering the question concerned with how double reading pairs could be strategically formed.

Based on our literature review, little focus has been given to protocols governing the pairing of readers. For example, should these pairings be left to convenience in terms of radiologists' timetables or would it be more effective to pair them according to a more meaningful criterion? If they are to be paired according to a specific criterion, then what criterion would be most effective: ability (sensitivity, specificity, or other measurements), experience, or profession (radiologists with

radiographers)? Only the last category has been touched upon in the literature, especially that increased demand for manpower has created a skill mix in many screening tasks (Taylor, 2009; Wivell, 2003). In our retrospective study, we ask:

1. Is there a significant benefit from pairing lower ability readers with higher ability readers (complementary pairs) rather than conveniently leaving pairings to a random arrangement? (Section 3.5.1)
2. If there is an improvement in performance between complementary double reading pairs compared to random pairs, what is the best way to create these complementary pairs? (Section 3.5.2)

3.3 Motivation

As suggested from previous work (Beam, 1996), the performance of a double reading pair will vary depending on certain factors concerned with the readers themselves, and not just how their decisions are combined. To help illustrate this, we are fortunate to use data from an independent UK clinical trial where 50 readers each read 180 mammograms – 60 with cancer and 120 normal (Taylor, 2004). This data is described in more detail in Section 3.4. We follow Beam et al. (1996)'s lead and choose an arbitrary reader, Reader15, from our dataset of 50 readers. We simulate the effect of pairing Reader15 with three different readers: Reader5, Reader33, and Reader44 from our dataset as shown in Figure 3-2. As per the OR Recall Rule, a patient is recalled if either reader chooses to recall. In the figure, each point represents the performance of a single reader or a double reading pair based on their sensitivity and specificity calculated based on their performance in reading the 180 cases in the dataset. "Performance" is defined as a reader's sensitivity and specificity, and "system performance" refers to a pair's sensitivity and specificity after double reading. Using the OR Recall Rule, this system performance may exhibit no, little or large gains in sensitivity compared to either individual reader's sensitivity. However, specificity of the pair will always be lower than the specificity of the individual readers, except in the best case in which it will be equal.

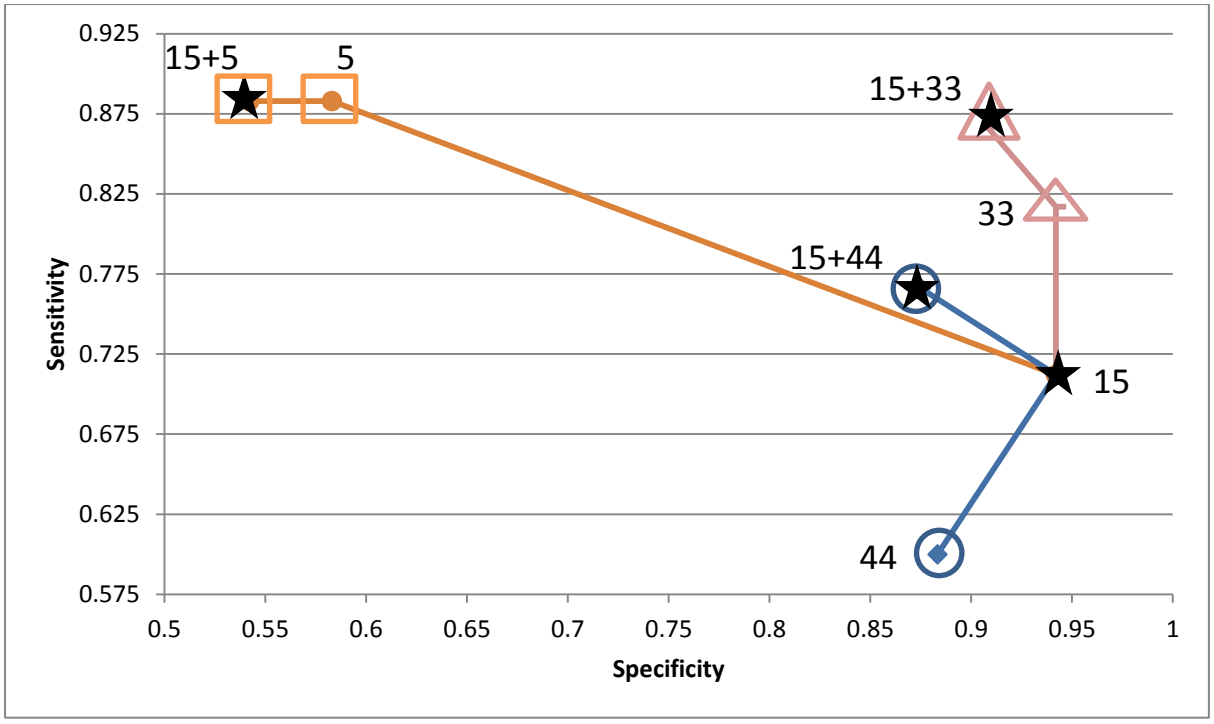


Figure 3-2: Illustration of Variable Effects of Double Reading Depending on Reader Diversity

Two readers' decisions are combined by a simple OR Recall Rule (recall if either chooses "recall") Reader15 is paired with: (1) Reader44 to show only a small gain in system sensitivity after the double-reading procedure – compared to the average sensitivity of individual readers in the dataset, (2) Reader5 to show system performance which is worse than one of the reader's independent performance, and (3) Reader33 to show system sensitivity that exceeds both readers' individual sensitivity, even though Reader33 is already the best reader in the dataset.

We choose these three particular readers to depict some interesting observations:

(1) *Pairing which yields system sensitivity that is below the average sensitivity of a single reader in this dataset*: We notice that Reader44 has a sensitivity lower than that of Reader15; in fact, Reader44 has a very low sensitivity compared to the average sensitivity in this dataset (0.75). Despite this, when Reader44 is paired with Reader15, the sensitivity of the pair is better than the sensitivity of Reader15 alone. However, after the added cost of double reading, the pair's sensitivity is still lower than the average sensitivity of a single reader in this dataset (0.75).

(2) *Pairing that results in system performance which is actually lower than the performance of one of the pair's readers:* When Reader15 is paired with Reader5, the sensitivity of the pair is the same as the sensitivity of Reader5. However, the pair's specificity is lower than that of Reader5. Thus, in this relationship, Reader5 is actually better off without the double reading procedure. This relationship is analogous to instances of automation bias in human + tool systems where system performance is actually lower with the tool than without it.

(3) *Pairing that yields system sensitivity that is higher than the sensitivity of the individual readers, even when one of the readers is already the best reader in the dataset:* When

Reader15 is paired with Reader33, who is the best reader in the dataset (according to Youden's index*), we notice a positive relationship in which the pair's performance shows improved sensitivity with a relatively small loss in specificity compared to each reader's individual performance. This latter combination is the type of relationship we seek in a double reading pair paired according to the OR Recall Rule.

*Note that Youden's index is defined as (sensitivity + specificity -1) and is one of many metrics that incorporate both sensitivity and specificity in a single index. It is chosen here somewhat arbitrarily, as a metric that would be easily available to hospital managers. However, in our extended analysis, we also assess accuracy according to other measures, such as positive likelihood ratio defined as (sensitivity / (1 – specificity)).

These observations may be relatively new to the double reading field. However, ideas of diversity and redundancy are accepted principles in many critical applications. Drawing on a paper by Popov et al. (2014), on the use of diversity in the development of software systems, we can help visualize the problem as shown in Figure 3-3. We define a demand space that represents all cancer cases in a given data set, and define a subset FN_A as the set of cancer cases that ReaderA fails to detect, and FN_B as the set of cancer cases that ReaderB fails to detect. Then, the set of cases that both readers fail to detect after double reading using the OR recall rule is the intersection of the two sets ($FN_A \cap FN_B$).

As shown in Figure 3-3, one of four scenarios can occur in the double reading procedure. In the most ideal pairing, scenario (a), the performance of both readers improves after double reading as they detect different cases, and because the sets are mutually exclusive, the pair are able to successfully detect all cancer cases. In scenario (b), the performance of both readers improves after double reading, but as there is an intersection between the two sets, there are still some cases that are undiscovered after double reading. On the other hand, in scenario (c) where FN_A is a subset of FN_B, ReaderA's performance does not improve after double reading. Finally, in scenario (d) where the two sets are equal, neither reader's performance improves from the procedure.

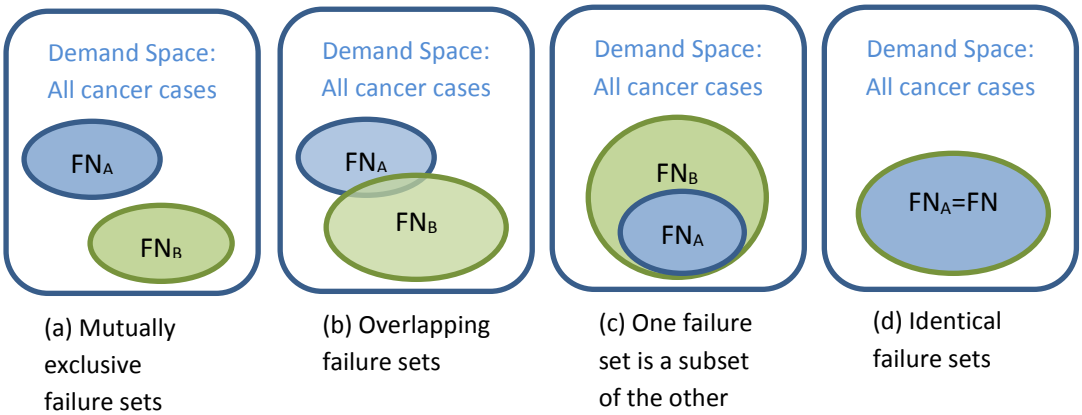


Figure 3-3: Possible Relationships Between the Failure Sets of Two Readers

Based on these four scenarios, we notice that the resulting sensitivity of the double reading pair could be improved by considering that it is affected by the size of the intersection between the two sets. This can be seen in turn as affected by two factors: the size of the original failure sets (sensitivity of readers A and B), and the overlap between the two sets (diversity between the two readers). Ideally, the smaller these failure sets are (i.e., the more sensitive the original readers are), and the more diverse they are from each other, the better the resulting sensitivity of the double reading pair will be.

Consideration of the resulting specificity of the double reading pair flips some of these concepts around. We redefine the demand space now as all normal cases in a given data set. FP_A is defined as the set of normal cases that ReaderA wrongfully recalls and FP_B as the set of normal cases that ReaderB wrongfully recalls. Now the false positives of the double reading pair will be the union, not the intersection, of the two sets (FP_A U FP_B).

Certainly in this purely human system there are confounding factors that complicate the above picture. Unlike the software context which is essentially deterministic (Salako, 2014), this system will exhibit much more noise. Research shows variability in readings of even the same case by the same person (Pitman, 2010). Also, as previously discussed, in many double-reading settings, the second reader is not blinded to the first reader's decision and naturally this may introduce a form of bias (Liston, 2003) or carelessness and/or prejudice by the second reader (Quekel, 2001). Even if the second reader is blinded to the first reader's decision, just knowing that they are the second reader may introduce some bias in their decision.

Despite these possible confounding factors, we note from this discussion that the system performance of the double-reading pair can vary depending on the performance of each reader and the diversity between the two. This motivated us to explore this variation empirically using data from a previous independent UK clinical trial.

3.4 Methods and Data

In our retrospective investigation, we were fortunate to be able to use data from a large independent UK clinical trial where 50 readers (30 radiologists, 5 breast clinicians and 15 radiographers) each read 180 mammograms - 60 with cancer and 120 normal (Taylor, 2004). The 60 cancers included in the study contained 20 false negative interval cancers, "cancers that had not been detected during screening using these mammograms but which were classified as false negatives when the mammograms were later reviewed by a panel of experts" and 40 cancers detected during screening (Povyakalo, 2013). The cases chosen included both easy and difficult cases, and contained a variety

of cancerous signs including masses, spiculated lesions, micro-calcifications and asymmetries (Povyakalo, 2013).

In the original study, each reader saw each case twice, once with and once without computer support. In both conditions, readers saw (1) "the actual films, positioned on a standard viewing roller" and (2) "a digitised version of the mammograms printed out on paper" (Alberdi, 2004). Digitised versions in the condition with computer support contained tool prompts. Each reader marked each of the 180 cases on a scale of 1 to 4 where (1: recall, 2: discuss but probably recall, 3: discuss but probably not recall, 4: no recall), but this scale was collapsed to a binary one (recall and no recall) for all our simulations. Also for purposes of our double-reading simulations, only readers' responses without computer support, mimicking how the films are read in everyday practice, were considered.

Researchers previously analysed this dataset for effects of CAD on reader decisions and published interesting results (Taylor, 2004; Alberdi, 2010). These were discussed in detail in Section 1.2.1. The 95% confidence interval for sensitivity with and without CAD was 0.76-0.80, while specificity with and without CAD was 0.81-0.87 and 0.81-0.86, respectively (Taylor, 2005). Most striking was the finding that use of CAD was associated with "a 0.016 increase in sensitivity for the 44 least discriminating radiologists for 45 relatively easy, mostly CAD-detected, cancers, [but] for the 6 most discriminating radiologists, with CAD sensitivity decreased by 0.145 for the 15 relatively difficult cancers" (Povyakalo, 2013).

From the 50 readers in this dataset, we identified the top ten and lowest 10 readers according to several ability factors: sensitivity, specificity, recall rate, and experience (measured in years). Thus, we created four High performing teams (H) and four low performing teams (L) of ten readers each. Using these eight teams we created six different types of *complementary* reader groups in which a member of a Team H, expected to be more effective, is paired with a member of a Team L, expected to be less effective. The complementary HL groupings are: (1) high and low specificity, (2) high and low sensitivity, (3) high and low experience, (4) high and low recall rate, (5) high sensitivity and high recall rate, and (6) high sensitivity and low specificity readers. In this paper, we will refer to these six groupings as forced diversity groupings. Each of these six groupings contains 100 reader pairs obtained from forming all possible unique double reading pairs between Team H and Team L of that group. The resulting decision of any given double reading pair is based on the OR recall rule where the pair recalls if at least one of the two readers chooses to recall (we consider ratings of 1 or 2 as a recall). It is worth noting that in this dataset, Groupings (5) and (6) are equivalent because the ten readers with the lowest specificity are the same ten readers that have the highest recall rates.

Similar to forced diversity groups, we define *unforced diversity groupings* to represent the random pairings that are common practice in many hospitals. A random pairing could result in two readers from team H (HH), two readers from team L (LL) or a reader from team H combined with a reader from team L (HL/LH). The union of all these possibilities would then be 100 possible HL reader pairs, 100 possible LH reader pairs, 100 HH reader pairs, and 100 LL reader pairs for a total of 400 pairs, including pairings of a reader with him/herself.

Another grouping often referred to in this paper is the *homogeneous grouping*, which represents the pairing of two readers from the same team (HH or LL). An example of a homogenous pair would be the pairing of two low sensitivity readers. Since each team has 10 readers, homogenous groupings have 100 possible double reading pairs, and only 45 unique pairings which exclude the pairing of a reader with him/herself.

3.5 Results

With the lack of focus in the literature on protocols governing the pairing of readers, we started our research by exploring whether there is in fact a benefit from forcing diversity rather than leaving pairings to a random arrangement. This is discussed in Section 3.5.1. Following this, we explored how best to create these forced diversity groupings in Section 3.5.2.

3.5.1 Comparison of Mean Sensitivities and Specificities of Forced Diversity, Unforced Diversity, and Homogeneous Groupings

For the three groupings described in Section 3.4: forced diversity, unforced diversity, and homogenous groupings, we measured the mean sensitivity and specificity of their corresponding pairs calculated over all 180 cases in the dataset.

Due to the fact that the forced diversity grouping is a subset of the unforced diversity grouping, we did not perform a t-test to test the significance of the mean of the forced diversity grouping and the unforced diversity grouping. However, we did perform Welch's Two Sample t-test to test for the significance of the mean sensitivities and specificities of the subgroup (forced grouping) and the subset (homogeneous grouping) of the overall group (unforced grouping) that does not include the subgroup. This corresponds to comparisons between the forced and homogeneous groupings, which are summarized in Table 3-1.

Table 3-1: Average Sensitivity of Forced Diversity, Unforced, and Homogeneous Pairs			
Grouping	**Average Sensitivity of Forced Grouping**	**Average Sensitivity of Unforced Grouping**	**Average Sensitivity of Homogeneous Grouping**
High and Low Specificity	0.869	0.856	0.842* *t(297.9)=4.07, p<0.0005*
High and Low Sensitivity	0.872	0.850	0.828* *t(251.9)=5.93, p<0.0005*
High and Low Experience	0.852	0.844	0.837* *t(265.6)=2.79, p= 0.00568*
High and Low Recall Rate	0.856	0.839	0.822* *t(290.5)=4.69, p<0.0005*
High and Low DA	0.868	0.864	0.859 *Non-significant t(253.7)=1.47, p=0.144*
High Sensitivity and Low Specificity = High Recall Rate and High Sensitivity	0.918	0.913	0.908* *t(262.5)=2.70, p= 0.00742*

* Symbolizes significant results

We note from the table that the general trend is that $\text{Sensitivity}_{\text{Forced}} > \text{Sensitivity}_{\text{Unforced}} > \text{Sensitivity}_{\text{Homogeneous}}$. We also note that the statistically significant changes are an increase in sensitivity in the Sensitivity, Specificity, and Recall Rate groups when diversity is forced.

We perform the same comparison for specificity and obtain the following results.

Table 3-2: Average Specificity of Forced Diversity, Unforced, and Homogeneous Pairs			
Grouping	**Average Specificity of Forced Grouping**	**Average Specificity of Unforced Grouping**	**Average Specificity of Homogeneous Grouping**
High and Low Specificity	0.631	0.653	0.675* *t(265.2)=3.27, p=0.00122*
High and Low Sensitivity	0.672	0.683	0.694 *t(264.6)=1.59, p=0.114*
High and Low Experience	0.722	0.726	0.730 *t(243.5)=0.815, p=0.416*
High and Low Recall Rate	0.623	0.644	0.665* *t(267.5)=3.28, p=0.00117*
High and Low DA	0.668	0.677	0.685 *t(304.9)=1.31, p=0.191*
High Sensitivity and Low Specificity = High Recall Rate and High Sensitivity	0.535	0.543	0.550 *t(230.1)=1.36, p=0.175*

* Symbolizes significant results

For specificity, we notice an opposite trend to sensitivity. In specific, we notice that $\text{Specificity}_{\text{Forced}} < \text{Specificity}_{\text{Unforced}} < \text{Specificity}_{\text{Homogeneous}}$. We also note that only groupings based on specificity and recall rate show a significant change in specificity.

3.5.1.1 Simulation of Forced Diversity and Unforced Diversity Groupings in a Virtual Hospital Setting

An important limitation to the analysis in the previous section is that it is based only on the top 10 and lowest 10 readers of each ability group from our dataset of 50 readers. In a real hospital setting, there will be mediocre readers not accounted for in this analysis. Thus, an important question that arises is whether these amplified results would still hold if they were also based on average rather than extreme readers.

To tackle this limitation, we:

1. Assumed a hospital consists of 6 radiologists based on the average facility size from (Beam, 1996), but rounded up to create even number teams that can be equally divided.
2. Randomly chose 10 of the 50 readers, and calculated all 210 possible unique combinations of 6-team hospitals from this pool.
3. For each hospital team, we divided the 6 doctors into high sensitivity and low sensitivity teams of 3 readers each.

4. Then we calculated the mean sensitivities and specificities of the unforced and forced diversity groups in each hospital.
5. Finally, we used a counter to determine which hospitals followed the general trends we observed in the previous section (i.e., $\text{Sensitivity}_{\text{Forced}} > \text{Sensitivity}_{\text{Unforced}}$ and $\text{Specificity}_{\text{Forced}} < \text{Specificity}_{\text{Unforced}}$)

The advantage of this analysis compared to the previous analysis in Section 3.5.1 is that it more closely mimics a real-world environment in terms of hospital size and varying reader abilities, as it includes the mediocre readers previously excluded. However, we only performed it based on the High Sensitivity + Low Sensitivity complementary group, and we used sampling with replacement.

The results appear consistent. In all 210 fictitious hospitals, forced diversity groups had a higher average sensitivity than unforced diversity groups. On the other hand, all 210 hospitals also had a lower specificity with forced diversity than with unforced diversity. We will explain this in a more general way in Section 3.5.1.2.

This bootstrap analysis that more closely mimics a typical hospital setting does not prove, but supports, our previous observation that forced diversity increases sensitivity and decreases specificity compared to random pairings. An important limitation of this modelling is that it is based on sampling with replacement.

3.5.1.2 Mathematical Examination of the Sensitivities of Forced and Unforced Diversity Groupings

Taking into light the results from the previous simulations, we settled to understand this phenomenon mathematically. Could we show, mathematically, that the sensitivity of the complementary double reading pair will always be greater than that of the random pair, as we observed empirically in our simulations?

To do this we relied on a publication by Salako et al. (2014). Salako et al. (2014) discuss a software project manager's choice between two alternative policies of the application of algorithms γ and δ. They define the decoupled alternative as the case where two teams of software developers A and B must each choose from one of the two algorithms independently. This may result in combinations: $\gamma\gamma$, $\delta\delta$, $\gamma\delta$, or $\delta\gamma$. In the double reading context, this decoupled alternative is equivalent to the unforced diversity groups. The second case the authors consider is the coupled alternative where the two software development teams are given freedom to each

choose one implementation algorithm but with the constraint that the two chosen algorithms must be different. In the double reading context, this is equivalent to the forced diversity groupings as it results only in γδ or δγ combinations.

Using the analysis in Salako et al., and applying it to the double reading context, we define *x* as a given patient case that is a cancer, and θ(x) as the likelihood that a random reader will fail on that case *x* (i.e. not recall it). To denote whether our readers are from group A or B, we define the labels $\theta_A(x)$ and $\theta_B(x)$ as:

$$\theta_A(x) := \theta_1(x|\text{reader group} = A) = \theta_2(x|\text{reader group} = A) \quad (1)$$

and

$$\theta_B(x) := \theta_1(x|\text{reader group} = B) = \theta_2(x|\text{reader group} = B) \quad (2)$$

Then for the unforced diversity groupings, the expression for the pair's likelihood of failure associated with demand *x* will be:

$$\frac{1}{4}(\theta_A(x))^2 + \frac{1}{4}(\theta_B(x))^2 + \frac{1}{2}\theta_A(x)\theta_B(x) \quad (3)$$

This is because there is a 0.25 probability that the pair will consist of two readers from group A, a 0.25 probability that the pair will consist of two readers from group B, and a 0.50 probability that they will be a mixed pair.

For the forced diversity grouping, the expression for the pair's likelihood of failure associated with demand *x* will be:

$$\theta_A(x)\theta_B(x) \quad (4)$$

This is because there is a 0.50 probability that the pair will consist of the first reader being from group A and the second reader being from group B and a 0.50 probability of the opposite order.

Now, to compare the likelihood of failure of the unforced and forced diversity groupings, we subtract expression (3) from expression (4) to yield:

$$-\frac{1}{4}(\theta_A(x) - \theta_B(x))^2 \quad (5)$$

As this difference will always be a non-positive number, this means the unforced diversity grouping's likelihood of failure will always be higher than that of the forced diversity grouping, except in the rare case where $\theta_A(x) = \theta_B(x)$ in which case the difference will be zero. Since the expected system probability of failure per demand (*pfd*) is given by a weighted sum of the system difficulty function over all demands, the expected sensitivity in the unforced diversity groupings will always be lower than the expected sensitivity of the forced diversity groupings.

This theorem explains the increased sensitivity of the complementary double reading pair. Similarly, it can be shown that the specificity of the complementary double reading pair will always be lower than that of the random pair.

An important condition of this theorem is that it is based on sampling with replacement. Also, this theorem does not refute the possibility that certain homogeneous pairs where a good reader is paired with another good reader (HH) may have a higher sensitivity than certain heterogeneous (HL) pairs, but it does show that any benefit from such HH groups will be masked by losses from the LL groups so that for any given case *x*, the forced diversity HL group will always have a lower probability of a false negative compared to the equivalent unforced diversity group (HH+LL+2HL).

3.5.2 Comparison of Forced Diversity Groupings

In Section 3.5.1 we observed that forced diversity groupings have a higher sensitivity compared to unforced diversity groupings, while empirical evidence showed that the decrease in specificity was often non-significant. Since in medicine high sensitivity is especially desirable, the subsequent natural question is then how to create these forced diversity groupings. We focus on comparing different forced diversity pairs created according to indicators of ability as recommended by NICE guidelines. This is one way that we can use to create the groupings, but future studies may also consider grouping according to other factors such as profession. In fact, Elmore et al. (2007) even suggested pairing readers according to their visual search patterns, although as the author noted, such patterns among readers are not easily determined. To compare the performance of the different groups we use three different methods: comparison of mean indices (Section 3.5.2.1), comparison using Information Entropy (Section 3.5.2.2) and regression analysis (Section 3.5.2.3).

3.5.2.1 Comparison of Mean Sensitivities, Specificities, Youden's Index and Positive Likelihood Ratios of the Forced Diversity Groupings

To tackle the question of which ability factor should best be used to create the forced diversity groupings, we calculated the average sensitivity and specificity for each of the six complementary groupings previously defined. We also calculated the average Youden's index defined as (sensitivity + specificity -1) and the average positive likelihood ratio defined as (sensitivity / (1 – specificity)) to use two metrics that incorporate both sensitivity and specificity in a single index. The results are shown in Table 3-3. To identify significance, we use 95% confidence intervals (CIs), presented in the table cells between parentheses.

Table 3-3: Performance Comparison of Forced Diversity Groups

Group	Average Sensitivity	Average Specificity	Average Youden's Index	Average Positive Likelihood Ratio
High and low specificity	0.869 (0.862, 0.876)	0.631 (0.620, 0.642)	0.500 (0.490, 0.510)	2.40 (2.34, 2.47)
High and low sensitivity	0.872 (0.867, 0.878)	0.672 (0.653, 0.691)	0.545 (0.528, 0.562)	2.95 (2.75, 3.15)
High and low experience	0.852 (0.844, 0.859)	0.722 (0.707, 0.737)	0.574 (0.561, 0.587)	3.30 (3.14, 3.47)
High and low recall rate	0.856 (0.849, 0.863)	0.623 (0.613, 0.634)	0.479 (0.470, 0.489)	2.31 (2.25, 2.37)
High recall rate and high sensitivity	0.918 (0.913, 0.924)	0.535 (0.518, 0.552)	0.453 (0.437, 0.469)	2.04 (1.96, 2.12)
High sensitivity and low specificity	0.918 (0.913, 0.924)	0.535 (0.518, 0.552)	0.453 (0.437, 0.469)	2.04 (1.96, 2.12)
High Youden's index and low Youden's index	0.868 (0.859, 0.878)	0.668 (0.650, 0.686)	0.536 (0.524, 0.548)	2.83 (2.68, 2.98)

Values between parentheses represent 95% confidence intervals.

We analyse confidence intervals where we know that the absence of an overlap between two intervals proves significance, although the opposite does not prove non-significance. We see from Table 3-3 that grouping high sensitivity and low specificity readers or high sensitivity and high recall rate produce the significantly highest average sensitivity: 0.918 (95%CI: 0.913, 0.924), while grouping according to experience produces a sensitivity of 0.852 (95%CI: 0.844, 0.859), significantly lower than all other groups. On the other hand, grouping according to experience produces the significantly highest overall specificity, 0.722 (95%CI: 0.707, 0.737) while combining high sensitivity and high recall rate (or high sensitivity and low specificity) produce the significantly lowest specificity, 0.535 (95%CI: 0.518, 0.552).

We notice that grouping by experience yields the highest average Youden's index 0.574 (95%CI: 0.561, 0.587), followed by grouping according to sensitivity 0.545 (95%CI: 0.528, 0.562). As with Youden's index, grouping by experience still maintains the highest positive likelihood ratio, followed by grouping according to sensitivity.

The high performance of the diverse experience groups may tempt us to suggest that this would be the best way to combine readers in this data set. However, we note that this is due to the group's significantly higher specificity while their sensitivity is significantly the lowest. In a clinical setting, higher sensitivity is often preferred over higher specificity (Dreiseitl, 2005; Halligan, 2011). The groups exhibiting the highest sensitivity also show a very high cost in specificity. Low specificity in a screening programme can be unsustainable and/or cause degradation of care. The sensitivity and Youden's index groups, second only to the experience group and significantly higher than all other groups, seem to maintain a reasonable balance between sensitivity and specificity.

We note from Table 3-3 that groupings can show a large magnitude of difference in performance. For example, in terms of sensitivity, there is a maximum difference of 0.066 and in the case of specificity a maximum difference of 0.187.

3.5.2.2 Comparison of the Expected Relative Entropy Provided by each of the Forced Diversity Groupings

We also compared the performance of the forced diversity groups using information theory. Information theory has been applied to a variety of clinical problems, and recently it has been used to create information graphs that can be used to evaluate and compare diagnostic tests. Besides being a visual comparison, information graphs offer an approach that balances between sensitivity and specificity, and also make explicit the relationship between pretest probability of disease and the information provided by a given test (Benish, 2002).

Entropy, a fundamental concept in Information theory is a measure of the uncertainty of the diagnosis, a random variable, while relative entropy is a measure of the distance between two probability distributions. Then the relative entropy of the post-test probability distribution with respect to the pretest probability distribution quantifies the amount of information gained by performing a diagnostic test. The expected value of this relative entropy or diagnostic information, I, can be expressed as a function of the pretest probability of disease P, sensitivity, Sn, and specificity, Sp as: (Benish, 2002)

$$I_E = P\,Sn \log_2 Sn + (1-P)(1-Sp) \log_2 (1-Sp) + P(1-Sn) \log_2 (1-Sn) + (1-P)\,Sp \log_2 Sp - c \log_2 c - (1-c) \log_2 (1-c) \quad (6)$$

where *c* is the probability of a positive test result:

$$c = (P)(Sn) + (1-P)(1-Sp) \quad (7)$$

To compare between the different forced diversity groups, we calculated and plotted diagnostic information as a function of pretest probability. The resulting graphs are shown in Figure 3-4. From the Figure we conclude that, for all values of *P*, the group that provides the most diagnostic information is the group paired according to experience followed by the group paired according to sensitivity, which is almost overlapping with the Youden index group. It is important to note that this ordering exactly follows the ordering of the groups according to their Youden index and Positive likelihood ratios discussed in the previous section.

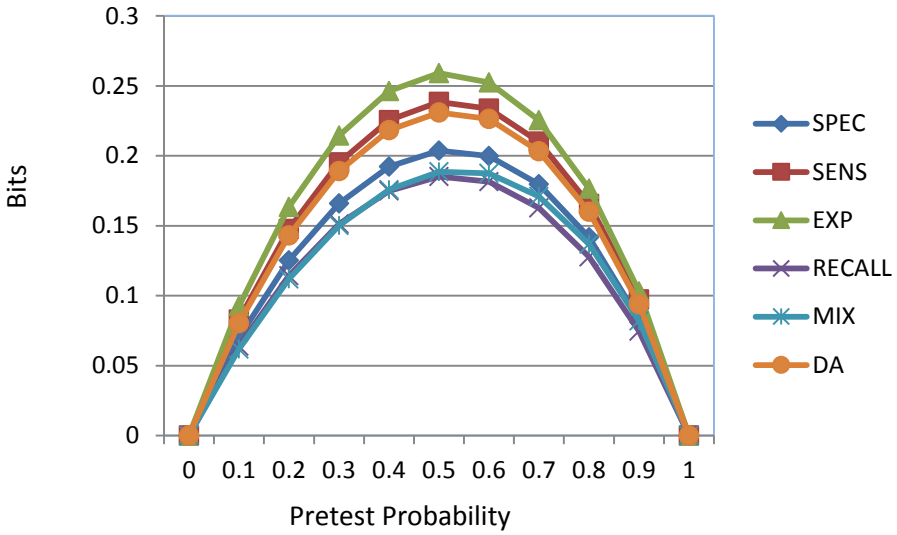


Figure 3-4: Expected Value of Diagnostic Information of the Forced Diversity Groups

3.5.2.3 Correlations Between Ability Measures

In order to help determine the most effective way to create forced diversity groups, it is not enough to look only at the differences in performance between the different groups as in Sections 3.5.2.1 and 3.5.2.2. Considering the relationships and dependencies between the various ability factors defined is also important. A simple method to determine this is to examine the correlations between the ability factors as shown in Table 3-4.

Table 3-4: Correlation Between Ability Measures					
	Sensitivity	**Specificity**	**Recall Rate**	**Experience**	**Discriminating Ability**
Sensitivity	1	- 0.61	+ 0.78	- 0.17	+ 0.25
Specificity		1	- 0.97	+ 0.23	+ 0.62
Recall Rate			1	- 0.23	- 0.41
Experience				1	+ 0.10
Discriminating Ability					1

Among the most important of these relationships is the high negative correlation (-0.97) between specificity and recall rate and the somewhat lower positive correlation between recall rate and sensitivity (0.78). The first may suggest that recall rate and specificity are redundant measures. Also of interest, is the very low correlation between experience and all other measures. This suggests that experience, measured in years, is not an accurate predictor of ability. In fact, the correlation between discriminating ability and experience is very low (0.10). If Youden's index is graphed against experience, we can visually see this very low correlation as in the graph in Figure 3-5. Graphs of sensitivity versus experience or specificity versus experience yield similar patterns.

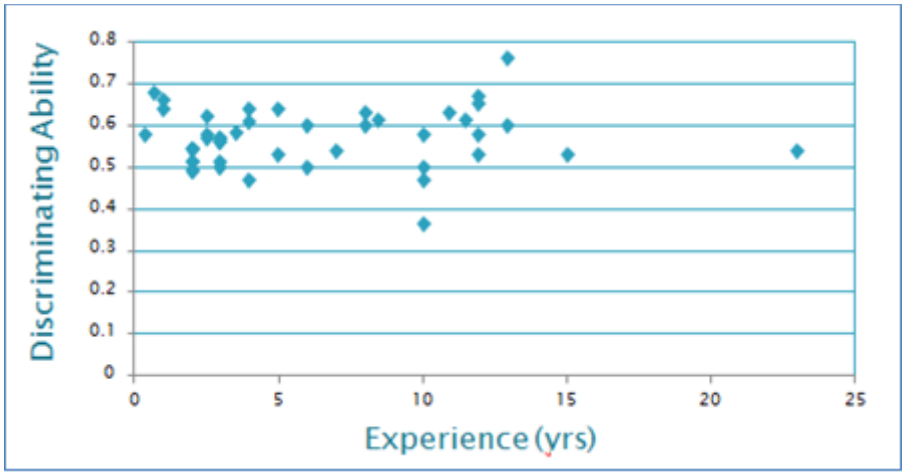


Figure 3-5: Correlation Between Experience and Youden's Index

In fact, the most experienced reader (by far since average experience was 6.2 years) had 23 years of experience and their Youden's Index was 0.533 (around the average in this dataset). On the other hand, one of the most novice readers with less than one year experience had the third highest discriminating ability: 0.672.

However, it is important to note that other measures of experience, not available to us in this dataset, might have been better correlated with ability. As noted by Elmore et al. (2007), not all

measures of experience are equally associated with the ability to interpret mammograms. For example, in their study, commonly regarded measures of experience such as number of years and average number of mammograms read per year were not correlated with sensitivity. However, a combined factor of the two, total number of lifetime mammograms read (average read per year * number years of experience), did yield a significant association with sensitivity.

3.5.2.4 The Use of Regression Analysis to Find the Best Determining Ability Factor

A different approach to determining the most effective ability measure to create forced diversity groups is using regression analysis. To do this, we started by setting the dependent variable as the Youden's index of the double reading pair. Then we defined the full model as the difference in each pair's reader abilities in terms of sensitivity, specificity, recall rate, and experience. We used this difference as a measure of the diversity between the two reader pairs and to examine which difference in the four measures had the strongest effect in determining the pair's resulting Youden index. Then, to reduce the dependent variables, we used backward selection based on AIC (Akaike information criterion), which is a relative estimate of the information lost when a given model is used to represent the data.

The smallest model that was equivalent to the 4-parameter model (ANOVA comparison results in a p-value=0.959) was based on the difference in sensitivities, recall rate, and experience of the two readers. Details of this model are shown in Figure 3-5.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5703753  0.0030149 189.183  < 2e-16 ***
deltarecall -0.0041507  0.0001500 -27.668  < 2e-16 ***
deltasens    0.4064089  0.0306775  13.248  < 2e-16 ***
deltaexp     0.0010367  0.0003184   3.257  0.00115 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06441 on 2066 degrees of freedom
Multiple R-squared: 0.2749,     Adjusted R-squared: 0.2739
F-statistic: 261.1 on 3 and 2066 DF,  p-value: < 2.2e-16
```

Although this model is a good fit to the data based on its residual standard error, adjusted R-squared and corresponding Q-Q plot, it is based on too many parameters to be truly useful in a clinical setting. However, it does reveal a few important points:

- Specificity is the least informative parameter in predicting a pair's Youden index. This coincides with the dependency investigation discussed in Section 3.5.2.3 as specificity and recall rate showed a strong negative correlation. However, it is interesting to note that removing specificity instead of removing recall rate significantly decreased the model's performance compared to the full 4-parameter model. Thus, according to this analysis, when removing one of these two redundant measures it was best to remove specificity as the redundant measure.
- Sensitivity and recall rate are the most significant factors in determining the resulting pair's Youden index.

Since the resulting model is based on too many factors to be truly useful in pairing readers in a hospital, we tried a different approach to determining the best ability factor to use to create the forced diversity pairs. If each of the six complementary groupings previously discussed in this paper is treated as an alternative model, then we can formally compare these non-nested models using the Vuong method. The null hypothesis is that the two models being tested are equally close to the data, while the alternative hypothesis prefers one of the two models.

If, as before, we set the dependent variable of the regression model to be the Youden's index of the double reading pair, then the models in order of superiority ($p<2e-16$) are pairing according to the:

1. Youden index of the two readers
2. Specificities of the two readers
3. Recall rates of the two readers
4. Sensitivity of one reader and the recall rate of the other reader
5. Experience of the two readers measured in years, or the sensitivities of the two readers

We observe from these non-nested comparisons that both sensitivity and experience are poor predictors of a double reading pair's resulting Youden index. This is especially interesting as pairing by experience is one of the groupings suggested by the UK NHS. We also note that if we were to base the pairings on only one ability measure, then the Youden's index of the two original readers followed by their specificities would be the best predictors of the pair's resulting Youden index.

How best to pair an entire pool of readers is a difficult question and poses many possibilities. In Section 3.5.2, we explored some of these possibilities using several different methods. A simple comparison of the resulting pairs' Youden's indices, positive likelihood ratios, and expected relative entropies suggested pairing readers' according to experience. However, a deeper look at this improvement showed that it was due to an increase in specificity at a high cost for sensitivity, a result that is generally undesirable in a medical context. Further analysis revealed that experience (measured in years), shows little correlation with discriminating ability. Instead, pairing according to sensitivity or discriminating ability seems to maintain a reasonable balance between sensitivity and specificity of the overall pair.

3.6 Study Limitations

An important limitation to this work is that results are based on the mechanical OR rule for combining readers' decisions while other methods such as dual reading or arbitration may show different effects. For example, although it may be beneficial to combine a reader with a high sensitivity with a reader with a lower sensitivity, this pair in a dual reading scenario may exhibit contrary effects. For example, if one of the readers possesses a dominating personality that controls the decision, the benefits of the complementary pairing may be overshadowed by such effects. In other words, our work has shown the benefits of mechanically combining two diverse components, but future work may take this a step further to consider how interactions between the two system components may affect these benefits, especially when the second component is an automated tool rather than human.

Thus, alternate methods to the mechanical OR rule for combining readers' decisions could be explored in future experiments or even simulated first. For example, Taylor et al. (2004) previously simulated double-reading with arbitration. In our work (Gadala, 2014) presented in Appendix 3A, we considered different ways that discussion, rather than the mechanical OR rule, could be simulated. We modelled the result of discussion according to four alternative scenarios described below:

- Ideal (always correct decision),
- Ideal limited (correct decision when readers disagreed),
- Authoritarian (decision of more experienced reader), and
- Confident (decision of more confident reader else OR recall).

Using these simulations to capture how we expect discussion may proceed between two readers, we were able to show that discussing cases identified as unsure by the first reader (rating of 2 or 3 on a 1 to 4 scale) could be significantly more effective than double reading all cases. We argue that future work may benefit from such simulations of discussions when looking at other methods of combining user decisions, and before tackling in an experiment.

Another important limitation to our work is that all the analyses were based on experience measured as number of years in practice as this was the only information available to us in this retrospective study. However, other measures of experience may prove more predictive in the performance of a double reading pair. As noted by Elmore et al. (2007), not all measures of experience are equally associated with the ability to interpret mammograms. For example, in their study, commonly regarded measures of experience such as number of years and average number of mammograms read per year were not correlated with sensitivity. However, a combined measure of the two, total number of lifetime mammograms read (average read per year * number years of experience), did yield a significant association with sensitivity. Incorporating other measures in future studies could prove useful in the comparison.

Our analyses are limited to the given data set and the criteria for creating complementary pairs may show different results for other populations. Future work may consider creating pairs not according to performance measures such as sensitivity and specificity, but according to mammogram-reading specific abilities. For example, readers could be paired according to who is more competent at identifying microcalcifications, or better at detecting masses, or those who quickly find asymmetries. This type of diverse pairing may be made possible using readers' self-reporting of strengths and weaknesses (Hartswood, 2003; Hartswood, 1999), or more formally using mathematical models which have been used in psychology not to detect different cut-offs employed by subjects in deciding when an item belongs to a particular category, but different criteria employed by subjects when categorizing items (Verheyen, 2013).

3.7 Contribution to Existing Knowledge and Overall Discussion

This work has taken a new perspective on double reading by asking not how opinions should be combined but which readers should be combined with whom. Results show that creating groups of complementary ability levels can significantly improve sensitivity, with an often insignificant effect on specificity, compared to random pairings. Results also suggest that pairing readers simply according to convenience could be significantly less effective. These results could be beneficial to hospital managers in improving the performance of their programs.

How best to pair an entire pool of readers is a difficult question and poses many possibilities. We explored some of these possibilities using several different methods for creating "diverse" pairs. A simple comparison of the resulting pairs' Youden's indices, positive likelihood ratios, and expected relative entropies suggested pairing readers according to experience. However, a deeper look at this improvement showed that it was due to an increase in sensitivity at a high cost for specificity, a result that is generally undesirable in a medical context. Further analysis also revealed that experience

(measured in years), shows little correlation with discriminating ability. Instead, pairing according to diversity in sensitivity or discriminating ability seems to maintain a reasonable balance between sensitivity and specificity of the overall pair. Whether these results would be reproduced in a specific hospital team of course depends on the specific readers available there (and which way their abilities differ for each case they examine); the importance of the present results is to show that it is worthwhile for administrators to check for these possible improvements.

The work presented in this chapter presented empirical and mathematical evidence of the benefits of diversity in improving the performance of human + human systems. Unlike the other experiments presented in this thesis, the system in this context is composed of two humans rather than a human and a tool. The main reason behind using a system composed of two humans was our limit in resources as this study is based on a retrospective analysis of existing data. Nevertheless, similar to results from the judge-advisory system literature looking at human + human systems, and which have been successfully integrated into the automation bias literature (Goddard, 2012; Madhavan, 2007), we argue that similar benefits from diversity can be obtained by applying such ideas to decision-support systems. Our work has shown the benefits of mechanically combining two diverse components, but future work may take this a step further to consider how interactions between the two system components may affect these benefits, especially when the second component is an automated tool rather than human.

Such results would have important implications, especially for designers of automated tools and especially because currently many tools seem not to be very diverse from their users. Rather than improving the overall performance of a tool, focus may be better shifted to assist users on cases where the most help is needed. In fact, as argued by Strigini et al. (2003), this may sometimes translate into accepting poor stand-alone tool performance as long as the tool prompts targets typically missed by the human. Diversity can also help reduce system false prompts and users' cognitive load and may, in fact, as suggested by a study from the JAS literature be preferred by users (Bonnacio, 2006). Future work, most interestingly, would benefit from showing that increased diversity between users and the tool can improve system performance even when this increase comes at the expense of decreased tool performance.

Chapter 4: Effects of Tailoring Tool Setting on Users of Varying Abilities and Cases of Varying Difficulty in a Spell Checker Environment

Some authors in the automation bias literature claim there is a need for adaptive automation, and automation that uses knowledge of individual differences to improve tool design and tailor the way advice is presented. Some of these authors ask inspiring questions such as whether rear-end collision detection alarms should be customized according to driver capabilities/age (Parasuraman, 1997a) or whether the sensitivity of medication management tools should be different for senior users (Ho, 2005) or whether, in general, tool settings should be tailored to each user or a group of similarly skilled users (Alberdi, 2009). Such questions deserve further attention. Taking this lead and combining findings from Chapter 1 that show different effects of automated tools depending on user ability and case difficulty, we propose to improve system performance by tailoring tool setting according to these factors. We define tool *setting* as a unique sensitivity / specificity combination and we suggest that if we can vary the setting of a tool so that one setting has a lower specificity (higher FP rate) than the other in return for a higher sensitivity (lower FN rate), while the other setting exhibits the opposite trade-off, then: (1) each group of similarly skilled users will perform better using a different tool setting and (2) the same similarly skilled group of users may perform better using a different tool setting in difficult versus easy cases.

To test these hypotheses, we conducted a repeated measures experiment using a spell checking tool. According to our knowledge of the literature, this is the first experiment to explore how system performance can be enhanced by calibrating the sensitivity and specificity of a tool for a range of user skill groups and case difficulty. Also, according to our knowledge, only one other study has published results on automation bias using a spell checking tool (Galletta, 2005) and we hope that by introducing the experiment in such a non-critical domain we may encourage future studies in automation bias using minimal resources before being extended to more critical domains.

During the experiment, participants were exposed to two different tool settings: one more specific and one more sensitive. Under each tool setting, each participant was asked to edit two different passages of text: one easy and one difficult, as measured by the frequency of difficult words in each passage. We rated user performance based on the amount of errors left in each passage after editing, while distinguishing between false positive and false negative errors. The analysis, as suggested by Alberdi et al. (2010), stratified users according to ability and spelling passages according to difficulty. User ability was determined by a dictation test, and questionnaires in the beginning, middle and end of the experiment helped deduce subjective measures of self-confidence, perception of tool reliability, tolerance for tool prompts, and other similar variables.

We found evidence of our general hypotheses that different groups of similarly skilled users perform better at different tool settings, and that the same group of users perform better using a different tool setting in difficult versus easy tasks, depending on their positive reaction to tool TPs, their negative reaction to tool FPs and the balance between the two. Furthermore, in agreement with our simple modelling of user behaviour:

- **high verbals** (defined as participants who scored above the average score in the experiment) performed better using the **sensitive tool** when editing **easy passages**, but the **specific tool** when editing **difficult passages**
- **very low verbals** (participants who scored worse than one standard deviation below average in the experiment) performed better using the **specific tool** when editing **easy passages**.

Contrary to our expectations, **low verbals** (participants who scored below the average score in the experiment) performed better using the **sensitive tool** when editing **difficult passages**.

Secondary results showed significant differences between users' general self-reported trust in spell checkers and their self-reported trust of the specific spell checker used in the experiment, and this is one reason that may explain some of the discrepancies in the literature on the effects of trust on automation use. Secondary analyses also revealed that high verbals were significantly better than low verbals in correcting spelling mistakes, but not in distinguishing false tool prompts. Among many possibilities, this may be because tool false prompts existed on words they were initially uncertain about and so were easily swayed by the tool's prompts, or that the tool's successful highlighting of difficult mistakes resulted in a level of trust that caused high verbals to accept some of these false tool prompts.

Interestingly, the majority of high verbals preferred working with the sensitive tool while the majority of low verbals preferred working with the specific tool. This majority vote in each category matches our expectation that high verbals would likely prefer the tool setting that highlights difficult mistakes even if it meant an increased number of FPs, while low verbals would probably prefer the tool that makes fewer FPs, perhaps because the additional TPs the sensitive tool highlights are not all recognizable to them.

The experiment resulted in trends which generally match our expectations, but some of these results revealed non-significant differences when analysed using ANOVA. Besides the possibility that the conjectured effect is absent, we also discuss several other possibilities for this non-significance including sample size, choice of tool setting, and the definition of high and low verbals.

An important contribution of this experiment is that it provides a useful basis for future, similarly designed studies. Experience with the generation of passages, discussion of design issues, and results

of ability indicators can be used to feed future studies in automation bias at the Centre for Software Reliability. Moreover, numerical results of the experiment can be used to inform experimental parameters when designing future studies, especially when choosing appropriate tool settings. Future work may also benefit from ideas we present including to: (1) better explore the significance of the results by increasing sample size and altering tool settings, (2) enrich the analysis by taking into account tool markings and orthographic word difficulty, (3) explore other prediction techniques besides modelling such as pilot observations, (4) implement the experiment in a different context using tools with pre-defined tool settings, and (5) provide insight into the causes of certain human errors using post-study questionnaires, eye tracking, and/or ethnographic observations.

Although this is a summary of the overall chapter, more details can be found in the pages that follow. Section 4.1 provides a brief literature review of studies in the automation bias and education literature about spell checkers and their effects. Section 4.2 states the hypotheses and the reasoning and modelling behind our expectations. Following this is a documentation of the experimental design in terms of stimuli, setup, procedure, outputs, participants, and passages. Section 4.4 contains the experiment results. Section 4.5 contains possible tracks for future work. Finally, Section 4.6 describes the experiment's contribution to existing knowledge and an overall discussion of the results.

4.1 Literature on the Use and Effect of Spell Checkers

In this section, a brief overview of the literature examining the effect of spell-checkers is presented. The review is divided into two categories: studies from the automation bias literature and studies from the education literature.

4.1.1 The Effect of Spell Checkers in the Automation Bias Literature

Among the automation bias literature, Galletta et al. (2005) are the only authors to publish results using a spell checking tool. In their interesting experiment, authors asked users to read a business letter and correct spelling and grammatical errors. Participants were divided into "high" and "low verbals" according to their score in a standard university entrance exam. Half the users saw the text with the checker and half saw it without. The performance of each user was evaluated based on the number of errors remaining in the document after their editing; thus, lower scores represented higher performance. Before summarizing Galletta et al.'s results, we carefully define word types to help all future discussions in this chapter. Words are classified into four categories according to two factors: whether they are spelled correctly, and whether they are marked by the spell checker:

Table 4-1: Definition and Examples of Word Types

Word Type	Definition	Example
Marked Misspelling	True positive by the tool	Incorrectly spelled word marked by the tool Example: *hapiness*
Unmarked Correct Word	True negative by the tool	Correctly spelled word not marked by the tool Example: *happiness*
Unmarked Misspelling	False negative by the tool	Incorrectly spelled word not marked by the tool such as an incorrectly used homophone Example: His effort was in *vein*
Marked Correct Word	False positive by the tool	Correctly spelled word marked by the tool. Such errors could be the result of a limited tool dictionary (that is still under construction or that does not include certain field-specific terminology) Example: *Czechoslovakia* Such errors could also be due to problems in the algorithm that detects the incorrect use of homophones Example: *There* is no problem

Based on this terminology, Galletta et al.'s (2005) results can be summarized as follows:

- Marked misspelling:
    - When the checker was turned off, high verbals had a lower score (better) than low verbals.
    - When the checker was turned on, high and low verbals both benefitted and came to the same score, but this implies that low verbals benefitted more.
- Marked correct words:
    - When the checker was turned off, high and low verbals had a similar score
    - When the checker was turned on, both high and low verbals' scores increased (worse). But low verbals fell for more of these errors. In other words, they were less able to differentiate between real versus false tool prompts; although this difference between high and low verbals was small.
- Unmarked misspelling:
    - When the checker was turned off, low verbals had a higher score (worse).
    - When the checker was turned on, both high and lower verbals had the same score. Thus, high verbals were degraded, and low verbals were almost unaffected, and *the authors* attributed this effect to "laziness" or "overreliance" in high verbals.

Although these results were exhibited in a non-critical system, they are not limited to this context and we note from the literature review in Chapter 1 that similar results were obtained when 50 readers read 180 mammograms to identify patients needing to be recalled (Alberdi, 2010). Results showed that both high and low discriminating readers demonstrated an increase in false positive errors due to false positive prompts from the tool, and that this effect was again more pronounced for less discriminating readers (Alberdi, 2010). Also, unprompted cancers were less likely to be recalled by high discriminating readers when using CAD compared to when not using CAD (Alberdi, 2010).

A recent Master's thesis at City University also considered automation bias in a spell checker context (Gray, 2012). In the experiment, readers read first without a checker, then with an accurate checker (only true negatives and true positives by the tool), and finally with a checker that was not 100% reliable (included false positives and false negatives by the tool). Gray (2012) found that all readers improved in performance or at least maintained performance when using the accurate checker. On the other hand, the imperfect checker improved the performance of low verbals, but surprisingly worsened the performance of half the high verbals. Also interesting was that the checker benefitted low verbals more than high verbals and that variance in user performance was lowest when the accurate checker was used.

4.1.2 The Effect of Spell-Checkers in the Education Literature

Outside the automation bias literature, in the education literature, there are also studies looking at the effect of spelling and grammar checkers on the performance and learning of users. Although the aim of these studies is different from our focus, the methodology, experimental design, hypotheses, and sometimes the results are useful to understand. In the following, we describe some of these studies most relevant to our needs.

In one study, students revising an essay using a checker were compared to those revising using a dictionary (Figueredo, 2006). The authors used two essays, one and a half pages each, and participants had fifteen minutes to read each essay. The essays were written by first year students in an English course and errors in low-frequency words were inserted into the text. The authors also created two versions of each essay so that each version had the same number of errors but a different set to avoid confounding factors. Participants did not only make corrections, but also provided oral descriptions of their revisions. This method of asking users about their performance, decisions and/or thoughts has been documented elsewhere and can take the form of: think-aloud records, questionnaires, or interviews (Fitzgerald, 1987).

Results showed that students corrected more mistakes in the checker condition than in the dictionary condition. Spelling ability, measured at the start of the task with the Diagnostic Spelling Potential Test, was correlated to the correction of spelling errors, but only in the dictionary condition. English major and graduate students corrected more spelling errors than freshmen. Authors attributed this difference to a possible lack of vigilance by freshmen, because there were no spelling ability differences between the groups.

In another study, Figueredo et al. (2004) examined differences in spelling error detection and correction as a function of type of spelling error. Errors were limited to incorrect suffixes and classified into three categories: phonological (e.g., *incredibul*), orthographic (e.g., *decisian*), and morphological (e.g., *extention*). Again authors inserted commonly misspelled words into student essays. Participants read an unmarked copy of the essay and then a copy where all errors were marked. Results showed that participants detected significantly more phonological errors, followed by orthographic and finally morphological errors. Detection of morphological errors also increased with spelling skill while FPs were inversely related to skill. Importantly, this showed that detection depended not only on spelling ability but also on error type.

Most users could correct most of the errors they detected, but a third of the errors that average spellers corrected in the marked copy, they did not detect in the unmarked copy (Figueredo, 2004).

This indicates that knowledge of the spelling of a word is not necessarily sufficient for the detection and correction of the error, and is consistent with previous work that found that users failed to correct 48% of errors that they had the knowledge of how to correct (cited by Figueredo, 2004). These results support the processing deficit hypothesis for spelling errors, which theorizes that users fail to correct an error because they fail to use their knowledge to *detect* the error, which is different from the knowledge deficit hypothesis which theorizes that users fail to correct an error because they do not have *knowledge* of its correct spelling (Figueredo, 2004). However, authors also highlight an important limitation in this interpretation of the results. It may be that users successfully correct an error in the marked copy, because when it is marked and there is only one reasonable unit of change, then the user can easily guess the correct spelling of the word – even without knowledge of its correct spelling. As future work, authors point to the importance of incorporating context and familiarity of text and how the two factors affect detection across error types.

The processing deficit hypothesis leads to an interesting question: why do people sometimes fail to detect spelling errors when editing, despite knowledge of the correct spellings? Evans (2013) lists several possible reasons in his interesting book titled, "Don't Trust Your Spell Check":

- Predictive coding in the brain which allows users to understand sent**nces with miss**g let***s or sentences with wrods whose letters hvae been mangled. In other words, humans don't read individual letters, but words as a whole. Thus, errors in a word's middle letters are often difficult to detect, especially compared to errors a word's exterior letters.
- When users are too familiar with the writing, their brain shows them not what they see, but what they expect
- When users become emotionally connected to the meanings and implications of the writing, this becomes a barrier in detecting spelling errors
- When people rely on spell checkers
- When people are distracted or uncomfortable (temperature, lighting, etc.) while they edit
- When people are pressed for time and thus forced to skim the text
- When errors are outside the main text (headings or captions) or in repeated text they can often be overlooked.

This list, which shows interesting overlaps with similar lists of faults in human reliability assessments, can be useful when designing a spell checking experiment as manipulating some of these variables can help increase people's error rate so as to amplify effects researchers would like to observe. For example, including more errors with middle letters than exterior letters, adding errors outside the main text, pressing users for time, manipulating the familiarity of the text, etc.

4.1.3 Orthographic Representation of Words as a Measure of Word Difficulty

Some words in the English language have a spelling which closely corresponds to the phonetic realization of the word, and others, especially archaic words or words with foreign borrowings do not exhibit such a correspondence. Between these two extremes are a group of words which contain an ambiguous segment involving some departure from straightforward phonetic mapping that may require understanding of morphophonemic structure (mis+spell requires retaining both s's) or application of established orthographic conventions (strap + ed requires doubling the p). This three-level division of words goes beyond the coarser division of regular and irregular spellings because such a two-level division does not take into account spelling pattern frequency or the application of a common rule or pattern. Table 4-2 gives a detailed definition of this three-level classification, suggested by Fischer (1985), along with examples.

Table 4-2: Word Types by Orthographic Representation

Word Type	Definition	Examples
Level 1	Orthographic representation closely follows the phonetic realization. These spelling patterns have a high frequency in the English language	Harp Adverb Retort
Level 2	Words which contain an ambiguous segment which requires understanding of morphophonemic structure or application of orthographic conventions	Thinned Clearance Inexhaustible
Level 3	Words which contain patterns which do not generally occur in English, or occur with a low frequency. These are mostly archaic or foreign words.	Fahrenheit Bourgeois Sergeant

In an experiment, Fischer (1985) showed that participants who differ in spelling ability differ in linguistic sensitivity, "the ability to apprehend the inherent regularities at various levels of linguistic representation". Participants were dictated 40 words from each of the three divisions in Table 4-2. Scores from this dictation were highly correlated to participants' scores in the Wide Range Achievement Test, a widely used conventional test of spelling ability; $r=0.84$, $p<0.001$. Furthermore, "good and poor spellers differed significantly in their ability to spell words at each of the three levels". Finally, the greatest difference between the two ability groups was in their performance on Level 2 words suggesting that linguistic abilities rather than rote memory were more significant in differentiating good and poor spellers.

Fischer published, in the appendix of his paper, 40 words of each level, balanced for syllable length (mean 2.8 syllables) and frequency of occurrence in written English (mean 6.1 occurrences

per 1014232 words). These words are available in Appendix 4A for reference as they play an important role in the design of our experiment.

4.2 Proposed Hypotheses to be Studied in the Spell Checker Context

As noted in our literature review in Chapter 1, future research in human factors and ergonomics aims to use knowledge of user differences to improve design and training efforts and similar research in the area of automation bias would appear worthwhile (Parasuraman, 2010; Oron-Gilad, 2005). Halligan et al. (2011) also mentioned a need for further research aimed at predicting the effect of tool assistance on individual users (Halligan, 2011).

In this section we explore one way that these differences can be incorporated into the tool's design. Crucial parameters in any alerting tool are the tool sensitivity and specificity, and the trade-off between the two poses a serious question to designers: which sensitivity/specificity combination will yield the best system (tool + human) performance? We will refer to this sensitivity/specificity combination as the tool *setting*. We argue that customizing the tool setting to specific users and cases could help improve overall system performance.

4.2.1 Tailoring a Tool's Setting According to User Experience

Parasuraman et al. asked an interesting question: should rear-end collision detection alarms be customized according to the driver's age group? (Parasuraman, 1997b). Ho et al. (2005) took this further by stating that because when using a medication management tool older users responded more often to false alarms than younger users, the tool's sensitivity should be different for the two user groups. Finally, Alberdi et al. (2010) suggested that for some applications, it may be possible to tune the sensitivity and specificity of an alerting tool for each individual user or to calibrate it for a range of different levels of user skill.

Combing these suggestions and the results in Section 1.4.1.3 that showed how users of varying experience react differently to tools' prompts, then perhaps the tool threshold a user from a certain ability group operates with should be adjusted in a manner to reduce the automation-related errors of that respective ability group. This may sometimes entail a higher rate of false prompts in return for a lower rate of false negatives or vice versa. As Sorkin et al. realized early on in 1985, optimizing an automated tool's performance may not always optimize the overall system's performance. There may be times when a lower tool sensitivity in return for a higher tool specificity may actually be beneficial for a certain class of users.

However, is a tool of varying settings a feasible design? According to personal correspondence with Dr. Greg Slabaugh, a professor at City University London who develops such tools, such an option is simple to implement. Interestingly, this is an option already incorporated in Hologic's latest CAD tool for the detection of breast abnormalities (The Smart Choice 2). Three separate operating points (sensitivity/false marker rate settings) can be selected for both microcalcifications and masses; nine combinations in all so that the software can be adapted to the radiologists' reading preferences at a specific site (The Smart Choice 2).

4.2.2 Tailoring a Tool's Setting According to Case Difficulty

The idea of varying tool settings can be extended further to be tailored not only for user ability but also for case difficulty. Relying again on an example in the traffic context, we note that the probability of a rear-end collision increases for heavy trucks compared to light trucks, for wet roads compared to dry ones, in dark hours compared to daylight, and for higher traffic density compared to lower traffic density roads (Kim, 2001). Based on such observations in the traffic context, the required minimum warning distance required by a tool before alarming, is adjusted, at least for vehicle types (Parasuraman, 1997b). De Vries et al. (2003) also interestingly noted that inaccurate traffic information was more harmful in the familiar (versus the new) setting and thus depended on the scenario and not just the user. Povyakalo et al. (2013) also noted that CAD's effectiveness was greatly affected by variations in case difficulty.

Identifying the difficult cases may not always be easy or feasible. However, one idea is based on our recent findings which showed that radiologists' statements of uncertainty in their pre-advice decision when reading mammograms were strong predictors of errors and changes in their post-advice decisions (Gadala, 2014). Thus, one way that a tool's threshold could be tailored according to case difficulty is by using uncertainty in a user's pre-advice decision to detect difficult cases and tailor the tool's setting accordingly. Such ideas lead us to believe that a tool's setting should be tailored not only to a user's ability but also to case difficulty

4.2.3 Statement of General Experiment Hypotheses

We define two general tool settings: one which has a lower specificity (higher FP rate) than the other in return for a higher sensitivity (lower FN rate), while the other setting exhibits the opposite trade-off. We also define two user ability groups: the highly skilled users and the less skilled users. "Better" performance refers to fewer total system errors, including false positives and false negatives. Then, using the ideas presented in Section 4.2.1 and 4.2.2, we suggest that:

- Highly skilled users will perform best using one tool setting, while less skilled users will perform best using the other tool setting.
- For a single user ability group, the tool setting that results in better system performance for difficult cases may be different from the tool setting that results in better system performance for easy cases.

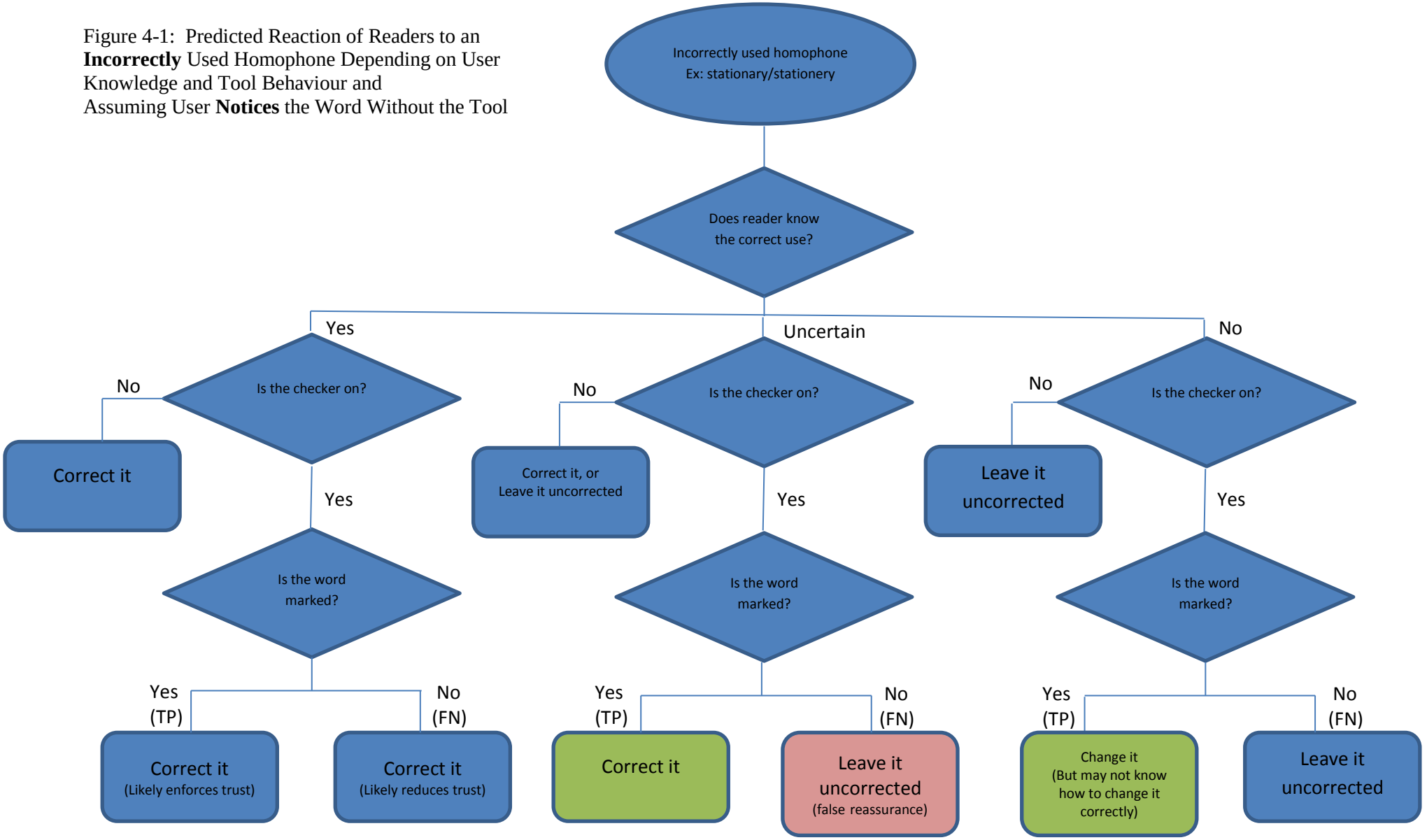
These general hypotheses state that different groups of similarly skilled users will perform better using different tool settings and depending on the difficulty of the task. However, they do not predict *which* tool setting each group of users will perform better using, and for which case difficulty. We use the modelling in Section 4.2.4 to aid us in making such predictions, and then restate the hypotheses in more specific terms in Section 4.2.5.

4.2.4 Interactions Between Tool Settings and User Knowledge

The question we are addressing is: which tool setting will better improve which users and for which task difficulty? For this, we need to consider two factors: (1) varying knowledge and vigilance of the *users*, and (2) different cases describing the *tool checker's* behaviour.

We consider, firstly, a scenario where the word "stationary/stationery" is used **incorrectly** in a passage with the checker turned off. Then we consider a scenario where it is used incorrectly in a passage with the checker turned on. In the latter case we assume two separate cases: the checker fails to mark the error, and another case where the checker correctly marks the error. Thus, we are considering three scenarios. However, in each of these scenarios, the user will react differently according to their own knowledge (and vigilance). To simplify our model, we consider three main possibilities: the user knows the correct spelling of the word, is uncertain about its spelling, or does not know the word at all. For each of these three knowledge scenarios, and assuming the user notices and analyses the word, the action that the user will probably take when the checker is turned off, when the checker fails to mark the error, and when the checker marks the error is outlined in the appropriate boxes in Figure 4-1 on the next page. An important assumption in these predictions is that the tool does not provide suggestions, only markings.

Figure 4-1: Predicted Reaction of Readers to an **Incorrectly** Used Homophone Depending on User Knowledge and Tool Behaviour and Assuming User **Notices** the Word Without the Tool

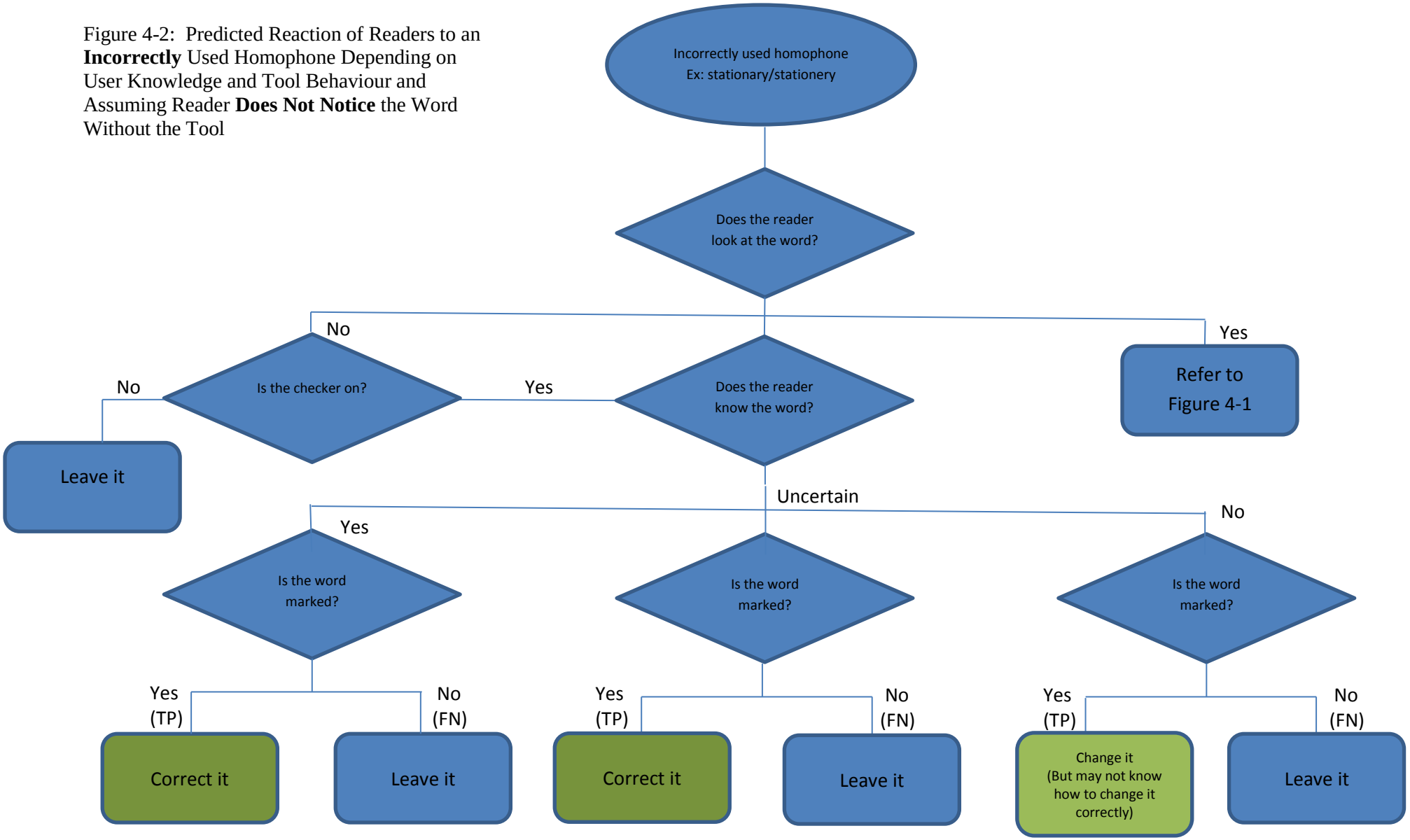


Thus, when we compare the scenarios when the checker is off and when it is on, we see that a high tool FN rate will have a negative effect on readers who are initially uncertain about a word's spelling (depicted by the red box). Also, according to our predictions, true positives from the aid will be helpful to readers who are initially uncertain about a word and may also be helpful to readers who do not know a word (depicted by the green boxes).

Thus, if tool FN errors occur on words that low verbals do not recognize and high verbals are unsure about, then we will get the effect that Galletta et al. (2005) found, which is that low verbals were unaffected by FNs and high verbals were negatively affected. This is different than Galletta et al.'s suggestion that high verbals were degraded by FNs because of lack of vigilance.

One of the main aims of spell checkers is to highlight misspellings that the reader fails to detect. To incorporate this scenario, we make a similar diagram that depicts the reaction of a reader to the same incorrectly used word but this time assuming that the reader did not notice the word when the checker was off due to reasons associated with vigilance/attention. The following figure shows how we expect the reader's behaviour to change when the tool is then turned on, and where we assume tool prompts are always seen and responded to.

Figure 4-2: Predicted Reaction of Readers to an **Incorrectly** Used Homophone Depending on User Knowledge and Tool Behaviour and Assuming Reader **Does Not Notice** the Word Without the Tool

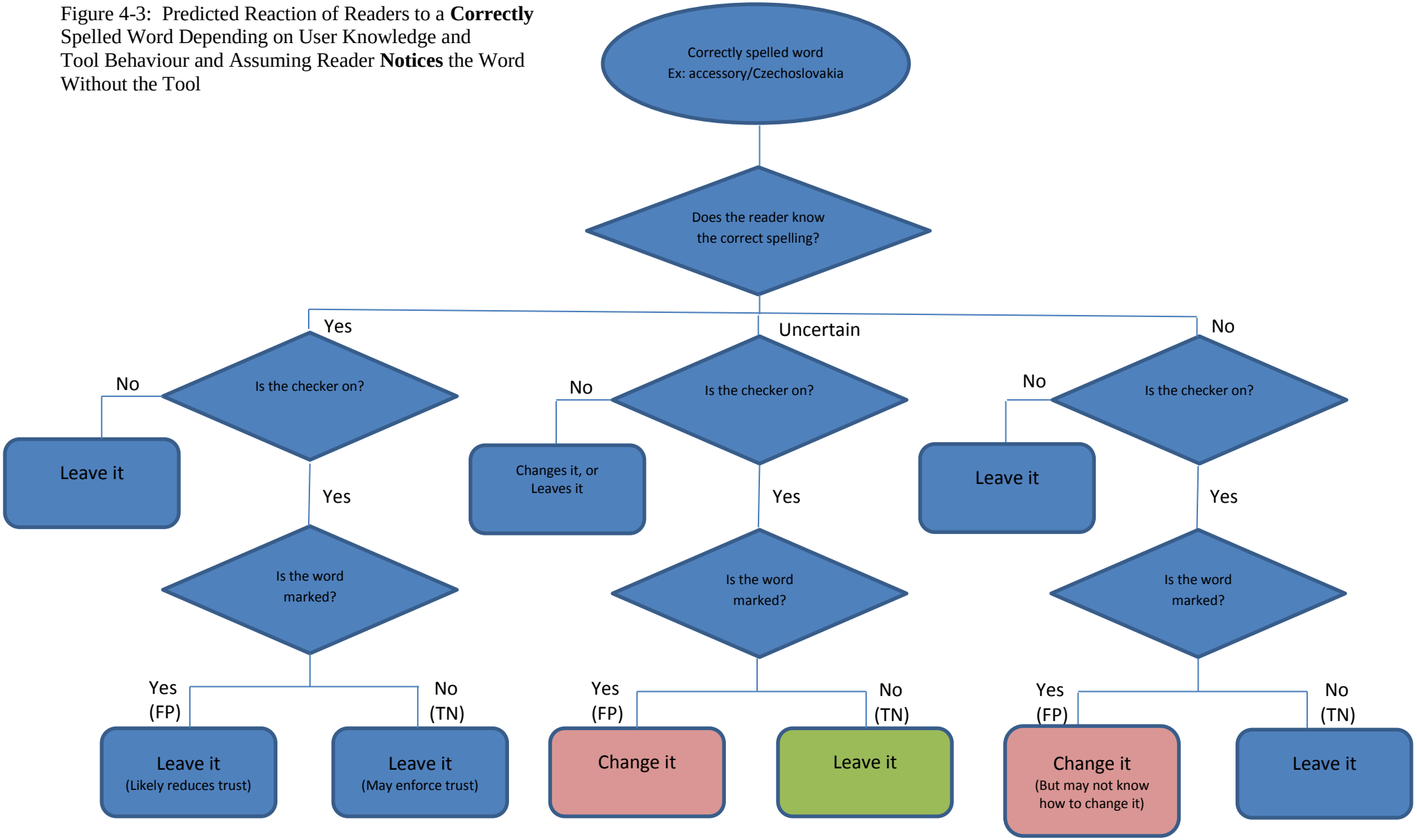


By assuming the reader had not noticed or not analysed the word with sufficient attention when the checker was turned off, and as depicted in Figure 4-2, we see that TPs from the tool are beneficial to all readers in this case. Indeed, this is the rationale for the implementation of most decision-support tools such as, for example, the use of CAD in cancer screening.

Note in the figure that we have introduced two shades of green boxes. This is to depict that we expect a higher benefit from the tool in the dark green coloured scenario than the benefit in the light green coloured scenario. Let's consider an example to clarify. When a reader initially does not notice a word, then there is approximately a 0% chance that they will correct the word without the tool. Turning on the tool to give a TP, raises this probability close to 100% if the user knows the word. Thus, there is a net benefit of about 100% - hence the dark green box. However, if the user does not know a word, then even when the tool is turned on and the word is marked, they may still choose not to correct it because they do not know how to, or they may guess the correction. In this case we may assume that there is approximately a 50% chance of correcting the word and thus the net benefit is approximately 50% - hence the light green box. These numbers do not represent exact probabilities of change, but approximate vicinities of our expectations and help us materialize the relative benefits of tool prompts in each scenario.

Figures 4-1 and 4-2 covered scenarios where a word was used incorrectly. Now, if we consider a difficult word such as "Czechoslovakia" or a commonly confused word such as "accessory" (where users may confuse which letters are doubled) and assume it is spelled *correctly* in a passage, there are several possibilities of tool behaviour. The checker may be turned off or on, and when turned on, the checker may incorrectly mark, or correctly leave the word unmarked. For each of these scenarios, a reader can go through one of the alternative scenarios depicted in Figure 4-3.

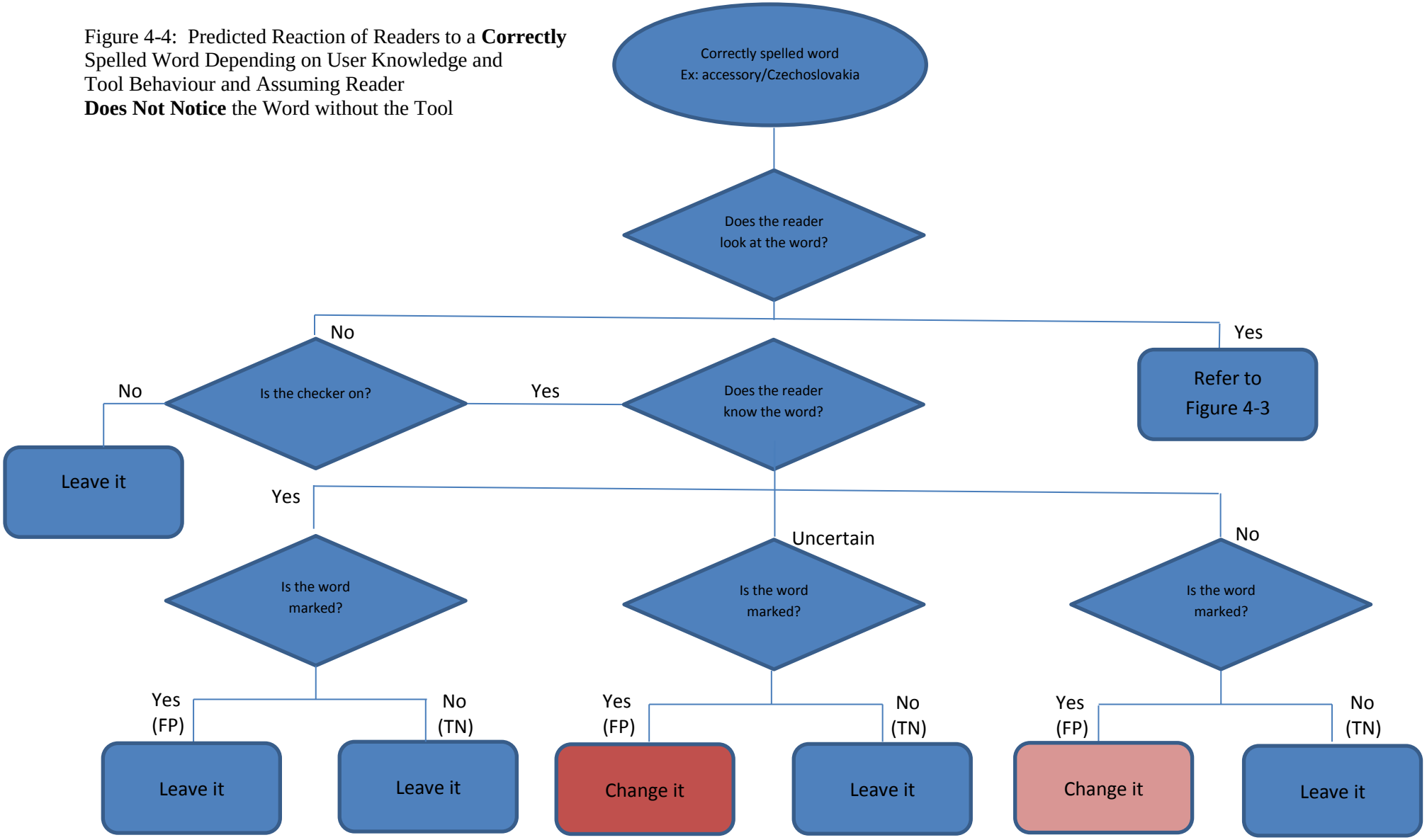
Figure 4-3: Predicted Reaction of Readers to a **Correctly** Spelled Word Depending on User Knowledge and Tool Behaviour and Assuming Reader **Notices** the Word Without the Tool



Thus, based on Figure 4-3, readers are most likely to change a correctly spelled word that is flagged (tool false positive) if they do not know it or if they are initially uncertain about it. A reader who is uncertain about a word's spelling and which the tool marks is likely to attempt to change the word. However, without the tool there may be only around a 50% chance that they change it, which rises to close to 100% with the tool on – thus resulting in approximately a 50% chance of degradation. A reader who does not know a word will likely not change it when the tool is off (0% chance of change), but may attempt to guess its spelling when it is marked by the tool (50% chance of change) – again representing about a 50% chance of degradation in performance. Hence both these boxes are shaded in light red unlike some of the dark red boxes in the next figure. Again, the numbers do not represent exact probabilities of change, but approximate vicinities of our expectations and help us materialize the relative effects of tool prompts in each scenario.

Again, the previous diagram assumes the reader notices or analyses the word with sufficient attention, but does not take into account the important scenario where the reader fails to notice the word when the checker is turned off due to reasons of vigilance/attention. Figure 4-4 looks at how such a reader's behaviour is likely to change when the tool is turned on.

Figure 4-4: Predicted Reaction of Readers to a **Correctly** Spelled Word Depending on User Knowledge and Tool Behaviour and Assuming Reader **Does Not Notice** the Word without the Tool

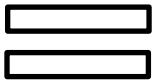
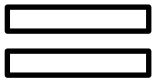
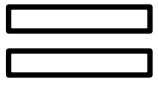
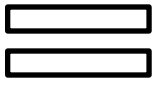
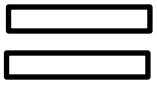
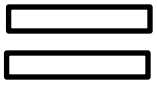
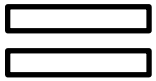
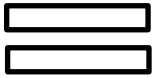
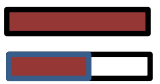


Combining the results from the predicted reactions of readers who notice a word and those who fail to notice a correctly spelled word shows that a reader is most likely to change a correctly spelled word that is marked (tool false positive) when they are initially uncertain about the word or if they do not know it, and that true negatives by the tool are most beneficial to those who have doubts about the word to begin with. Similar to before, we introduce two shades of colours to account for a difference in the predicted effect in performance.

We can use the predictions outlined in the previous four diagrams to predict the better tool setting for each group of users. First we note that for each reader, there will be words they: (1) know, (2) don't know, or (3) are uncertain about. We assume that: (1) there is somewhat uniform knowledge within each of the two skill groups and (2) for any given word, low verbals always have less than or equal knowledge compared to high verbals. "Knowledge" is used here as a simplistic descriptor of users' abilities, specific to each word.

To see a difference in behaviour between the two skill groups at different tool settings, we need to focus on words which the two groups have different knowledge about. These words are binned into three types: A, B, and C as defined in Table 4-3. The effect of each tool marking is visually depicted either as no effect (empty rectangle), strong positive effect (green coloured rectangle), weak positive effect (half- filled green coloured rectangle), strong negative effect (red coloured rectangle), or weak negative effect (half-filled red coloured rectangle) - corresponding to the colours described in the previous four figures. For each error type, the effect is divided into two columns, the first shows the predicted effect when the reader notices the word when the checker is off and the second shows the effect when the reader fails to notice the word when the checker is off. In effect, this table is just a single visual summary of the effects previously presented in Figures 4-3 to 4-6.

Table 4-3: Summary of the Effects of Tool Markings on Users of Varying Abilities and Words of Varying Difficulty Forecasted by the Models in Figures 4-1 to 4-4

Word	Effect of Tool TP		Effect of Tool FN		Effect of Tool TN		Effect of Tool FP	
	Reader Notices Word When Checker is Off	Reader Doesn't Notice Word When Checker is Off	Reader Notices Word When Checker is Off	Reader Doesn't Notice Word When Checker is Off	Reader Notices Word When Checker is Off	Reader Doesn't Notice Word When Checker is Off	Reader Notices Word When Checker is Off	Reader Doesn't Notice Word When Checker is Off
A. High verbals know Low verbals are uncertain								
B. High verbals know Low verbals don't know								
C. High verbals are uncertain Low verbals don't know								

Difficulty ↓

These basic trends help us predict general tendencies of user behaviour according to word type, user knowledge and tool markings. Furthermore, they provide one possible explanation why Galletta et al. (2005) found that low verbals were unaffected by FNs and high verbals were negatively affected. Galletta et al. (2005) attributed this difference to lack of vigilance by high verbals, but Table 4-3 shows that if tool FN errors were based on words that low verbals did not know and that high verbals were unsure about, then we would get the effect observed by Galletta et al. (2005) – without a lack of vigilance by high verbals.

Furthermore, just as results obtained in a spell checking context by Galletta et al. (2005) complemented results obtained in the more critical domain of cancer screening by Alberdi et al. (2010), we expect that the modelling presented in this section could also be extended to applications outside the spell checking context and specifically to decision support tools that provide users with alerts or prompts. For example, this modelling could be used to reason why Alberdi et al. (2010) noticed a higher degradation in performance due to tool false positives in low discriminating readers compared to high discriminating readers (possibly because these tool false prompts were on cases that high discriminating readers were certain about), or why unprompted cancers were less likely to be recalled by high discriminating readers when using CAD compared to when not using CAD (possibly because these unprompted cancers represented cases which high discriminating readers were initially uncertain about but low discriminating readers did not identify or notice even without the tool).

4.2.5 Experiment Hypotheses Restated

Both Alberdi et al. (2010) and Galletta et al. (2005) found that:

- High and low discriminating users showed an increase in false positive errors due to false positive prompts from the tool, but this effect was more pronounced for less discriminating users.
- False negatives by the tool affected high discriminating users by making them more likely to miss targets compared to when they were not using the tool, while they had no/a smaller effect on low discriminating users. Alberdi et al. (2010) were able to identify that this was true for *difficult* false negatives.

Using these observations from previous publications, and assuming enough "difficult" errors exist in a passage, then groups of similarly skilled users should perform better, where better is defined as fewer total human errors (sum of human false negatives and human false positives) using different tool settings if for the more sensitive tool setting:

- High verbals were more positively affected by the tool's decreased FN rate than negatively affected by the tool's increased FP rate,
- But low verbals were more negatively affected by the tool's increased FP rate than positively affected by the tool's decreased FN rate.

This can be represented mathematically in Tables 4-4 and 4-5, using variable notation. In Table 4-4, we define x as the number of human FPs committed by low verbals when using the specific tool, and y as the number of human FNs committed by low verbals when using the sensitive tool. We also use σ and Δ (where we define $\boldsymbol{\sigma<\Delta}$) to respectively represent incremental changes in the number of human FNs and FPs committed by low verbals when they change the tool setting they use.

Table 4-4: Performance of Low Verbals

Type of Errors	Tool Setting	
	Specific Tool (higher FN rate, lower FP rate)	**Sensitive Tool (lower FN rate, higher FP rate)**
Correctly spelled words that reader will change (Human FPs)	x	$x+\Delta$, $\Delta>0$
Incorrectly spelled words that reader will leave (Human FNs)	$y+\sigma$, $\sigma<\Delta$	y
Total Human Errors	$\boldsymbol{x+y+\sigma, \sigma<\Delta}$	$\boldsymbol{x+y+\Delta}$

Thus, Table 4-4 presents our prediction that low verbals will make more human FPs $(x+\Delta)$ using the sensitive tool compared to when using the specific tool (x). However, we also expect that low verbals will commit more human FNs using the specific tool ($y+\sigma$) compared to when using the sensitive tool (y). If, as we expect, $\sigma<\Delta$, then low verbals will thus have fewer human errors in total when using the specific tool ($x+y+\sigma$) compared to when using the sensitive tool ($x+y+\Delta$).

Similarly, in Table 4-5, we define c as the number of human FPs committed by high verbals when using the specific tool and d as the number of human FNs committed by high verbals when using the sensitive tool. We also use η and γ (where we define $\eta<\gamma$) to respectively represent incremental changes in the number of human FPs and FNs committed by high verbals when they change the tool setting they use.

Table 4-5: Performance of High Verbals

Type of Errors	**Tool Setting**	
	Specific Tool (higher FN rate, lower FP rate)	**Sensitive Tool (lower FN rate, higher FP rate)**
Correctly spelled words that reader will change (Human FPs)	c	$c+\eta,\ \eta<\gamma$
Incorrectly spelled words that reader will leave (Human FNs)	$d+\gamma,\ \gamma>0$	D
Total Human Errors	$c+d+\gamma$	$c+d+\eta,\ \eta<\gamma$

Thus, Table 4-5 presents our prediction that high verbals will make more human FPs $(c+\eta)$ using the sensitive tool compared to when using the specific tool (c). However, we also expect that high verbals will commit more human FNs when using the specific tool ($d+\gamma$) compared to when using the sensitive tool (d). If, as we expect, $\eta<\gamma$, then high verbals will thus have fewer human errors in total when using the sensitive tool $(c+d+\eta)$ compared to when using the specific tool ($c+d+\gamma$).

These hypotheses rely on two key behaviours: (1) high verbals will not fall for some of the tool FPs, and (2) low verbals will not be negatively affected by the specific tool's increased FN rate. From the modelling in the previous section, we see from Table 4-3 that these two behaviours are more likely true for: words of Type A/B, and words of Type C, respectively.

However, when editing difficult text, where we assume there is a high occurrence of difficult tool FPs, we expect that high verbals will show an opposite trend. Difficult text will contain more words that now even high verbals may not know or will be uncertain about. In this case, their behaviour will resemble that of low verbals, meaning there will be more FPs from the tool that

high verbals will accept and fewer TPs from the tool that will be helpful. Thus, we predict that when editing difficult text, high verbals will commit fewer human errors in total when using the specific tool compared to when using the sensitive tool.

Now, we can restate the general hypotheses introduced at the start of Section 4.2, but with specific predictions of which tool settings will be better for each user ability group and task difficulty:

1. When editing easy or difficult text, most **low verbals** will commit **fewer errors** (total sum of human FNs and human FPs) using the more **specific tool** setting compared to the more sensitive tool setting.
2. When editing easy text, most **high verbals** will commit **fewer errors** (total sum of human FNs and human FPs**)** using the more **sensitive tool** setting compared to the more specific tool setting.
3. When editing **difficult** text**,** most **high verbals** will commit **fewer errors** (total sum of human FNs and human FPs) using the **specific tool** compared to the more sensitive tool setting.

4.2.6 Example Scenario

These hypotheses may become clearer with an example. Suppose we define two different spell checker tool settings, one which is more sensitive but less specific and the second which shows the opposite trade off as described below:

Specific tool
Weaker algorithm in detecting homophone errors
Larger dictionary

Sensitive tool
Stronger algorithm in detecting homophone errors
Smaller dictionary

The specific tool has a larger dictionary which means it will make fewer tool FPs compared to the sensitive tool. However, because its homophone detection algorithm is weaker, then compared to the sensitive tool, the specific tool will commit more tool errors (FPs and FNs) in homophone errors.

On the other hand, the sensitive tool has a smaller dictionary which means that more correctly spelled words will be wrongly marked by the tool; thus resulting in more tool FPs compared to the specific tool. However, because its homophone detection algorithm is better, then compared to the specific tool, the sensitive tool will commit fewer tool errors (FPs and FNs) in homophones.

Now if a high verbal were to use the sensitive tool, they would experience more tool FPs than when using the specific tool, which would likely result in more human FPs. However, we predict that they wouldn't fall for all of these FPs because as high verbals some of these words would be words of Type A and B, which they are certain of their spellings (as defined in Table 4-3). Although they would likely commit more human FPs, the sensitive tool also highlights more of the difficult homophone errors compared to the specific tool. This helps the high verbals fix these errors and results in fewer FNs compared to the specific tool. So long as this benefit in increased TPs is greater than the degradation due to increased FPs, then high verbals will perform better using the sensitive tool. This trade-off depends not only on our prediction of user behaviour but also on the prevalence of word difficulty and rate and type of tool errors.

Now if a low verbal were to use the sensitive tool, then like the high verbals they would attempt to change some of the correctly spelled words marked by the tool, and because they are low verbals, they would commit more FPs than high verbals because they do not have the same knowledge as high verbals to allow them to distinguish true versus false tool prompts. Then, even when the sensitive tool marks some difficult homophone errors, the low verbals are unable to correct them because they don't know how to change them or why they are wrong (ex: stationary/stationery or complementary/complimentary). Thus, optimization of the homophone algorithm does not significantly help these low verbals and they commit a similar number of FNs as when using the specific tool, but with an increased human FP rate. Thus, their overall performance will be better when using the specific tool.

4.2.7 Assumptions in the Model

As is true in any model, the model we presented makes a number of simplifications to the real world problem, but describes at a microscopic (single word) level, the forms of diversity between various users and the tool. Some of these simplifications are related to the predicted reactions of the readers in different scenarios and others are related to the predicted behaviour of the high and low verbal groups using the sensitive and specific tools. For reference, both types of assumptions are listed here.

Assumptions made when predicting the behaviour of readers in different scenarios:

- Readers that *know* a certain word will not be swayed to change it if the tool marks it or swayed not to change it if the tool fails to mark it. We predict this swaying in opinion is only true for readers who are initially *uncertain* or *do not know* a certain word.
- False prompts from the tool will not distract the reader from the tool's true prompts nor devalue them.

Assumptions made when predicting the behaviour of high and low verbals when using different tool settings:

- There is somewhat uniform knowledge within each of the two skill groups
- For any given word, low verbals always have less than or equal knowledge compared to high verbals

4.3 Experimental Design

We test the effects of tailoring the tool setting on users of varying ability and case difficulty using a spell checking tool. This tool is called the Automated Automation Bias Spell Testing Tool (AABST) and was previously developed by undergraduate students at City University London and later debugged and improved by Cristina Gray (2012) as part of her Masters project. Figure 4-5 is a screenshot from the tool showing a passage that needs to be edited and which contains words marked (in red) by the tool. It also shows the text field (in the right-hand corner) where words can be edited after being clicked by the user.

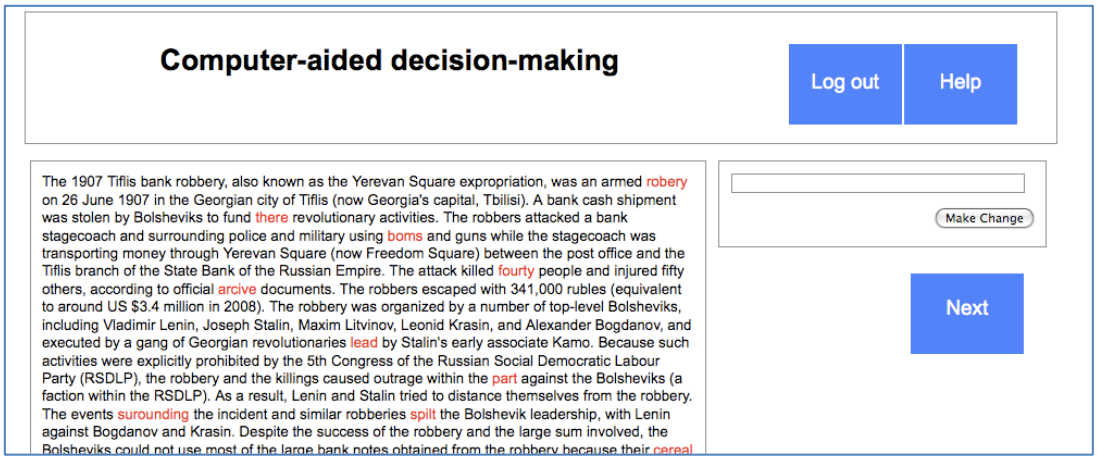


Figure 4-5: Screenshot from the Automated Automation Bias Spell Testing Tool (Gray, 2012)

In the following sections, we describe the experiment in more detail starting with the experimental: stimuli (Section 4.3.1), setup (Section 4.3.2), procedure (Section 4.3.3), outputs (Section 4.3.4), participants (Section 4.3.5) and passages (4.3.6). Unlike Chapter 2, the discussion of design issues are not presented in a separate section, but in their respective sections. This experiment received final ethical approval from the Ethics Committee of the School of Mathematics, Computer Science and Engineering in July 2015.

4.3.1 Experimental Stimuli

In the experiment, we are interested in manipulating the following variables:

- Tool setting of the spell checker
- Text difficulty

Due to the random recruiting of participants, we also expect the following variables to naturally exhibit variations, and we decided only to assess but not manipulate these variables:

- Spelling ability
- Self-confidence
- Trust in the tool

4.3.2 Repeated Measures versus Matched Pairs Setup

Two possible set-ups exist for the experiment: repeated measures or matched pairs. Their advantages and disadvantages are compared in Table 4-6.

Table 4-6: Comparison Between Repeated Measures and Matched Pairs Setup

Setup	Advantages	Disadvantages
Repeated Measures All participants edit 4 passages: 2 easy and 2 difficult at 2 different tool settings each	Requires fewer participants Could reveal differences in the case that readers read better at different tool settings according to factors other than their ability (e.g. certain characteristics or perceptions we can deduct from the questionnaire) Does not require a double setting commitment	Requires similar passages of "equal" difficulty Requires more work from each participant
Matched Pairs Half the high verbals edit 2 passages – one easy and one hard – at a given tool setting, and the other half edit the same passages at the second tool setting	Passage is unified	Requires more participants Assumes participants are "equal" or close enough to being so If a pre-task is used to split the participants, there may be drop-outs or less volunteers due to this extra commitment

Based on the comparison above, a repeated measures design was preferred, especially because experience of previous researchers suggested that the recruitment of participants can sometimes be challenging, especially in a double setting commitment. However, it was also preferred, because we wanted to test our hypotheses independent of any individual differences. Most importantly, it was preferred to account for the possibility that user ability might be better defined according to a different criteria than our division based on the assessment task (in fact, this turned out to be the case as will be discussed in the Results Section).

4.3.3 Experimental Procedure

Potential participants were recruited via invitation emails, leaflets, class announcements, and posters. A sample of the poster used is included in Appendix 4B. These methods of advertisement included a link to a "recruitment survey" which contained the Participant Information Sheet (Appendix 4C) and the Consent Form (Appendix 4D) followed by a few demographic questions. With the contact information participants provided in the recruitment phase, participants were sent location details of the on-campus computer lab where the experiment would take place. When they arrived, they were briefed before seeing the first two passages under the first tool setting. Then they completed a short questionnaire followed by two more passages under the second tool setting. Again these passages were followed by a short questionnaire. The final step was an assessment task before a short de-briefing. An overall view of the system from the user's perspective is presented in the user journey in Figure 4-6, where only normal flow through the system is shown.

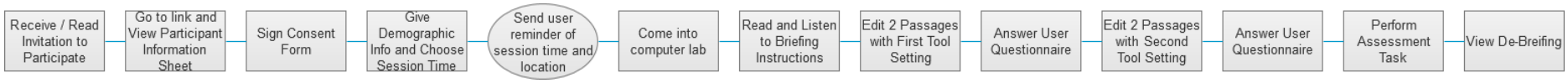


Figure 4-6: User Journey Depicting Flow Through the Spell Checker Experiment

In the following, we present some details describing each of these individual steps through the system along with some of our reasoning in case of design issues that aroused and caused discussions and evaluations of respective costs and benefits.

Recruitment:

The recruitment survey was hosted on Qualtrics and was used to collect demographic information about potential participants. Below is a list of the questions included in this stage:

- What is your gender?
- Are you a student?
  - I am enrolled in the _____ School.
  - I am a/an _____ student.
    - undergraduate/postgraduate
  - In my current degree, I am in my _____ year of study.
- How old are you?
- My first language is _____.
- I have been speaking English for the past _____ years.
- I consider myself a/an _____ speller
    - poor/average/good/excellent
- In general, I trust spell checkers
    - 5-point Likert scale (Strongly Agree to Strongly Disagree)
- Not including texts for my studies, I read _____ books in a year (including online books)
    - None/One or two/3-10/10-40/More than 40
- Regarding newspapers, I _____.
    - Rarely read one/Read one a couple times a month/Read one every day/Read more than one every day

These factors were carefully chosen based on previous publications because they were seen either to affect people's ability to detect and correct spelling errors, or affect the likelihood that people would be affected by automation bias, or both. For example, Fitzgerald et al. (1987) previously found an association between reading habits and spelling ability.

At this stage, participants were also provided with a list of different timings for when the experiment would take place in one of the on-campus computer labs. In the recruitment survey, participants were able to choose a time that suits their schedule, or choose the option to be contacted via email for an alternative time.

Although this experiment was based on an online application that could technically be completed from any computer, we decided against providing participants with a direct link to the experiment.

Although this setup would have likely resulted in more participants, and thus a higher power in the analysis, our main concern was that users would use a dictionary to check the spelling of words. This would have completely contaminated the results. We considered modifying the AABST tool to technically disallow users from opening other applications while completing the experiment; however, even such a precaution could not prevent participants from using a dictionary on another device or using a hard-copy dictionary.

In fact, even whilst in the lab we observed one participant refer to an online dictionary while completing the tasks and had to discard their results. This concern is likely further exacerbated by the incentive (£80 Amazon voucher) we offered to any participant who successfully edited all errors. However, the reason we added this incentive is because in many non-critical domain experiments, participants may exhibit laziness or fail to take the experiment seriously. We acknowledge this concern, but we also see benefits in carrying out the experiment in such a non-critical domain as it represents a low-resource option that allows us to test our hypotheses and because it has previously shown results extendable to other more critical domains such as mammogram reading (the similarity in the results of Galletta et al. (2005) and Alberdi et al. (2010)). The incentive was one way for us to address this concern. We also included, in the briefing, both written and verbal reminders for participants to do their best to help us collect reliable data.

Another reason that we chose the lab set up rather than an Internet-based experiment is that we wanted to control the amount of time each participant had to edit each passage and there was currently no option to do this using the AABST Tool. There were other advantages to having the experiment in the lab including more control and thus consistency on external factors such as lighting, screen size, temperature, etc. as well as a higher probability that participants would follow the briefing instructions as they were verbally read to them. In Gray's experiment (2012), many participants failed to use the Mozilla Firefox browser despite warnings in the briefing. Unfortunately, those participants' responses were wasted as the tool does not successfully record responses made in Internet Explorer.

Briefing:

A few days before participants' scheduled experiment time, participants were sent a reminder email which contained:

- Exact location of the computer lab, including a map
    - This was especially important for participants from outside the university

- A reminder to bring headphones with them and that we would have extra pairs in case they can't or forget
- A chance to reschedule in case they were no longer able to make the chosen time

Upon arrival in the lab, participants were greeted and given the link and individualized login credentials to the AABST Tool. The welcome screen of the AABST Tool contained the following information about the experimental procedure, the tool's behaviour, and types of corrections expected. Furthermore, this exact information was also read, by the experimenter, to the participants:

> Welcome and thank you for your interest in our study!
>
> In this first part of the experiment, you will be shown four different text passages. For each passage, you must correct as many SPELLING errors as you can before the time expires.
>
> You will be aided by a unique spell checker that aims not only to detect invalid English words, but also incorrect usage of words and especially homophones based on sentence structure and context. Homophones are words that are pronounced the same but differ in meaning and spelling such as "to"/"two"/"too". The tool is not 100% accurate and has a limited vocabulary. This means that it may sometimes highlight, in red, words or homophones that have been correctly spelled or used, and that it may also sometimes fail to highlight mistakes.
>
> Please remember:
>
> 1. To change a word, you need to click it. It will then appear in the box on the right hand side of the screen. Change it to the desired spelling. Then press "Make Change" to see the change appear in the passage.
> 2. You have a limited time to edit each passage, so please pay attention to the stopwatch projected on the lab screen. The experimenter will announce when there are 3, 2, and 1 minute left for each passage.
> 3. There is no need to correct grammar, punctuation (example: apostrophes or dashes), or to insert/delete words.
> 4. Where there are differences between American and British spellings, only the British spelling of the word is assumed correct in this study.
> 5. It is very important not to use the "Back" or "Refresh" buttons in your web browser.

6. The experiment aims to measure the effect of different spell checking tools - not to judge your spelling ability.
7. Do your best to edit all errors to help us collect reliable data, but also to entitle you to be entered into a draw for an £80 Amazon voucher.

We note that the briefing contains a rational explanation for the tool's behaviour. The tool made more false positives and false negatives than an average spell checker and we needed to rationalize this so as not to lose participants' trust. The way we did this was to explain that the tool had a limited vocabulary (thus false positives) and did not have a perfect homophone detection algorithm (thus false negatives and false positives).

Passages:

Participants read one easy and one hard passage, at each of the two different tool settings, where easy and hard refer to the amount of difficult, misspelled words in each passage (Level 2 and 3 versus Level 1 words). We will discuss our definition of task difficulty further in Section 4.3.6.4. Furthermore, a briefing between the two settings (written and verbal by the experimenter) was included to introduce the second setting as a separate tool and thus help erase any learning effects.

Tool setting, and the order of easy and difficult text was randomized. Text between the two settings was also switched to prevent confounding factors (Figueredo, 2004). Randomization of the tasks will be further described in Section 4.3.7 while the choice of passages and errors will be discussed in Section 4.3.6.

Questionnaire:

During the experiment in the lab, users were presented with two questionnaires, one following each of the two tool settings. These were hosted on Qualtrics. Using a 5-point Likert scale, participants were asked about:

- Self-Confidence: I rate myself as an _____ speller
- Subjective, self-reported, trust: I trust this tool
- Tolerance: I am annoyed by this tool's flagging of correctly spelled words (false positives)
- Perception of tool reliability: I find this tool reliable

Unique to the second questionnaire were three questions to measure participants' understanding of each tool setting:

- Compared to the first spell checker, I found the second spell checker failed to highlight ____ mistakes (more/fewer)
- Compared to the first spell checker, I found the second spell checker incorrectly highlighted ____ correctly spelled words (more/fewer)
- I preferred working with the ___ spell checker (first/second/I did not notice any difference between the two tools)

Assessment Task:

Participants' scores in the assessment task were used as a measure of their spelling ability and helped us stratify users for the analysis stage. We debated several different types of assessments including: (1) a standard test such as the Diagnostic Spelling Potential Test (Figueredo, 2006) or the Wide Range Achievement Test (Fischer, 1985), (2) a passage for editing without any tool markings, (3) a posteriori analysis, or (4) a dictation of words based on Fischer's publication (1985). Table 4-7 summarizes the advantages and disadvantages of each of these options.

Table 4-7: Options for Assessing User Spelling Ability		
Option	**Advantages**	**Disadvantages**
Wide-Range Achievement Test (dictation)	Robust and reliable as it is a widely used conventional test of spelling ability Easy to administer and design	Doesn't test ability to *detect* errors Doesn't include homophones Must be purchased
Passage without tool markings	Tests ability to *detect* and correct errors Has the flexibility to include homophones Enriches the analysis with a "no tool" condition	Difficult to guarantee that our choice of passage will accurately distinguish between high and low verbals because of confounding factors in the passage such as difficulty, subject material, type of errors, etc. Difficult to include a range of unrelated words
A Posteriori	Easiest option Includes homophones	Is clouded by effects of the tool on different users, especially that previous publications have shown that sometimes high verbals perform worse with tool support
Multiple choice or dictation of Level 1-3 words (identified by Fischer (1985))	Scores in the dictation of these words have shown high correlation with scores in the Wide Range Achievement Test, ($r=0.84$, $p<0.001$) Errors inserted in the passages are of the same word types and thus correlation between performance in the assessment and passage editing tasks are expected to be high Less work for participants compared to editing another passage	Doesn't test ability to *detect* errors Doesn't include homophones
Standardized test scores (as done by Galletta et al. (2005))	Reliable tests that rely on the design of those more knowledgeable in the language field Least work for participants	Complicates ethics application May deter some people from participating

Based on these trade-offs, we chose to dictate, using audio clips from dictionary.com, and a survey created in Qualtrics, words from the list provided by Fischer et al. (1985). This option meant less work for participants, but results whose reliability would likely be comparable to a robust test such as the WRAT and whose correlation would likely be high with the ability of users to correct errors in the passages (as errors are of the same word types). This latter advantage could also enrich the results by giving us the ability to look at users' reactions to tool support in each of the three word levels compared to their ability in those levels assessed in the dictation. Furthermore, the mix of Level 1 to 3 words meant that the assessment task measures a range of difficulty levels while relying on different spelling techniques: users' phonetic realization of sounds (Level 1 words),

their understanding of common orthographic and morphological conventions (Level 2 words), and their recall of uncommon English patterns (Level 3 words).

The greatest disadvantage of this choice is that it does not assess users' abilities to *search* a document for errors, and this distinction is important as noted by the study by Figueredo et al. (2006) which showed that people may fail to detect a spelling error even when they possess spelling knowledge of the word. However, the option that would allow such a distinction was to write a passage without tool markings and ask users to edit it. However, we avoided this option because of possible confounding factors that may affect users' results in this task besides the extra work for participants and the difficulty in creating such a reliable passage.

The dictation was performed in the final part of the experiment to avoid discouraging participants from the start of the experiment. This was especially feasible because we did not use participants' scores to assign them to certain conditions and relied on random assignment. A list of the thirty words we used, ten from each level, are available in Appendix 4E.

De-briefing:

The final screen in the ABBST Tool contained a thank you statement and a statement about the importance of the user's contribution. It also included contact information for participants that wished to provide feedback. Participants were also given the choice to receive (after all participations were complete), a copy of the corrected passages via email. This idea was inspired by comments from some of Gray's (2012) participants who mentioned that they would have liked to receive information about their performance in the study.

4.3.4 Experimental Output Variables

From the briefing, assessment, passages, and questionnaires discussed in the previous section, we collected the following outputs for analysis.

From the **recruitment** phase, we collected mostly demographic information about each participant:

- Email
    - Inputted twice in case of errors
- Gender
- Age
- Whether they are a native speaker of English

- How many years they have been speaking English
    - Important measure for distinguishing ability of non-Native English speakers
- Self-confidence in their spelling ability
- General, self-reported trust in spell checkers
- Educational background
    - Whether they are students, postgraduate/undergraduate, current year of study, and department they are studying in
- How many non-academic books they read per year
    - Has been previously associated with spelling skill (Fitzgerald, 1987)
- How often they read a newspaper

From the **assessment**, we calculated a score for each reader according to the number of words they correctly spelled in the dictation.

From the four **passages** (easy and difficult passages at two different tool settings each) that each reader edited using the AABST Tool, we collected a list of all the changes the participant made to each passage. From this list, we derived a simple count of:

- Human TPs
    - Number of incorrectly spelled words the subject changes
    - These can be further divided according to the tool's behaviour into marked misspellings and unmarked misspellings
- Human FNs
    - Number of incorrectly spelled words the subject fails to fix
    - These can be further divided according to the tool's behaviour into unmarked misspellings and marked misspellings
- Human FPs
    - Number of correctly spelled words the subject changes
    - These can be further divided according to the tool's behaviour into marked correct words and unmarked correct words

From the two **questionnaires** (after each tool setting), we had a score, on a 5-point Likert-scale, for each of the following factors, measured for each tool:

- Tolerance of false positives committed by the tool
- Trust in the tool
- Self-confidence in spelling ability
- Perception of tool's reliability

From the second questionnaire, for each user, we also had:

- Preferred tool setting
- Which tool they thought made more false negatives
- Which tool they thought made more false positives

4.3.5 Experiment Participants

We recruited participants using posters, leaflets, email lists, and class announcements. The assessment was used to distinguish high and low verbals. We did not apply any inclusion criteria except that participants had to be over 18 due to ethical constraints. We aimed to recruit a range of spelling abilities from different backgrounds, schools, and ages as we wanted to amplify differences between high and low verbals to help us approach real-world environments like a cancer screening program where there are varying user abilities. Also, the fact that some foreigners may perform better than natives is not a problem because users were divided using their score in the assessment, not their level of study or nativity.

The experiment took a total of 45 - 60 minutes to complete. Participants were offered an £8 Amazon voucher, sent via email as an incentive for their time. We also offered an incentive of an £80 Amazon voucher for anyone who successfully corrected all spelling errors. Unfortunately, despite these incentives it was not easy to recruit participants. People often cancelled their bookings or failed to show up to the time slots they selected.

4.3.5.1 Sample Size

We calculated the target sample size we needed to recruit by relying mostly on Galletta et al.'s (2005) study as it is the most similar to our experiment. Using the following steps, we found that we need to recruit approximately 47 participants.

1. We set Type I error = 0.05 (standard in the literature) where Type I error describes the probability that the difference we detect is due only to chance.
2. We set Type II error to 20% (i.e., power = 0.80 which is typical in the literature) where Type II error describes the probability that we fail to detect a difference when in fact there is a difference
3. Next we calculate the expected effect size, which is the difference in the value of the variable in the two experimental groups and is measured based on previous studies using Cohen's f (since the previous study was based on an ANOVA analysis).

- The effect we are measuring is the difference in performance between high and low verbals as they correct words of varying difficulty using two different spell checker tools which vary in their thresholds. One of the closest publications in the literature is that by Galletta et al. (2005) where high and low verbals edit a passage for spelling and grammatical errors with or without a tool.
- However, the study is different than our experiment because it encompasses spelling and grammatical errors, the difficulty of the errors and their effect on performance are not reported, and most importantly the difference in performance measured is between users with the tool versus users without the tool, not between users using two tools of varying thresholds. These factors are expected to decrease the difference between the two groups although we do not know by how much.
- Using the results in the ANOVA Table published by Galletta et al., we calculate Cohen's *f* as follows:
  - The most important effect we are interested in is the interaction between user ability and use of the checker. The variance explained by this effect is 81.387. The error variance is 455.784.
  - Using G* Power to calculate, this means that $\eta 2=0.1515$ and Cohen's $f=0.423$.
  - According to the literature, this is a large effect size (>0.4)

4. Using the calculated effect size, we plug in the remaining variables in G* Power to estimate the required sample size.
   - Test family: F test
   - Statistical test: ANOVA Fixed effects, special main effects and interactions
   - Effect size: 0.423
   - Type I error probability: 0.05
   - Type II error probability: 0.80
   - Number df: (Ability groups-1)x(Checker groups)=(2-1) x (2-1)=1
   - Number of groups: 4 (Ability groups x Checker groups=2x2=4)
   - Output = 47 participants
5. However, as stated above, due to differences between the study we are planning and Galletta's study, we expect that the effect size could be smaller – although we do not know by how much. We create the plot in Figure 4-7 to see how differences in the effect size will affect the number of required participants while maintaining the same confidence level and power. From the plot we see that if the effect size were not large, but medium (f=0.25 as defined in the literature), then the required sample size would increase from 47 to approximately 125. This is a more difficult number to attain but still possible, and could be considered a goal if no significant differences are detected after reaching 47 participants.

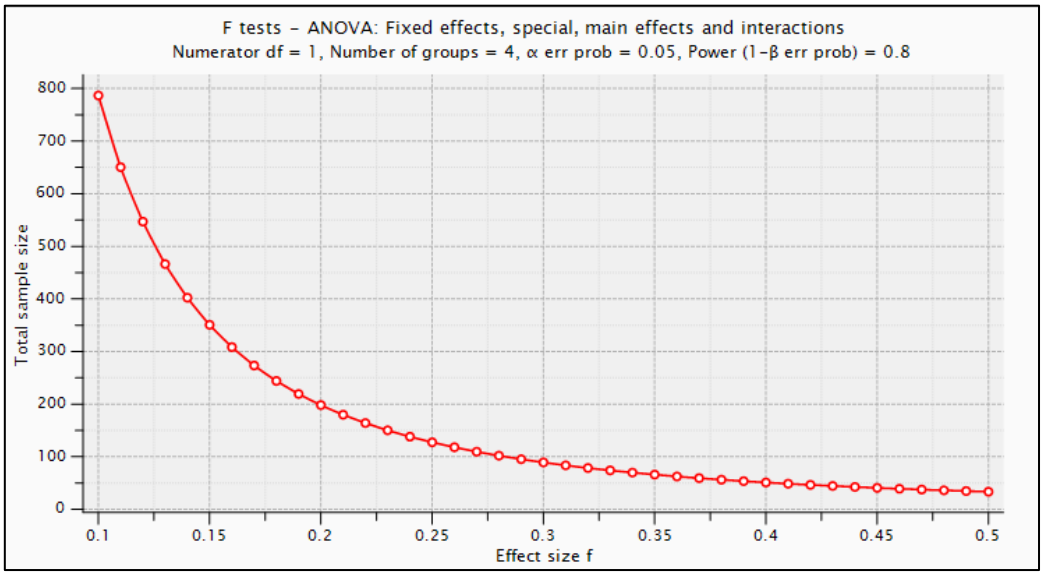


Figure 4-7: Relationship between Effect Size and Required Sample Size

4.3.6 Passages of Text for Editing Task

Perhaps the step with the most design issues was the choice of passages and errors for users to edit. We were faced with a series of decisions about the origin, difficulty, errors, and randomization of the passages. In the following, each will be discussed in a separate subsection.

4.3.6.1 Choice of Target Words

As noted in Section 4.2.5, our hypotheses rely on two key behaviours: (1) high verbals will not fall for some of the tool FPs, and (2) low verbals will not be negatively affected by the tool's increased FN rate. In turn, these two behaviours depend on the prevalence of errors in passages, the number and types of tool errors and are more likely true for: words of Type A (high verbals know, low verbals are uncertain about) or B (high verbals know, low verbals don't know), and words of Type C (high verbals are uncertain about and low verbals don't know), respectively.

This reliance on differences between high and low verbals' skills in each of the levels encouraged us to rely on Fischer's (1985) list of words because, as previously discussed in Section 4.1.3, there was already evidence from his experiment that high and low verbals show significant differences in performance between each of the three levels, and that this difference is greater for words in Level 2 and 3.

This meant that we needed to write four passages of text that contained target words from the list provided by Fischer (1985). This was a bit of a challenging writing exercise. We also needed to include a significant amount of homophones in the passages. For this, we relied on websites and Evans' (2013) list of common errors. Some of these sites had the added advantage of listing these homophones according to difficulty, but we also relied on our own judgement for this division, with mostly infrequent homophones classified as more difficult.

For the sake of realism and ecological validity, we used the average number of errors committed by Fischer's participants in each word category as a way to guide us on the number of these target words that should be misspelled. This naturally resulted in more spelling errors in Level 3 words compared to Level 1 words and the exact percentages of misspelled words was approximately 50%, 30%, and 10% in Levels 3, 2, and 1, respectively. This also resulted, across passage difficulty, in equal frequencies of misspellings in each word category.

4.3.6.2 Choice of Tool Setting

Deciding how many of these incorrect words would be highlighted by the tool was a difficult decision. We rationalized, firstly, that the number of FNs by the two tools would essentially be equal because an incorrectly spelled word will not exist in either dictionary and so will be highlighted by both tools. Except for cases like gilt/guilt or tongs/tongues where the word was changed into another word which was valid and only existed in the larger dictionary setting. In this case, the specific tool exhibited slightly more tool FNs compared to the sensitive tool.

The number of tool FPs and the number of FNs due to homophone errors was decided using a think-aloud process imagining user's reactions to each word using the flowcharts in Figures 4-1 to 4-4. We aimed to balance the number of these errors to best observe the trends we expected: that when using the sensitive tool, high verbals would be more positively affected by the tool's decreased FN rate than negatively affected by the tool's increased FP rate, but that low verbals would be more negatively affected by the tool's increased FP rate than positively affected by the tool's decreased FN rate.

We acknowledge that this process relied on subjective processing and predictions. However, in Section 4.4.3.6, we present a more formal approach based on mathematical formulae for choosing the tool setting. This formal approach requires numerical probabilities of users' reactions rather than general trends as we have from the flowcharts in Figures 4-1 to Figure 4-4. Instead, this formal approach could be implemented in future experiments while feeding the numerical results from this study into the formulae to select suitable tool settings.

Furthermore, we aimed to choose tool settings that when combined with the percentage of errors in each word level would result in a whole (rather than fraction) number of words both in the easy and difficult passages, because with such a small number of target words, a single word can really change the apparent sensitivity/specificity of the tool. For example, easy passages contained 6 Level 3 words and difficult passages contained 12. From Fischer's publication, participants misspelled 50% of these words, which results in a whole number of incorrectly spelled words for both passage difficulties (3 and 6). We assigned a sensitivity of 1.0 to the sensitive tool; thus, 3 and 6 words were highlighted, respectively. We assigned a sensitivity of 2/3 to the specific tool, which again resulted in a whole number of highlighted words for each setting (2 and 4). This consistency was maintained across all three levels of target words and homophones.

Finally, we note that in a real-world setting, tool settings are often pre-determined by tool designers. Then, users and their managers need to decide which of these tool settings is best for each group of users. However, in designing this experiment, we were essentially faced with the opposite scenario: looking for tool settings that would exhibit user differences. For example, Hologic's latest CAD tool for the detection of breast abnormalities has three separate operating points (sensitivity/false marker rate settings) which can be selected for both microcalcifications and masses; nine combinations in all (The Smart Choice 2). In such a real-world setting, the head of a screening program or radiology department or even a researcher asking which tool setting to choose for each user is not faced with the subjective judgement process we had when designing our own experiment.

4.3.6.3 Error Generation

In his book, Evans (2013) listed a number of reasons why people make spelling mistakes. These reasons can give us ideas about how to generate some of the errors so that they appear as natural as possible:

- Quick typing

  Note: single letters introduced could change a word into another acceptable homophone (find and fined) or another correctly spelled, but incorrectly used word (hoping and hopping)
- Auto-correct
- Lack of knowledge on how to use certain words
- Common misspellings which a user may not realize are wrong (ex: appartment)
- Distractions while writing

- Errors made by optical character recognition from scanned or bitmapped documents
- Tiredness or boredom

Based on this list, some ways we could *generate* errors are:

1. Look for lists of commonly misspelled words and their common misspellings
2. Search for lists of commonly confused homophones
3. Look for auto-correct databases
4. Have an experiment where we dictate the text to a group of people and use the results for our experiment. Increase errors by having them type under:
    i. Distractions
    ii. Time pressure
    iii. As they multi-task
    iv. When they are tired or bored
5. Produce bitmap outputs at low resolution and then use optical character recognition to produce the passages

Although options 4 and 5 would probably yield the most natural results, they do not give us the ability to limit the errors to the target words identified in Fischer's publication and ensure a common number of target words across passages. Instead, we searched for the target words in lists of commonly misspelled words. For target words not in these lists, we focused on common pattern errors. For example, replacing words ending in ance with ence (clearance), or ible with able (exhaustible). Evans' (2013) list of common mistakes was also useful in identifying homophone errors such as loose versus lose and stationary versus stationery.

4.3.6.4 Passage Difficulty

When writing the passages, we were faced with an important question: what do we truly mean by easy versus difficult text? Three possible meanings were discussed:

- Difficult passage contains more words of Type 2 and 3 and more difficult homophones
- Difficult passage contains same amount of Type 2 and 3 words as the easy passage but overall is harder to understand because it's about a more complex topic
- Difficult passage contains same amount of Type 2 and 3 words as the easy passage but overall is harder to understand because it's about an unfamiliar topic

Another level of difficulty is in the type of misspelling. For example, some misspellings are easier to identify because they represent an implausible sequence of letters in the English

language. Such misspellings may be easily corrected, despite being in a "difficult" word. For example, switching the order of the letters 'i' and 'e' may be a more difficult error to notice and more confusing to correct in an easy word such as "friend" compared to replacing the 'b' in the difficult word "bureaucracy" with the nearby letter 'v' on the keyboard to produce "vureaucracy".

We chose to define difficulty in terms of the number of target words from Level 2 and 3 and difficult homophones it contained. Misspellings were mostly generated by switching often confused patterns rather than creating obvious keyboard-type errors. This was because, to match our modelling, we wanted to avoid scenarios where users could simply guess the correct spelling of a word without previous knowledge of it.

We also decided against adding another confounding factor to the analysis which was the familiarity and complexity of the topic, which previous researchers have commented on (Figueredo, 2004). Difficult passages contained 12 target words from Level 3, 12 from Level 2, and 13 homophones which were mostly difficult. Easy passages contained 6 target words from Level 3, 6 from Level 2, 8 from Level 1, and 13 homophones.

4.3.6.5 Source of Passages

There were several possible sources we could rely on to generate passages for editing. Gray (2012) used Wikipedia articles after gaining permission from the company to slightly modify some of their articles by inserting errors. Fitzgerald (1987) used student essays. We explored these options as well as others before settling on independently writing the passages. We discuss the advantages and disadvantages of the different options we considered in Table 4-8. We use the term *target words* to refer to the specific list of words by Fischer (1985) that we chose to use in our experiment for the reasons outlined in Section 4.3.6.1. *Target words* also refer to homophones in the passages.

Table 4-8: Advantages and Disadvantages of Possible Sources for Passages		
Option	**Advantages**	**Disadvantages**
Independent writing of passages	No copyright issues Most freedom in error type, content and length	Could be difficult to get good content Hasn't gone through formal editing for mistakes unnoticed by the writer
Books or novels	Less work than independently writing passages	Copyright issues unless very old Difficult to find target words consecutively in a short passage
Proofreading examples (on websites or in books)	Already tailored for our same need in terms of interest and high frequency of errors in a short passage	Permission required Still need to manually insert our target words
Newspaper articles	Interesting material	Permission required Difficult to find target words consecutively in a short passage
Wikipedia articles	Interesting material Have copyright permission (Gray, 2012)	Difficult to find target words consecutively in a short passage Couldn't find a way to search the entire domain - only individual pages
Student essays	Used in previous publications Easier to obtain permission Likely already contains some mistakes	Difficult to find target words consecutively in a short passage

The main reason we chose to independently write our own passages was that we had a list of 120 specific words from Fischer's publication (1985) which we wanted to insert into the passages along with a reasonable number of homophones. Relying on Fischer's list ensured that errors: would be strongly correlated to the assessment task, would show strong differences between the performance of high and low verbals, and would give us a clear definition of error and passage difficulty. These were important advantages and it would have been very difficult to find a passage which contained these words or synonyms (so we could replace them with the specific word) with a high enough frequency and close enough proximity. The passages were written by Marwa Gadala and then edited by Professor Lorenzo Strigini. Passages needed to be interesting, but neutral, neither offending nor causing distress or discomfort to any of the participants. For interesting topic ideas, we considered relying on:

- Proofreading exercises
- Newspaper/magazine articles
- Old TOEFL/IELTS exams
- Typing practice passages
- Books
- TED talks

The final topics were: the battle against Malaria (based on a recent newspaper article and TED talk), four habits of highly effective people (based on Stephen Covey's famous book), a business letter (similar to Galletta's idea and based on a recent newspaper story), and how to save a wet phone (based on a number of magazine articles). The former two were the more difficult passages while the latter two were the comparably easy passages. For reference, a copy of the passages are included in Appendix 4F with all target words underlined.

4.3.6.6 Timing for each Passage

Previous work has shown that increased time pressure is related to higher levels of automation bias (Alberdi, 2009; Goddard, 2010) which has been explained by users' increased satisficing behaviour and reduced vigilance (Goddard, 2012a), users' shift of more effortful processing to heuristic-based decision-making (Skitka, 2000), or over-reliance on automation to provide correct output so that attention can be channelled towards other tasks (Goddard, 2012b). Thus, to amplify automation bias effects, we needed participants to have enough time to finish reading the passages but to still feel pressured for time. We did not want to have cases where people failed to correct an error at the end of a passage simply because they didn't have time to reach that part, but we did want to induce some of the mechanisms described by previous research.

It was challenging to find a time that would achieve this as we had a range of spelling abilities. We needed to ensure that even the least verbal user would be able to at least complete reading the passage. To choose an appropriate time, we ran the experiment first on a number of friends and family members of varying spelling abilities to get a reasonable time, which we found was 6 minutes for the easy passages and 8 minutes for the difficult passages. Then, in the first experimental session that we ran, at the end of each passage we would ask the participants whether they had completed the passage. Based on their feedback, we increased the time for the more difficult passage by 1 minute to a total of 9 minutes. All sessions thereafter were run on this basis, and as an experimenter observing the participants, most participants seemed busy for the entire timeframe, only very few finished before the allotted time, and none complained of being unable to finish reading the passages in the allotted time.

4.3.7 Randomization of Conditions

To avoid effects due to ordering, randomization was required in our experiment on several levels: tool setting, passage difficulty, and specific passage text. Recall that our main experiment set up is as follows:

Tool Setting 1		Tool Setting 2	
User edits easy passage	User edits hard passage	User edits easy passage	User edits hard passage

1. We want to ensure that the order that the two tools are presented in does not affect the results. Thus, we need two experimental conditions, where 1 and 2 refer to the sensitive and specific tool, respectively:
    - User sees Tool 1 then Tool 2
    - User sees Tool 2 then Tool 1

2. We also want to ensure that the order of the text difficulty does not affect the results. Thus, we need four conditions (E=Easy, H=Hard, |=differentiates between Tool 1 and Tool 2):
    - EH|EH
    - HE|HE
    - EH|HE
    - HE|EH
3. We also want to ensure that the actual text chosen for each tool setting does not affect the results. For each difficulty level, we have two passages (A and B). This means that when combined with difficulty level, we have four possibilities: EA, EB, HA, HB. Thus, for each group in (2) above, we have 4 possibilities (2x2x1x1). For example, EH|EH in (2) above can be either:
    - EAHA|EBHB
    - EAHB|EBHA
    - EBHA|EAHB
    - EBHB|EAHA

Thus, to cover all these possibilities we would need 2x4x4=32 conditions. However, we need to reduce this number because it's almost as large as our sample size so we may end up with groups with only one participant and each condition currently requires a separate installation of the tool and separate database which all need to be collated at the end – thus this poses some technical difficulties. To reduce the 32 versions:

1. We reduce the four versions in (2) above (EHEH, EHHE, HEHE, HEEH) to only two versions: either the user sees hard text first or the user sees easy text first:
    - HE|HE
    - EH|EH

2. We add a constraint to (3) above that for each tool setting, the user must see both passages from the same set (either Set A or B but no mixing). Thus, for each group in (2) above, instead of 4 possibilities (for example, EH|EH in (2) above can be either: EAHA|EBHB, EAHB|EBHA, EBHA|EAHB, EBHB|EAHA) we have only two options:
    - EAHA|EBHB
    - EBHB|EAHA

This reduces the number of versions/conditions we need to eight (2x2x2). Because the AABST Tool unfortunately does not support more than one active experiment at a time, the tool had to be cloned and hosted on eight separate sites, one for each version, and results later collated. However, this extra effort was necessary to avoid confounding factors due to the order of the tool settings, the order of passage difficulty, and the specific passage used in each setting.

4.3.8 Issues in the AABST Tool and How They Were Resolved

The AABST Tool, which we used to record users' edits to the four passages, was first developed by undergraduate students at City University and later modified by Cristina Gray as part of her Master's degree. Despite these modifications which greatly improved the original tool, we had a few obstacles while implementing our experiment. Due to technical and time constraints it was not an option to modify the tool itself, and instead we needed to think of work-around solutions. Table 4-9 summarizes the most significant of these issues and the temporary solutions we used to resolve them.

Table 4-9: Issues in the AABST Tool and How They Were Resolved	
Issue	**Resolution**
Creating a participant account does not work consistently	We create participant accounts after the recruitment phase and before participants come into the lab so that if there is a problem, we just re-attempt creating the same account
Tool only allows one experiment to be active at a single time, but due to randomization needs, we need eight versions of the experiment	We create eight clones of the tool and host each version on a separate website. This was necessary despite complicating data analysis when collating results from the different databases.
Deleting a sub session within a session, or a task within a sub session, causes an irreversible error	If while defining the experiment we make an error in one of the sub sessions, we delete the entire session and start over
Tool doesn't provide a way to brief participants between passages or tool settings. It only allows briefings at the beginning and end of the experiment.	We create tasks in the experiment which look exactly like a passage but which contain briefing text. Also, as an experimenter, we read out these briefings verbally. Unfortunately, these are editable if users try to edit them.
If a participant uses Internet Explorer, the tool's database does not record the changes and they do not appear on the screen.	We include a statement in the briefing about using Mozilla Firefox and manually check, as an experimenter, at the start of each session, that all participants are using the correct browser.
Tool does not support user questionnaires, which we need after each tool setting.	We host these questionnaires on Qualtrics, but unfortunately this means that participants need to open several windows/tabs during the experiment.
Tool does not support blank lines in the definition of the passages, but the passages are long and consist of several paragraphs.	We divide paragraphs using a series of dashes to provide a visual division that is technically feasible in the definition of the .xml files used as input to AABST.
Tool records all changes made on any word in the passage and several changes to the same word are logged as separate entries.	We manually check the list of changes for duplicates and remove them.

For future experiments, these issues may serve as a useful starting-point for improving the AABST Tool.

4.4 Results

In this section, we present results of the experiment described in Section 4.3. We start with demographic statistics (Section 4.5.1), ability-related data (Section 4.5.2), most importantly results from the actual passage editing tasks (Section 4.5.3), and finally results of user perceptions (Section 4.5.4).

4.4.1 Demographic Statistics

A total of 47 people participated in the study over a total of twenty 1-hour lab sessions. 79% were female and 77% were students from several different universities. City University students were from a range of different schools including the School of Mathematics, Computer Science and Engineering, CASS Business School, the School of Law, the School of Arts and Social Science, and the School of Health Sciences. The majority of student participants were postgraduates (78%). The average age of participants was 30 but some were as young as 18 and the oldest was 58. About half the participants were native English speakers (55%) and a range of self-assessed spelling abilities were represented. Non-native English speakers spoke a diverse range of languages including: Russian, Ukrainian, Arabic, Farsi, Cantonese, Indonesian, Italian, Urdu, Greek, Lithuanian, Spanish, Bulgarian, Telugu, Sinhalese, and Polish. These demographic results are summarized in Table 4-10 along with other interesting factors such as the distribution of participants' reading habits. Numbers in parentheses represent actual number of participants in each category.

Table 4-10: Demographics of Study Participants

Characteristics	Mean ± SD
Gender (Female)	79% (37)
Student (Yes)	77% (36)
Postgraduate	78% (28)
Age (years)	30 ± 8
First language (English)	55% (26)
Years speaking English	22 ± 11
Self-reported spelling ability	
Poor	2% (1)
Average	13% (6)
Good	60% (28)
Excellent	26% (12)
Non-academic books read per year	
None	4% (2)
1 or 2	11% (5)
3 to 10	28% (13)
10 to 40	40% (19)
> 40	17% (8)
Newspaper reading habits	
Rarely	21% (10)
Few times per month	34% (16)
Every day	32% (15)
More than one every day	13% (6)

4.4.2 Ability Statistics

The assessment task, the dictation of 30 Level 1, 2 and 3 words from Fischer's (1985) list, resulted in a mean score of 21 ± 4.8. The maximum score achieved was 30 and the minimum was 9. When

analysing the results, we discovered that many participants had problems distinguishing the exact word being dictated, even though they were provided with an audio clip that they could repeat as many times as they liked. We decided against penalizing users for mistakes resulting in valid spellings as there was no way of knowing whether the participant would have correctly spelled the word had they heard it correctly. However, in future studies, the dictation should include not just an audio clip of the word, but also a sentence putting this word into context in order to avoid having to accept responses such as those in Table 4-11. Also included in Table 4-11 are cases of alternate spellings which we accepted for some words (for example, dialog and dialogue).

Table 4-11: Accepted Spellings in the Dictation Task

Word	Accepted Spellings
Hollow	Hello, parlour, hallow, polo
Inactive	Enactive, be inactive
Conjugate	Congregate
Gauge	Gage, engage
Dialogue	Dialog
Jeopardy	Japanese
Asymmetric	Isometric
Gastronomy	Astronomy
Manoeuvre	Maneuver
Rendezvous	Rendez vous, wonderful
Relieve	Relief, relieved, we leave
Clannish	Planish, plenish, planet
Youthful	Useful
Recurring	Reoccurring

We analysed different factors as indicators of spelling ability. This analysis was motivated by curiosity, but also to expose any peculiarities in the recruitment sample and also to help organize the classification or randomization of participants in potential future studies. Spelling ability was measured according to results in the 30-word dictation. To measure association with ability we used: independent t-test for nominal variables, 1-Way ANOVA for ordinal variables, and Pearson's correlation for continuous variables. Factors that had a significant effect on participants' score in the dictation were: (1) whether English was the participant's first language, (2) number of years the participant had been speaking English, (3) self-reported spelling ability and (4) number of non-academic books the participant read per year. Thus, as expected, there was a strong difference between native English speakers and those that spoke English as a second language, as well as a positive correlation between ability and number of years of spoken English. Interestingly, we also noted a strong association between spelling ability and book reading habits as previously noted by Fitzgerald et al. (1987). Also interesting was the strong association between self-reported spelling ability and true ability suggesting that this could be a reliable measure when stratifying participants in future studies. No significant differences in ability were

found based on gender, graduate level of study (matching Galletta's (2005) results), age, school, or newspaper reading habits. The complete results of these explorations are shown in Table 4-12.

Table 4-12: Indicators of Spelling Ability

Factor	Association Test Used	Significance
Female/Male	Independent t-test	t(45)=-1.03 p = 0.310
Undergrad/Postgrad	Independent t-test	t(34)=-.286 p = 0.777
Native English speaker/Not*	Mann-Whitney U Test because normality assumption was violated	U=144 p = 0.005
Age	Pearson correlation	0.114 p = 0.451
Years speaking English*	Pearson correlation	0.516 p < 0.0005
School	1-Way ANOVA	p = 0.266
Books read per year*	1-Way ANOVA	p = 0.004
Newspapers reading habits	1-Way ANOVA	p = 0.197
Self-reported spelling ability*	1-Way ANOVA	p = 0.001

*Significant result

Similar to results shown by Fischer (1985), we note that, as expected, the highest mean frequencies of errors were on words of orthographic representation Level 3 and the lowest on those of Level 1. We defined two groups of user ability, high and low verbals, as those above and below the mean score in the assessment task. We found strong differences between the performance of high and low verbals in each level, and these differences increased with level difficulty. We tested the significance of these results using the independent t-test for scores in Level 3 as the normality assumption was not violated, but the Mann-Whitney U Test for scores in Level 1 and 2 as their distributions violated the assumption. The differences in the performance of high and low verbals in each level were significant: $U=67.5$; $p<0.0005$ for differences in Level 1, $U=67.5$; $p<0.0005$ for differences in Level 2, and $t(45)=-6.35$; $p<0.0005$ for differences in Level 3. These significant results are positive indicators about our design decisions, which aimed to choose groups of words of varying difficulty level that high and low verbals performed significantly different on. A visual representation of these results is shown in Figure 4-8.

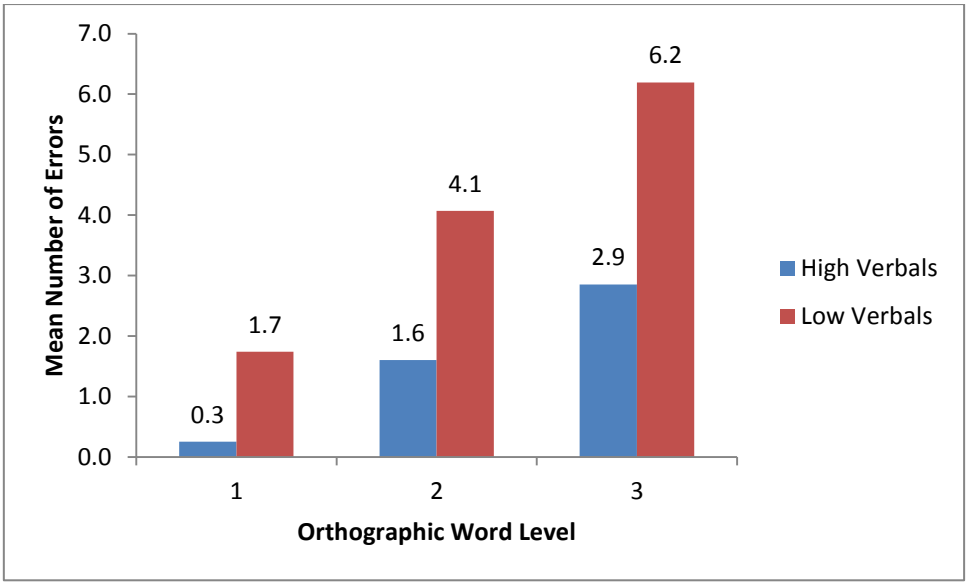


Figure 4-8: Mean Number of Errors by Orthographic Word Level and User Ability

4.4.3 Results of Passage Editing

Demographic and ability data show a wide range of user abilities, which reveal expected differences according to the orthographic representation of a word. However, the most important results are those from the actual passage editing tasks and these are presented in this section.

4.4.3.1 Correct Action versus Correct Result

We divide our analysis of the results of passage editing into two main categories based on a distinction which we noticed when closely analysing participants' changes. From the unique corrections of some participants, we noted that sometimes a user would change a word that was correctly spelled and replace it with another word that is also correctly spelled. In this case, the user's action is deemed incorrect as they have attempted to change a correctly spelled word suggesting that they had reason to believe it was incorrectly spelled. However, the resulting word is correct, both in spelling and within the context of the sentence. For example, changing "preventive" to "preventative" or "afflict" to "affect" or "incidence" to "incidents". We call these corrections *acceptable FPs*.

By the same token, sometimes a user would change an incorrectly spelled word but their attempt would result in a word which is still spelled incorrectly. In this case, the action is correct while the result is incorrect. Many times these changes looked like guesses from the users. For example, changing "ellicit" to "elicit" rather than "illicit" or "cite" to "site" rather than "sight". We label these corrections as *attempted TPs*.

Another interesting correction we noticed is when users would change an incorrectly spelled word by replacing the word rather than correcting it. We label these changes as *acceptable TPs* and although the corrected word is not what we expected, we always count these types of corrections as TPs because both the action and result of the user are correct. For example, changing "continuance" to "commitment/continuity" or "disperse" to "dispense" or "sergeant" to "agent".

This fine distinction between users' actions versus results affects all further analyses. We limit our presentation to analyses based on correct user results, not correct action. This is for two main reasons. Firstly, the category of acceptable FPs contained many changes of words from spellings with an "s" to spellings with a "z" such as victimize, minimize, and organize. Some of the passages naturally contained a significantly higher frequency of these words than others and this frequency would have become a confounding factor in the results.

Secondly, we take into account the lessons learned from previous findings (Alberdi, 2008) and limit our analyses to correct user results, not correct action. Alberdi et al. (2008) found that 13.5% of recall decisions by radiologists and radiographers in their experiment were due to features other than those indicating actual cancer. For example, a radiologist who takes the correct action of recalling a patient who has cancer stating that there is a problem in the left breast when in fact the only mass the patient has is in the right breast. In essence, such an action would likely result in a biopsy on the wrong breast and the cancer would likely go undiscovered. The ultimate result for the patient is a false negative just as the ultimate result in our case is a misspelled word, and we decided it was important to take this extra level of detail suggested by the authors into account and judge participants' performance based not on action, but result.

4.4.3.2 Identifying User Groups

We define four different user groups. Both the mean and median score in the dictation test were 21.1. Participants who scored above this threshold were considered *high verbals* (20 participants (43%)) and those who scored at or below it were considered *low verbals* (27 participants (57%)). We also created two other user groups: very high verbals and very low verbals as we considered it important to look at these two extremes of the spectrum for trends in performance as done by Alberdi et al. (2010) and which resulted in the discovery of interesting results. We define *very high verbals* as those having a score above the mean+SD=26 (7 participants (15%)) and *very low verbals* as those having a score below the mean-SD=16 (5 participants (11%)).

4.4.3.3 Graphs of Mean User Performance

In this section, we present graphs of the mean performance of each of the four ability groups and in the next section we discuss the statistical significance of these results. Each graph presents the mean score of users in each of the four groups using the specific tool setting and the sensitive tool setting. We provide a graph for each type of word change: TPs, FPs, and Total Errors, where total errors is the total number of human FNs and FPs left in each passage after editing. Finally, we provide separate graphs for easy and difficult passages as our hypotheses suggest that we expect different trends based on passage difficulty.

We also restate the hypotheses presented in Section 4.2.5 here to ease comparisons.

1. When editing easy or difficult text, most **low verbals** will commit **fewer errors** (total sum of human FNs and human FPs) using the more **specific tool** setting compared to the more sensitive tool setting.
2. When editing easy text, most **high verbals** will commit **fewer errors** (total sum of human FNs and human FPs**)** using the more **sensitive tool** setting compared to the more specific tool setting.
3. When editing **difficult** text**,** most **high verbals** will commit **fewer errors** (total sum of human FNs and human FPs) using the **specific tool** compared to the more sensitive tool setting.

Graphs of Mean User Performance When Editing Easy Passages

We notice that for easy passages, the sensitive tool helped increase the performance (lower scores are better) of low verbals, high verbals, and very high verbals. However, the specific tool helped increase the performance of very low verbals. This serves as a manifestation of our general hypothesis that different user ability groups will perform better using different tool settings. Furthermore, all but the low verbals' behaviour matched the direction of our expectation, because as we recall from Section 4.2.5, we expected that when editing easy passages, high verbals will perform better using the sensitive tool setting and low verbals will perform better using the specific tool setting.

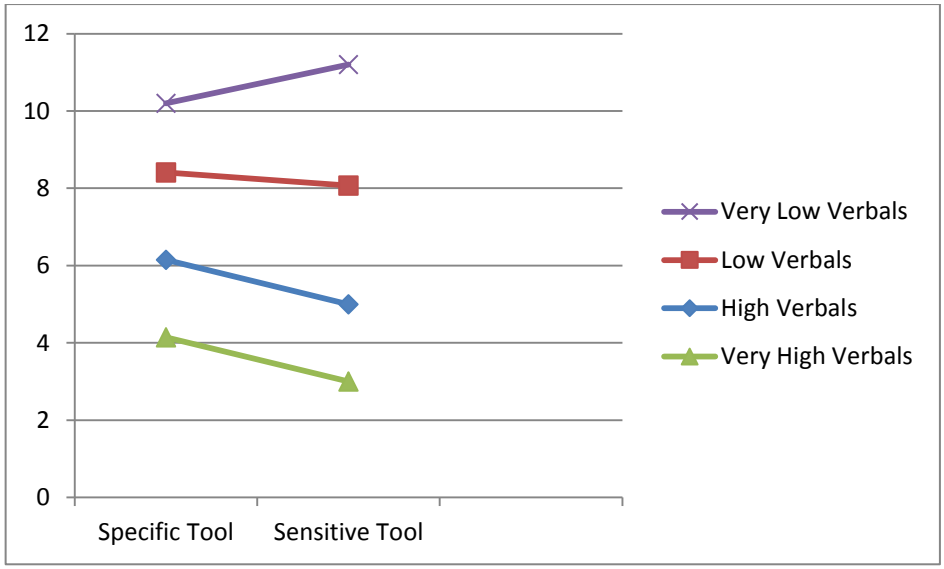


Figure 4-9: Average Number of Errors Left in Easy Passages
Each line represents the performance of a certain user group when using the specific versus sensitive tool. Overall, when editing easy passages, all user groups, except Very Low Verbals, made fewer total errors (FPs and FNs) when using the sensitive tool compared to when using the specific tool.

To analyse the reasons behind these differences more closely, we need to consider the individual effects of human FNs and FPs. We notice from Figure 4-10 that all groups except very low verbals were helped by the sensitive tool's increased number of TPs. However, low verbals actually had fewer TPs when using the sensitive tool. We can better understand this seemingly strange behaviour if we consider Table 4-3, which summarized our modelling, and where we predicted that a tool's increased TP rate on the most difficult words has a stronger positive effect on high verbals compared to low verbals. These are likely words which even when highlighted by the tool, a very low verbal has difficulty correcting because they do not know why it is incorrect, or how to correct it, or both (for example, the difference between the two homophones: discretely and discreetly).

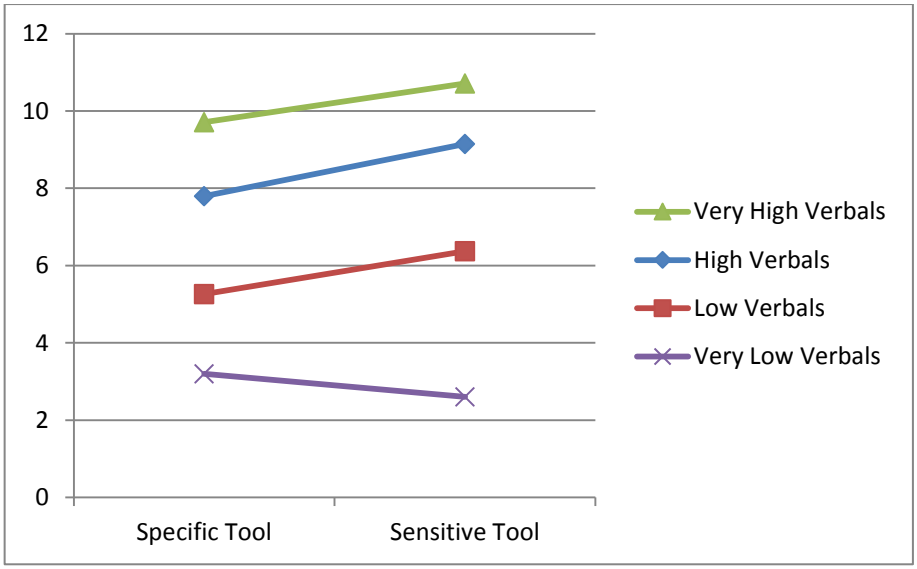


Figure 4-10: Average True Positives of User+Tool System in Easy Passages
Each line represents the performance of a certain user group when using the specific versus sensitive tool. Overall, when editing easy passages, all user groups, except Very Low Verbals, corrected more misspellings when using the sensitive tool compared to when using the specific tool.

The increased number of tool FPs had a negative effect on all users except the very high verbals. Again if we refer to Table 4-3, which summarizes our modelling, we note that increased tool FPs have a negative effect on users who do not know or are uncertain about the spelling of a word, but not those who know the spelling of a word. Thus, if we assume that only these group of very high verbals knew the correct spelling of some of the incorrectly marked words, then all other users would be negatively affected as we see in the graph.

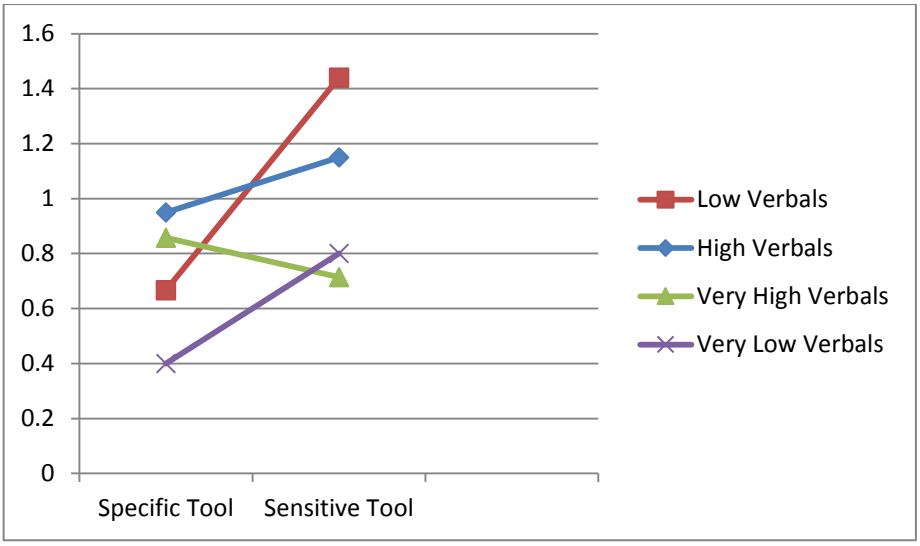


Figure 4-11: Average False Positives of User+Tool System in Easy Passages
Each line represents the performance of a certain user group when using the specific versus sensitive tool. Overall, when editing easy passages, all user groups, except Very High Verbals, changed more correctly spelled words when using the sensitive tool compared to when using the specific tool.

Thus, overall, high verbals were helped by the sensitive tool's increased TP rate more than they were negatively affected by its FP rate, probably because they were able to distinguish many of the tool's FPs and because they were able to recognize the correct spelling of difficult words correctly highlighted by the tool. However, very low verbals were negatively affected by the sensitive tool's increased FP rate and did not benefit from its highlighting of difficult errors; thus resulting in a better overall performance using the specific tool.

Graphs of Mean User Performance When Editing Difficult Passages

If we consider the same performance measures in the difficult passages, again we find that different user ability groups performed better using different tool settings, thus providing more evidence of our general hypothesis that different user ability groups will perform better using different tool settings. Moreover, both groups of high verbals which performed better using the sensitive tool when editing easy passages now performed better using the specific tool and the very low verbal group who performed better using the specific tool setting when editing easy passages now performed better using the sensitive tool setting. This switch in user behaviour is evidence for our second general hypothesis that the better tool setting for each user ability group will also depend on the task difficulty. The behaviour of both groups of high verbals matches our expectation, as they performed better using the specific tool when editing difficult passages. However, both groups of low verbals performed better using the sensitive tool, which is opposite to our expectation.

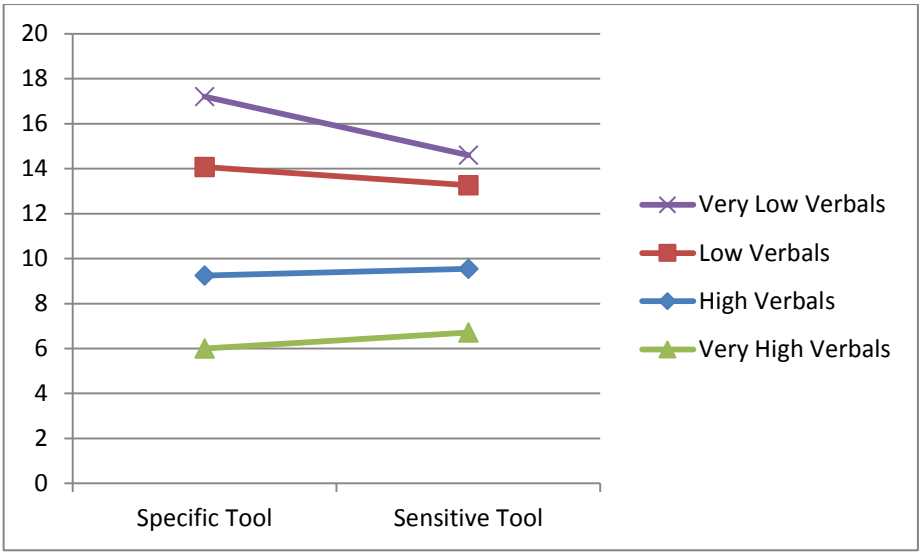


Figure 4-12: Average Number of Errors Left in Difficult Passages
Each line represents the performance of a certain user group when using the specific versus sensitive tool. Overall, when editing difficult passages, Low and Very Low Verbals made fewer total errors (FPs and FNs) when using the sensitive tool compared to when using the specific tool. High and Very High Verbals, instead, made fewer total errors when using the specific tool.

We look deeper at the causes of these overall trends in user behaviour when editing difficult text. We note that all users benefitted from the sensitive tool's increased TP rate although very high verbals showed almost no improvement probably because they were able to correct the errors even without the tool's support. We also note, as would be commonly expected and as predicted in our model, that benefit increased inversely according to user ability so that very low verbals showed the most benefit followed by low verbals, then high verbals and finally very high verbals.

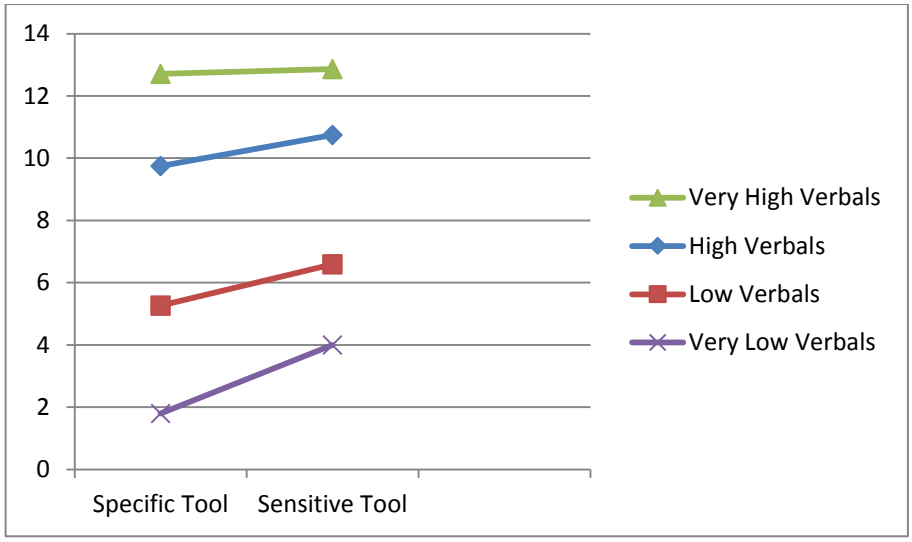


Figure 4-13: Average True Positives of User+Tool System in Difficult Passages
Each line represents the performance of a certain user group when using the specific versus sensitive tool. When editing difficult passages, all user groups corrected more misspellings when using the sensitive tool compared to when using the specific tool.

On the other hand, all participants were negatively affected by the sensitive tool's increased FP rate. Surprisingly high verbals were more negatively affected than low verbals. If we look again at Table 4-3 to help explain this, we see that this may be because these FPs represented words that high verbals were uncertain about and so were tricked by the tool's false markings, but low verbals did not know so left unchanged.

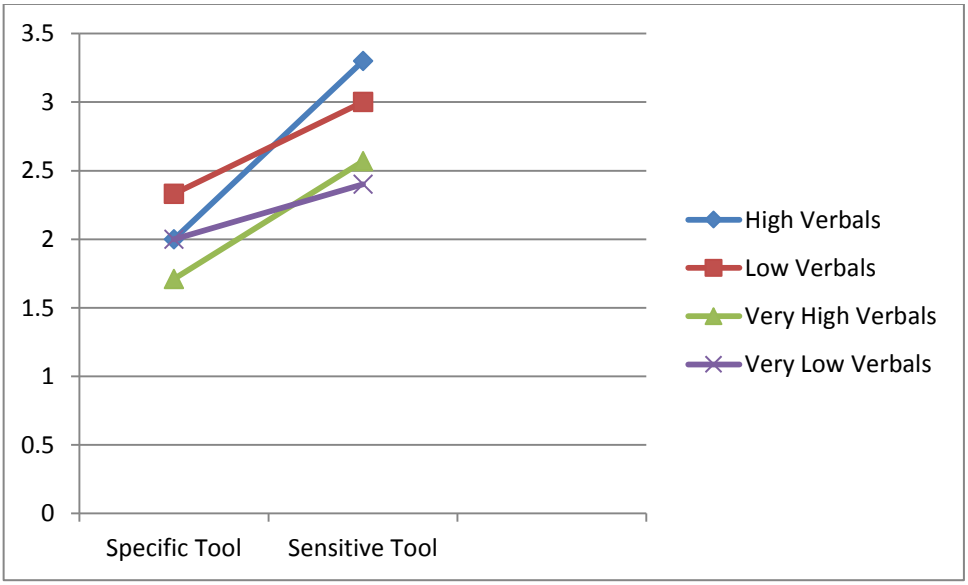


Figure 4-14: Average False Positives of User+Tool System in Difficult Passages
Each line represents the performance of a certain user group when using the specific versus sensitive tool. When editing difficult passages, all user groups changed more correctly spelled words when using the sensitive tool compared to when using the specific tool.

Overall, the results from both easy and difficult passages suggest that: (1) different user ability groups perform better using different tool settings based on their positive reaction to TPs, their negative reaction to FPs and the balance between the two, and that (2) this balance depends on the difficulty of the task. Moreover, in agreement with our more specific hypotheses: (1) high verbals perform better using the sensitive tool when editing easy passages, but (2) the specific tool when editing difficult passages while (3) very low verbals perform better using the specific tool when editing easy passages. Contrary to our expectations, both groups of low verbals perform better using the sensitive tool when editing difficult passages.

4.4.3.4 Significance of User Performance Differences

To test whether the differences presented in Section 4.4.3.3 were significant, we used Repeated Measures ANOVA (Wiredu, 2014). We divided analysis by difficulty (easy versus difficult passages) as they contain a different number of target words and so would probably exhibit significant differences in an ANOVA. We checked the normality assumption using the Shapiro Wilk Test and the sphericity assumption using Mauchly's Test of Sphericity. Where there was a between-subjects factor in our ANOVA, we also checked equality of covariance matrices using Box's Test (Wiredu, 2014). Where assumptions were violated, we used the non-parametric Mann Whitney U Test for independent groups and the Wilcoxon signed rank test for dependent groups (Wiredu, 2014).

Easy Passages

A repeated measures ANOVA determined that the mean number of remaining errors in the easy passages after editing (human FPs + human FNs) differed significantly between the two tool settings at the 90% confidence interval; $p=.064$, and differed significantly between high and low verbals; $p<0.0005$. However, the interaction of tool setting and user ability was non-significant; $p=0.301$.

We analysed this overall effect more closely by looking at the effect of each error type. A repeated measures ANOVA determined that the mean number of true positives in easy passages differed significantly between the two tool settings; $p=0.001$, and also between high and low verbals; $p<0.0005$. However, similar to the total number of errors, the interaction between tool setting and user ability was non-significant; $p=0.737$.

On the same token, we considered the number of human false positives. However, the distribution of false positives violated the assumptions of normality, even after removing all

outliers. Thus, instead of Repeated Measures ANOVA we used the non-parametric Mann-Whitney U Test to determine whether ability had a significant effect on the number of human false positives. Interestingly, the number of false positives in easy passages was not significantly affected by ability ($U=1060$, $p=.871$). We used the non-parametric Wilcoxon signed rank test to determine whether tool setting had a significant effect on the number of human false positives. The number of human false positives in easy passages was significantly affected by tool setting ($Z=-2.14$, $p=.032$). These results suggest that although tool setting had a significant effect on the number of human false positives, compared to low verbals, high verbals were not significantly more successful in distinguishing false tool prompts.

Difficult Passages

A repeated measures ANOVA determined that the mean number of remaining errors in the difficult passages after editing (human FPs + human FNs) did not differ significantly between the two tool settings at the 90% confidence interval; $p=.668$, but differed significantly between high and low verbals; $p<0.0005$. The interaction of tool setting and user ability was non-significant; $p=.355$.

We analysed this overall effect more closely by looking at the effect of each type of change. A repeated measures ANOVA determined that the mean number of true positives in easy passages differed significantly between the two tool settings; $p=0.05$, and also between high and low verbals; $p<0.0005$. However, similar to results from the easy passages, the interaction between tool setting and user ability was non-significant; $p=0.775$.

Similar to results from the easy passages, the distribution of false positives again violated the assumption of normality, even after removing all outliers. Thus, instead of Repeated Measures ANOVA we used the non-parametric Mann-Whitney U Test and Wilcoxon Signed-Rank Test to determine whether tool setting and/or ability had a significant effect on the number of human false positives. Interestingly, and similar to results in easy passages, the number of false positives in difficult passages was not significantly affected by ability according to the Mann Whitney U Test; $U=1069.5$, $p=.935$. However, it was affected by tool setting according to the Wilcoxon signed rank test; $Z=-2.95$, $p=0.003$.

4.4.3.5 Significance of User Performance Differences Between Tool Settings Depending on Passage Difficulty

To test whether user performance differs significantly between tool settings depending on passage difficulty, we need to compare the difference for each of our ability groups separately. The hypothesis would be significantly true if: (1) the total number of errors was significantly different between the two tool settings for easy passages, (2) between the two tool settings for difficult passages, *and* (3) the better tool setting was different between the two case difficulties.

For high verbals, the difference in the total number of errors in easy passages was significant at the 90% confidence interval; Z=-1.715, p=0.086, but the difference between the total number of errors in the difficult passages was non-significant; Z=-0.328, p=0.743. We note, however, that the trend was overall different as high verbals performed better using the sensitive tool in easy passages, but the specific tool in difficult passages.

For low verbals defined according to the dictation, the difference in the total number of errors in easy passages was non-significant; Z=-0.837, p=0.403, and the difference between the total number of errors in the difficult passages was non-significant; Z=-1.043, p=0.297. We note that the trend was not supported as low verbals performed better using the sensitive tool in both types of passages.

Thus, although observing changes in trends suggests that the optimal tool threshold for a given user ability group changes depending on case difficulty, these results are not significant. Possible reasons for this are further discussed in the next section.

4.4.3.6 Possible Reasons for Non-Significant Results

The results from the ANOVA tests suggest that the positive trends matching our hypotheses may be due merely to chance. We acknowledge this possibility, but we also discuss other possibilities for these non-significant results, which may prove useful when exploring future work.

Sample Size

There is a possibility that we may not have found significant differences due to the size of our sample. Our pre-experiment calculations presented in Section 4.3.5.1 indicated that a sample of 47 participants would have likely been enough to reveal differences between “Tool” and “No

Tool" settings as was the case in Galletta's experiment. However, differences due to varying tool settings are expected to be much smaller and thus a larger sample size, around 125 as suggested by the calculations in Section 4.3.5.1, may help to reveal such differences. Because we observed trends that match our hypotheses, we believe it is worth exploring them further by recruiting more participants in a future phase of the experiment.

Furthermore, our comparison of mean error types for the individual user groups using paired t-test support this need for a larger sample. We performed a paired t-test on the mean scores for each user ability group divided according to error type (total human errors, human TPs, and human FPs) and difficulty (easy and difficult). Thus, for each of the four user ability groups, we had results from six paired t-tests. We note that among the 12 comparisons in the very high verbal group and the very low verbal group, only one comparison was significant at the 90% confidence interval: the amount of TPs between the two tool settings in hard passages for very low verbals; $t(4)=2.56$, $p=0.06$. We note that these two extreme ability groups each contained a very small number of participants: 7 very high verbals and 5 very low verbals.

However, the larger groups: low verbals (27 participants) and high verbals (20 participants) showed more significant differences. At the 90% confidence interval, high verbals showed significant differences between the number of TPs between the tool settings for easy passages; $Z=-2.448$, $p=0.014$, the number of total human errors between the tool settings for easy passages; $t(19)=-1.71$, $p=0.104$, the number of false positives between the tool settings for hard passages; $t(19)=3.03$, $p=0.007$, and the number of false positives between the tool settings for easy passages $t(15)=2.334$, $p=.034$. While low verbals showed significant differences between the number of TPs between the tool settings for easy passages; $t(26)= 2.29$, $p=0.03$, the number of false positives between the tool settings for easy passages; $t(24)=2.092$, $p=0.047$, and the number of true positives between the tool settings for hard passages; $t(26)=1.72$, $p=0.098$. We note that we do not use these multiple comparisons to determine the significance of certain results, as this would likely require a correction, such as the Bonferroni *correction*, to account for multiple t-test comparisons. However, we use these comparisons only to suggest that this increased number of significant results in the larger groups may be due to the increased number of participants in them, and we use this to support our suggestion for the possible need of a larger sample size in order to observe significant overall differences.

User Reactions to False Prompts

We note from the analysis that high verbals, compared to low verbals, were significantly better at identifying true positives. However, results also showed that high verbals were not, as we

predicted, *significantly* better than low verbals at distinguishing tool false positives. This may be another contributing factor to the overall non-significant effect we obtained in the total number of human errors. We recall from Section 4.2.5, that our hypotheses that high verbals will perform better using one setting compared to a different setting for low verbals is built on two important blocks: that high verbals will benefit more than low verbals from the sensitive tool's additional tool TPs, but will be able to distinguish most of the sensitive tool's additional FPs. If this latter block is in fact untrue or not significantly true, then this may be another reason why the overall error results showed non-significant results.

Choice of Tool Setting

We note from the ANOVA tests that tool setting had a significant effect on the number of TPs and the number of FPs, but not on the overall number of errors. This leads us to consider whether our choice of tool settings resulted in smaller overall effects or effects that cancelled out. Recall that the hypotheses are built on the basis that the *balance* between the benefit from tool TPs and the degradation due to tool FPs will show opposite directions for different ability groups. We consider the possibility that the overall balance of errors could be one of the reasons that we did not see overall significant differences and that a different choice of tool settings may change this result.

As described in Section 4.3.6.2, we chose the two tool settings for the experiment using a rather subjective method. This could change if we choose a different application such as breast cancer screening and rely on a tool with pre-defined manufacturer tool settings. This could also change in the same spell checker context if we use data from this experiment to derive tool settings that could better address our hypotheses. In what follows we present a mathematical framework for how such a derivation might occur.

A general form of the equations used to write the number of human errors of a certain type (FN/FP) committed by a certain ability group (high/low verbals) is as shown in the example:

Number of FNs committed by a low verbal

= (Number of Type 1, incorrectly spelled and marked words) x P(FNLow | Type 1, incorrect, marked)

+ (Number of Type 1, incorrectly spelled and unmarked words) x P(FNLow | Type 1, incorrect, unmarked)

+ (Number of Type 2, incorrectly spelled and marked words) x P(FNLow | Type 2, incorrect, marked)

+ (Number of Type 2, incorrectly spelled and unmarked words) x P(FNLow | Type 2, incorrect, unmarked)

+ (Number of Type 3, incorrectly spelled and marked words) x P(FNLow | Type 3, incorrect, marked)

+ (Number of Type 3, incorrectly spelled and unmarked words) x P(FNLow | Type 3, incorrect, unmarked)

The number of correctly/incorrectly spelled words of type *x* marked/unmarked by the tool (red terms above) are a function of three variables:

- How many target words of type *x* the passage contains
  - In future experiments, we may set these using the same ratios defined in Section 4.3.6.4 for easy and difficult passages
- How many of these words are spelled incorrectly
  - In future experiments, we may determine these as before, by using the average number of times words of type *x* were spelled incorrectly by participants in Fischer's publication (1985)
- How many of these words are marked by the tool
  - This is basically the tool setting for type *x* words, and represents the unknown we are trying to solve for. In this experiment we chose this using a think-aloud process. But it would be desirable to choose the tool setting so that it is close to what we expect will give us a significant difference

The blue terms were, before running this experiment, completely unknown as they depend on users' behaviours and reactions according to word type and tool marking. These blue coloured values are hard to predict – they depend on factors such as probability of the person: noticing the word, knowing the word, trusting the tool, etc.

We suggest using data from our results to inform the blue value variables and solve for the tool settings that could better reveal some of our hypotheses. The unknowns we would need to solve for are listed in Table 4-13.

Table 4-13: Unknowns to Solve for			
Sensitive Tool (A)		**Specific Tool (B)**	
Sensitivity		Sensitivity	
	~~Type 1 Words~~		~~Type 1 Words~~
	Type 2 Words		Type 2 Words
	Type 3 Words		Type 3 Words
	Homophones		Homophones
Specificity		Specificity	
	~~Type 1 Words~~		~~Type 1 Words~~
	Type 2 Words		Type 2 Words
	Type 3 Words		Type 3 Words
	Homophones		Homophones

Note that we crossed out Type 1 words to simplify the problem and assumed for simplicity that these words would be available in both dictionaries so that for both settings, sensitivity = specificity = 1.

To solve for these unknowns, we can write a number of inequalities, which are based on our hypotheses and where Setting A and B correspond to the sensitive and specific tool settings, respectively:

1. High verbals and easy task:
   FP(Setting A)-FP(Setting B) < FN(Setting B) – FN(Setting A)
2. Low verbals and easy task:
   FP(Setting A)-FP(Setting B) > FN(Setting B) – FN(Setting A)
3. High verbals and difficult task:
   FP(Setting A)-FP(Setting B) > FN(Setting B) – FN(Setting A)
   a. Different than equation 1 because of the frequencies of these words and the different behaviour of users under these conditions (blue values)
4. Sensitivity (Homophone, A) > Sensitivity (Homophone, B)
5. Specificity (Homophone, A) > Specificity (Homophone, B)
6. Specificity (Type 3, A) < Specificity (Type 3, B)
7. Specificity (Type 2, A) < Specificity (Type 2, B)
8. Sensitivity (Type 3, A) >= Sensitivity (Type 3, B)
   a. All words spelled wrong are marked by both tools unless the misspelling results in another valid word not included in the smaller dictionary setting. For example, “guilt” spelled incorrectly may result in “gilt”, which may only exist in the larger dictionary.
9. Sensitivity (Type 2, A) >= Sensitivity (Type 2, B)
10. FPs in Setting B by high verbals < FPs in Setting B by low verbals

Mathematical constraints:

- Sensitivity and Specificity must be between 0 and 1 to make sense.

As is true with any mathematical definition of a real-world problem, we have made a number of simplifications/assumptions:

- Blue valued probabilities are the same for both tool settings which ignores user reactions due to issues of trust or the ignoring of excessive prompts
- Blue valued probabilities do not change with time as the user learns more about the tool

However, despite these assumptions, we believe that for future experiments which do not rely on tools with pre-set thresholds, such a mathematical approach for choosing appropriate tool settings may help reveal significant results in case they exist. Moreover, with increased resources, using more than two tool settings is also likely to increase chances of finding significant differences.

Definition of High and Low Verbals

An important basis for our analysis is our definition of high and low verbals. Users were divided into these two categories based on their performance in the dictation test. As discussed in our design of the experiment in Section 4.3.3, we chose the words in this dictation test to closely match the orthographic representation and difficulty levels of words in the editing tasks. However, we note from the literature that users may sometimes fail to correct an error because they neglect to use their knowledge to *detect* the error, not because they do not have *knowledge* of its correct spelling (Figueredo, 2004). If many of our participants' errors in the tasks were due to such errors and not due to errors in spelling knowledge, then the assessment task may not accurately represent ability. Another way to look at this issue is that in Section 4.2.4, we presented a model of how high and low verbals will react to words and tool prompts. However, classifying users into these two groups according to the dictation test may not represent the most accurate division of users according to the behaviours and divisions we expected in our model.

An alternative method we use to classify users into high and low verbals is according to their performance in editing the four passages. We summed the number of errors (human FPs + human FNs) in each of the four passages to compute what we define as an "Editing Score" for each participant. We compare this score to users' "Dictation Score" in the scatterplot in Figure 4-15.

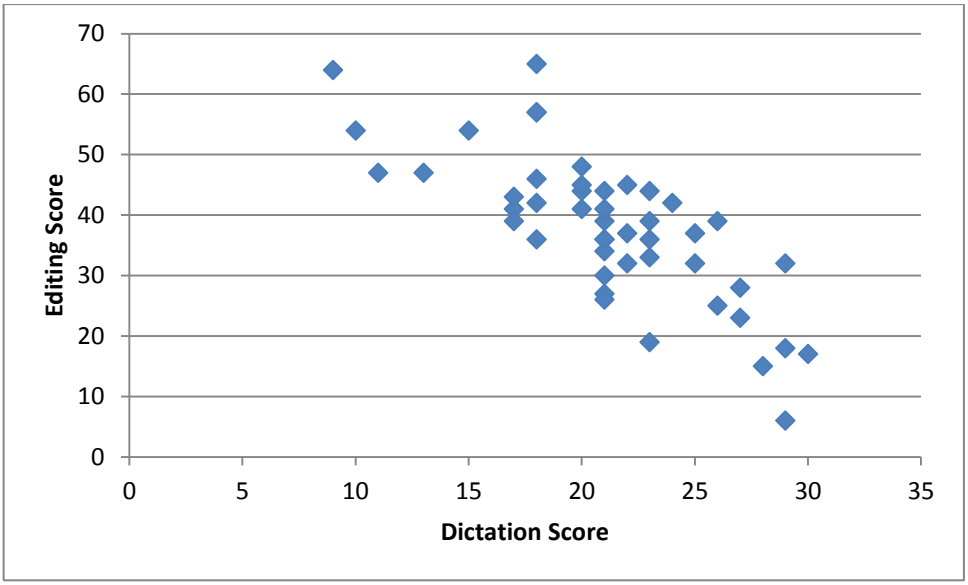


Figure 4-15: Relationship Between Participants' Performance in Dictation Versus Editing Tasks

As expected, we note a negative relationship between the two scores (higher editing scores translate to more errors left in the passages and thus poorer performance). Pearson's correlation between the two variables is -0.768, $p<0.0005$; and although this suggests the two scores are highly related, they are not identical and so we consider in what follows, an analysis of our hypotheses based on a new aposteriori definition of high and low verbals based on their performance in editing the passages rather than in the dictation. High verbals are now defined as those achieving editing scores lower than the mean editing score, and low verbals as those achieving scores greater than the mean editing score.

Most importantly, High and Low verbals defined according to the dictation do not show significant differences in terms of their ability to distinguish tool false positives; $U=269$; $p=0.983$. However, when defined according to the editing tasks, high verbals were significantly better at avoiding false positive errors; $U=158$; $p=0.012$. This is an important behaviour because we use this specific trait to distinguish between the behaviour of high and low verbals in our modelling where we argue that high verbals will be better at distinguishing false tool prompts.

Now, a repeated measures ANOVA determined that the mean number of remaining errors in the easy passages after editing (human FPs + human FNs) differed significantly between the two tool settings at the 95% confidence interval; $p=.042$, and differed significantly between

high and low verbals; p<0.0005. Furthermore, the interaction of tool setting and user ability was also significant; p=.004.

This result is very important as it shows using a different definition of high and low verbals provided strong statistical evidence for our hypothesis. This is depicted in the graph below where, as we predicted in our hypothesis, when editing easy passages, high verbals performed significantly better using the sensitive tool and low verbals using the specific tool.

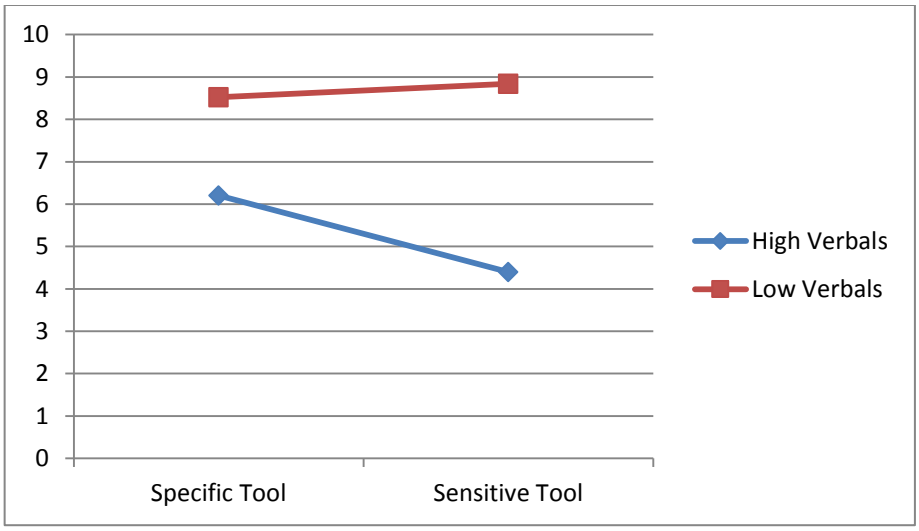


Figure 4-16: Average Number of Errors Left in Easy Passages Using Aposteriori Classification of User Ability

However, a Repeated Measures ANOVA on the difficult passages showed no significant differences in the mean number of remaining errors after editing (human FPs + human FNs) according to tool setting (p=0.578), a significant difference according to ability (p<0.0005), and a non-significant difference in the interaction of tool setting and user ability (p=0.859).

4.4.4 Results About User Perception

In the responses to the post-experiment questionnaire, the number of users who preferred the specific tool setting (19 participants) exactly matched the number who preferred the sensitive tool setting, while 9 participants expressed that they did not notice a difference between the two settings. This may suggest no strong user preferences for a certain tool. However, a deeper analysis of the results based on user ability suggests otherwise. 67% of high verbals preferred the sensitive tool while 61% of low verbals preferred the specific tool. This majority vote in each category matches our expectation that high verbals would probably prefer the tool setting that

highlights difficult mistakes even if it meant an increased number of FPs, while low verbals would probably prefer the tool that makes fewer FPs, perhaps because the additional TPs the sensitive tool highlights are not recognizable to them. These results are visually depicted in Figure 4-17.

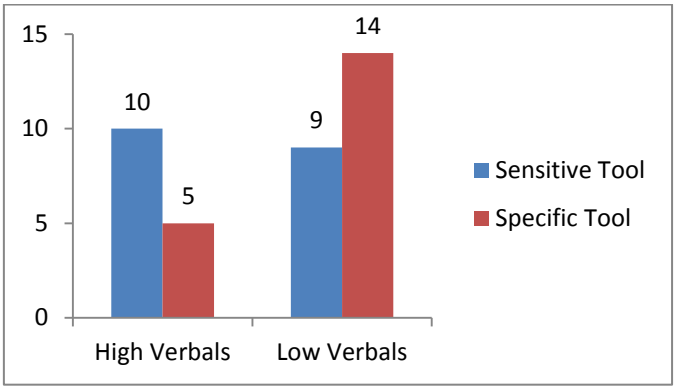


Figure 4-17: Self-Reported Tool Preference According to User Ability

When participants were asked at the end of the study to distinguish which tool made more false negatives and which made more false positives, 66% of users identified the correct tool.

As the distribution of user perceptions of the two tools was not normal, we used the Wilcoxon Signed-Rank Test to compare user perceptions of the two tools. The Wilcoxon Signed-Rank Test revealed no significant differences between users' perceived reliability of the sensitive tool and specific tool; Z=-1.12, p=.263. Wilcoxon Signed-Rank Test also revealed no significant differences between users' trust in the sensitive tool and specific tool; Z=-.86, p=.39. Wilcoxon Signed-Rank Test revealed no significant differences between users' annoyance from false positives by the sensitive tool and specific tool; Z=-.836, p=.403.

At the 95% confidence interval, these differences remain non-significant even when analysed according to ability, except for high verbals who had a significantly higher perceived reliability in the sensitive tool compared to the specific tool; Z=-1.81, p=.07 and a significantly higher trust in the sensitive tool compared to the specific tool; Z=-2.07, p=0.038. This increased trust and perceived reliability may be because the tool successfully highlighted misuses of difficult homophones. This increased trust may also explain why high verbals did not show, compared to low verbals, a significantly higher ability in distinguishing false tool prompts.

Finally, we note an important difference between overall user trust in spell checkers and user trust in the particular spell checkers that were presented with in the experiment. Kendall's tau-b, a suitable correlation for ordinal variables with many ties (Wiredu, 2014), showed a moderately

negative, but significant correlation between overall trust and trust in the sensitive spell checker (Kendall's tau-b=-0.304, p=0.036). Many studies in the literature fail to make a distinction between overall trust in DSSs, overall trust in a certain genre of DSSs (for example, all spell checkers) and trust in the specific DSS used in the experiment. As these results show, these values are not always equal, and this is one reason that may help explain some of the discrepancies in the literature on the effects of trust on automation use.

4.5 Considerations for Future Work

In this section, we present a range of ideas that build on the experiment we conducted. Some of these ideas aim to better explore the significance of the results (Section 4.5.1), to enrich the analysis of the existing data (Section 4.5.2), or to try different techniques (Section 4.5.3) and contexts (Section 4.5.4) to predict the tool threshold that could best improve user performance. Finally, in Section 4.5.5, we list several ideas that could be used to give deeper insight into the causes of certain human errors within the spell checker context.

4.5.1 Ideas to Better Explore the Significance of the Results

We previously discussed possibilities that may explain why some of the positive trends we observed were non-significant. These reasons represent a good starting point for future studies. Firstly, to increase the sample size. Secondly, to use a more mathematical approach to determine appropriate tool settings. Thirdly, to consider different ways of defining high and low verbals. These ideas were presented in detail in Section 4.4.3.6.

4.5.2 Future Analyses

Future analyses of the existing data may benefit from taking into account the effect of the tool markings on errors. Currently, all human false positives are grouped into a single category. However, these can be classified into false positives marked by the tool and false positives not marked by the tool. A similar classification can be applied to false negatives to enrich the results by separating out possible causes of the errors.

Future analyses may also take into account human errors based on word type (Level 1, 2, 3, or homophones) and how this compares to users' performance in the corresponding levels in the dictation assessment. This could give insight into benefits of diversity between tool and user abilities as was discussed in depth in Chapter 3.

4.5.3 Other Techniques to Predict Suitable Tool Thresholds

In this experiment, we relied on simple modelling to predict the tool threshold that would result in the best system performance for each group of similarly skilled users. A different approach would be to run a pilot study, observe user behaviour and performance according to different indicators and to use this to predict the suitable threshold for a new set of users. The use of pilot observations rather than modelling to predict, among a subset of tool settings for a certain tool, the tool setting that would be most appropriate for each user could be an interesting track for future studies.

4.5.4 New Contexts

We suggest our hypotheses may benefit from explorations in new contexts or the use of different decision-support tools. Introducing the experiment in a non-critical domain such as spell checking represented a chance to test the hypotheses using minimal resources, but may have given us, as researchers, almost unlimited freedom in the choice of the tool setting. Perhaps working with an already existing tool with tool settings pre-defined by tool designers could help inform designers in their design choices and managers in their use of an existing tool.

4.5.5 Ideas to Help Deduce the Causes of System Errors

In this section, ideas that could give us deeper insight into the causes of certain errors that users may commit in the experiment are discussed. Knowing the reason why a participant did not change an incorrectly spelled word or changed a correctly spelled word would be beneficial in a number of ways:

- We may find that varying the tool setting helped improve the performance only of certain errors
    - For example, an error due to vigilance versus knowledge, or familiar versus unfamiliar words
- Different causes translate to real-world contexts such as reading mammograms in a breast-cancer screening program where there are detection versus classification errors
- We may be able to quantify sources of human errors
    - In the comprehensive review on automation bias presented by Parasuraman et al. (2010), authors mentioned that “it is necessary to decompose the underlying determinants of commission errors” and to date, there is little empirical evidence to allow for an assessment of the different causes of such errors.

4.5.5.1 Questionnaires or Ethnographic Methods to Deduce Causes of Errors

Although it is beneficial to better understand the causes of system errors, this is not necessarily an easy task. Using Fitzgerald et al.'s (1987) list of methods (questionnaires, interviews, and think-aloud records) that provide further insight on users' decisions and thoughts, combined with Gray's (2012) finding that participants were interested in knowing their score after completing the spelling correction tasks, participants may be asked at the end of a spell-checking experiment whether they would like to know how they did, and then asked about their errors, or a subset of their errors. For example:

- Why did you leave [insert human FN] uncorrected?
    - I didn't see it
    - I was unsure if it was incorrectly spelled
    - I was unfamiliar with its correct spelling
    - It wasn't marked by the tool
    - Other _____
- Why did you change [insert human FP]?
    - I was unsure about its spelling
    - I didn't know this word
    - I believe it's spelled incorrectly
    - I trusted the tool's prompt
    - Other _____

A few factors to consider with such a questionnaire style are:

- There is an increase in the effort of participants (but with the incentive of finding out their score)
- Participants' second look will likely affect their answers, especially that Jacoby et al. (1990) showed that "a single reading of a word can influence its later spelling even when the word is not recognized as one that was read earlier".
- The questions and choices of answers need to be carefully worded and thought out to ensure that they do in fact enrich the analysis and present all possible mutually exclusive options.

It's also possible to gain insight on the causes of some errors using ethnographic research methods. Such methods were previously used by Alberdi et al. (2014) and Hartswood et al. (2003) to yield interesting perspectives. These methods may include set-ups that allow for detailed observations, interviews, or think-aloud protocols. One idea which we believe would yield interesting insights in terms of explanations suggesting overreliance and complacency is to provide participants with

access to a dictionary and to observe which words they look up and the frequency with which they use the dictionary.

4.5.5.2 Eye-Tracking to Deduce Causes of Error

Another way that results could be enriched to provide insight into causes of human errors is by using an eye tracker to examine which words people looked at and for how long. This could help researchers:

- Better understand how people edit a document (word for word, or starting with marked words, etc.)
- Discover behavioural differences between how high and low verbals edit a document and which may contribute to differences in system performance.
    - For example, Dreiseitl et al. (2005) used eye-tracking to observe dermatologists of varying experiences diagnose a set of lesion images. Although authors did not find significant performance differences in the experience groups in terms of sensitivity and specificity, the expert group exhibited visibly lower ratings for all eye measures. They spent a disproportionately lower time in fixations, took a more holistic approach, and did not scan for the presence or absence of features as much as the less experienced radiologists (Dreiseitl, 2005).
- Determine whether a human false negative was due to detection versus classification (i.e., the user did not notice the word versus the user noticed the word but decided it was not spelled incorrectly suggesting they discounted certain information but did not ignore it)
- Discover differences between how people look at marked versus unmarked words
- Discover differences in how people edit text with and without a checker

Interestingly, eye tracking analysis is based on the important assumption that there is a relationship between fixations, user gaze and what the user is thinking about (Wickens, 2005; Juhasz, 2003). However, sometimes fixations do not necessarily translate into a conscious cognitive process (Reichle, 2010). For example, during a search task, a user may fixate on a target but still miss its presence. Sometimes this happens because user expectations rather than users' actual experiences modulate what they see. Researchers refer to this as "inattentional blindness" and using only eye-tracking, this cause for errors would be difficult to distinguish from the active discounting of information (Memmert, 2006; Parasuraman, 2010).

4.6 Contribution to Existing Knowledge and Overall Discussion

According to our knowledge of the literature, this is the first experiment to explore how system performance can be enhanced by calibrating the sensitivity and specificity of a tool for different levels of user skill and of case difficulty. This idea builds on suggestions in the literature that adaptive automation and tools that are individualized to take into account user differences can enhance system performance. This idea also builds on the analysis of the literature for common trends of tool prompt effects on users of varying ability and cases of varying difficulty.

One of the closest studies resembling our work is that by Zheng et al. (2001) who recruited radiologists who interpreted the same mammograms without CAD and then with CAD under five separate tool settings to observe effects of overall system performance. However, all cases were difficult, and unlike our study, effects of user skill and case difficulty on the appropriateness of the tool threshold were not considered.

We found evidence of our general hypotheses that groups of similarly skilled users perform better at different tool settings, and that the same group of users perform better using a different tool setting in difficult versus easy tasks depending on their positive reaction to tool TPs, their negative reaction to tool FPs and the balance between the two. Furthermore, in agreement with our simple modelling of user behaviour:

- **high verbals** (defined as participants who scored above the average score in the experiment) performed better using the **sensitive tool** when editing **easy passages**, but the **specific tool** when editing **difficult passages**
- **very low verbals** (participants who scored one standard deviation below average in the experiment) performed better using the **specific tool** when editing **easy passages**.

Contrary to our expectations, **low verbals** (participants who scored below the average score in the experiment) performed better using the **sensitive tool** when editing **difficult passages**.

Secondary results showed significant differences between users' general trust in spell checkers and their trust of the spell checker in the experiment, and this is one reason that may explain some of the discrepancies in the literature on the effects of trust on automation use. Secondary analyses also revealed that high verbals were significantly better than low verbals in correcting spelling mistakes, but not in recognising false tool prompts. Among many possibilities, this may be because tool false prompts existed on words about which they were initially uncertain, so that they were easily swayed by the tool's prompts. Another possibility is that the tool's successful highlighting of difficult mistakes may have resulted in a level of trust that caused high verbals to accept some of the false tool prompts.

Interestingly, the majority of high verbals preferred working with the sensitive tool while the majority of low verbals preferred working with the specific tool. This majority vote in each category matches our expectation that high verbals would likely prefer the tool setting that highlights difficult mistakes even if it meant an increased number of FPs, while low verbals would probably prefer the tool that makes fewer FPs, perhaps because they lack the knowledge to decide how to respond to the additional TPs the sensitive tool highlights.

The experiment resulted in trends which generally match our expectations, but some of these results revealed non-significant differences when analysed using ANOVA. Besides the possibility that the conjectured effect is absent, we also discussed several other possibilities for this non-significance including sample size, choice of tool setting, and the definition of high and low verbals.

Another important contribution of this experiment is that it provides a useful basis for future, similarly designed studies. Experience with the AABST Tool, with the generation of passages, discussion of design issues, and results of ability indicators can be used to feed future studies in automation bias at the Centre for Software Reliability. Most importantly, as discussed in Section 4.4.3.6, numerical results of the experiment can be used to inform experimental parameters when designing future studies, especially when choosing appropriate tool settings. Moreover, the use of a spell checking tool to study effects of automation bias has only been previously reported in one other published study, and we hope that by introducing the experiment in such a non-critical domain that we may encourage future studies in automation bias using minimal resources.

Future work may also benefit from ideas presented in Section 4.6 including ideas to: (1) better explore the significance of the results by increasing sample size and altering tool settings, (2) enrich the analysis by taking into account tool markings and orthographic word difficulty, (3) explore other prediction techniques besides modelling such as pilot observations, (4) run the experiment in a different context using tools with pre-defined tool settings (to help users effectively choose amongst tool settings already pre-set by manufacturers), and (5) provide insight into the causes of certain human errors using post-study questionnaires, eye tracking, and/or ethnographic observations.

We believe that the concept of varying tool settings does not require infeasible tool designs or tool designs that will not be seen except in the far future, because at least one tool operating at several tool settings has already been released into the market (The Smart Choice 2). We have examined the questions of whether this option could in fact result in better system performance and how to best choose the appropriate setting for each user ability groups and case difficulty level, but these questions still remain open to future explorations.

Conclusion

The phrase "*Automation bias*" refers – as used in this thesis - to decision errors made with the tool that are not made without the tool. Many times, automation bias can result in performance (in terms of correct decisions) which is actually worse with the tool than without it. From our literature review in Chapter 1, we noted that although there is abundant literature on automation bias mediators and some literature on mitigation strategies, there is a lack of research on the cognitive causal explanations of automation bias and on adaptive mitigation strategies that result in tools that adapt to the needs and characteristics of individual users. The goal of this thesis was to help reduce the rate of automation bias through a better understanding of its causes and the proposal of innovative adaptive mitigation strategies via the design, analysis, and discussion of three experiments presented in Chapters 2 to 4.

In Chapter 2 we presented an experiment in glaucoma treatment which aimed to, for the first time in the literature:

(a) Provide evidence that the mere availability of advice can alter user behaviour, even before advice is presented
(b) Provide evidence of the homeostasis mechanism - users altering the decision threshold that they apply before they see tool advice - as a possible cause of some automation bias errors
(c) Explore how users' cognitive processes vary depending on when CDSS advice is presented
(d) Evaluate the effect of the glaucoma risk calculator on the management decisions of ophthalmologists and optometrists, when provided with realistic and informative patient cases

Such an understanding of causal mechanisms and effects of tool advice are addressed to help designers, managers and medical professionals design more effective AB mitigation strategies, establish more effective protocols for the use of automated tools and assess the benefit of the glaucoma risk calculator in the management of glaucoma suspects, respectively.

This experiment provided evidence of: (1) a rise in users' decision threshold when they know they will receive advice and before presentation of this advice, and which may account for some automation bias errors previously observed but unexplained in the literature, (2) the effect of the presence of a decision support tool on user decisions, even before its presentation, and (3) differences between self-reported trust and perceived tool reliability and differences between objective measures of difficulty defined in terms of variance and subjective measures of difficulty defined in terms of

uncertainty. Study results showed no significant differences in user decisions depending on the sequence of presentation of advice, even when analyses incorporated error type. This latter result, although different from our expectation, may add to the evidence that users are not being overreliant or complacent and thus encourage more explorations of alternative mechanisms, such as the homeostasis mechanism.

From a medical viewpoint, the study provided evidence of: (1) strong disagreement with NICE guidelines even among UK professionals and even after use of the calculator, (2) a significant effect of the calculator on user decisions, increased user confidence, and increased intra-user consistency in decisions, (3) differences in treatment recommendations depending on where a participant practices, and (4) lack of consideration for patient age and PSD in clinicians' treatment decisions. Results showed no significant differences in treatment recommendations based on specialty or profession.

In Chapter 3, we presented a retrospective study of the benefits of diversity in double-reading pairs in a breast cancer screening context. Simulations of these pairs aimed to, for the first time in the literature:

(a) Provide evidence of the benefits of creating diverse reader pairs compared to random pairs matched based on convenience
(b) Suggest ways to create effective complementary double-reading pairs

This work provided a new perspective on double reading by asking not how opinions should be combined but which readers should be combined with whom. Results showed that creating groups of complementary ability levels can significantly improve sensitivity, with an insignificant effect on specificity, compared to random pairings. These results could be beneficial to hospital managers in improving the performance of their screening programs. How best to pair an entire pool of readers is a difficult question and poses many possibilities. We explored some of these possibilities using several different methods. Pairing according to sensitivity or discriminating ability seemed to maintain a reasonable balance between the sensitivity and specificity of the overall pair. Outside our exploration of the initial hypotheses, our analysis of the data interestingly revealed that experience, measured in years of practice, was not an accurate measure of user ability.

The work presented in Chapter 3, presented empirical and mathematical evidence of the benefits of diversity in improving the performance of human + human systems. Similar to results from the judge-advisory system literature looking at human + human systems, and which have been successfully integrated into the automation bias literature (Goddard, 2012; Madhavan, 2007), we argue that similar benefits from diversity can be obtained by applying such ideas to decision-support systems. Our work showed the benefits of mechanically combining two diverse components, but

future work may take this a step further to consider how interactions between the two system components may affect these benefits, especially when the second component is an automated tool rather than human, and to show that increased diversity between users and the tool can improve system performance even when this increase comes at the expense of decreased tool performance.

In Chapter 4 we presented an experiment, using a spell checking tool, which exposed participants to two different tool settings (sensitivity / specificity combinations), and aimed to, for the first time in the literature:

(a) Provide evidence that each group of similarly skilled users will perform best using a different tool setting
(b) Provide evidence that for a single user ability group, the tool setting that results in better system performance for difficult cases may be different from the tool setting that results in better system performance for easy cases
(c) Predict, among a subset of tool settings for a certain tool, the tool setting that would be most appropriate for each user group, via modelling

This adaptive mitigation strategy should help create tools that adapt to individuals' needs and characteristics to improve system performance and may motivate designers and managers to implement tools with variable and easily adjustable settings, and to predict, based on modelling and/or pilot studies, the appropriate setting, among a subset of tool settings for a given tool, the setting that is most appropriate for each group of similarly skilled users and cases of similar difficulty level.

This experiment supported our general hypotheses that groups of similarly skilled users perform best at different tool settings, and that the same group of users perform best using a different tool setting in difficult versus easy tasks depending on their positive reaction to tool TPs, their negative reaction to tool FPs and the balance between the two. Furthermore, in agreement with our simple modelling of user behaviour, (1) high verbals (defined as participants who scored above the average score in the experiment) performed better using the sensitive tool when editing easy passages, but (2) performed better using the specific tool when editing difficult passages; while (3) very low verbals (participants who scored worse than one standard deviation below average in the experiment) editing easy passages performed best using the specific tool. Contrary to our expectations, low verbals (participants who scored below the average score in the experiment) editing difficult passages performed best using the sensitive tool. Outside our exploration of the initial hypotheses, our analysis of the data interestingly, and contrary to expectations, revealed that high verbals, as defined based on their performance in the assessment task, were not significantly better than low verbals in distinguishing tool false prompts.

Although we list, in this conclusion, a number of contributions of this work, we summarize the *major* contributions of this research in a single sentence for each chapter. This thesis successfully adds to the literature on automation bias:

- evidence of the mere presence of tool advice, even before its presentation, as well as evidence that supports the existence of a newly introduced cognitive mechanism to explain some automation-induced errors previously unexplained in the literature
- empirical evidence of the benefits of applying diversity specifically to double-reading pairs in a breast cancer screening programme, while considering possible rules for forming these pairs
- benefits of a mitigation strategy by which system performance is enhanced by calibrating the sensitivity and specificity of a tool for a range of user skills and case difficulty, and exemplifies how modelling of user behaviour can be used to predict, among a subset of tool settings for a certain tool, the tool setting that would be most appropriate for each user ability group and class of case difficulty

Future Work

We present here a range of research ideas that build on the experiments we conducted. Some of these ideas aim to better explore the significance of the results, to enrich the analysis of the existing data, to provide deeper insights, or to try different techniques and contexts. More details explaining these ideas can also be found in Sections 3.6 and 4.5.

Future work extending the **glaucoma** experiment should, firstly, explore the repeatability of the results using other decision support tools to check evidence of the homeostasis mechanism, the effect of the presence of tool advice, and the non-significant effect of changing the sequence of advice. In future experiments, experimenters could make the homeostasis mechanism more evident by: (1) enforcing a protocol that does not allow users to change their pre-advice decision from "target" to "no target", (2) decreasing the system's acceptable alarm rate, (3) manipulating participant incentives, or (4) manipulating the dataset to include more borderline cases (close to the user's decision threshold). Future research would also benefit from a study of the relative importance of the different explanations of AB, including the homeostasis mechanism, and especially in relation to different contexts and mediators. Lastly, from a medical perspective, some of our findings contradict those of previous studies by Boland et al. (2008), thus suggesting that the exact effect of the glaucoma risk calculator remains undetermined and would benefit from further studies.

Future work extending the **double-reading** simulations would benefit, most importantly, from taking the first step of showing the benefits of mechanically combining two diverse components, and

extending them to consider, experimentally, how interactions between two system components may affect these benefits, especially when the second component is an automated tool rather than human. Most interesting would be to show that increased diversity between users and the tool can improve system performance even when this increase comes at the expense of decreased tool performance.

From a medical perspective, future work in the double-reading context should:

- Study the effects of pairing readers according to complementary ability in a *clinical* setting
- Consider other measures of experience beyond number of years in practice
- Explore whether the second reader should be blinded from the first reader's decision, and how much information should be shared – the recall/no recall decision, or a complete highlighting of suspicious areas on the mammogram
    - Explore, in the case that decisions are not blinded, who should read first and whether the second reader's knowledge that they are in fact second significantly affects the reader's decision or vigilance
- Consider creating pairs not according to performance measures such as sensitivity and specificity, but according to mammogram-reading specific abilities. For example, according to who is more competent at identifying microcalcifications, masses, or asymmetries
- Simulate pairs' decisions according to alternate methods to the mechanical OR rule such as discussion or arbitration

Future work extending the **spell checker** experiment should:

- Better explore the significance of the results by increasing sample size and altering tool settings
- Enrich the analysis by taking into account which words were marked by the tool as well as orthographic difficulty of the words
- Explore, for the purpose of deciding the tool settings for different users, other prediction techniques besides modelling, such as pilot observations
- Implement the experiment in a different context, especially using tools with pre-defined tool settings (to help users effectively choose amongst tool settings already pre-set by manufacturers)
- Provide insight into the causes of certain human errors using post-study questionnaires, eye tracking, and/or ethnographic observations

In conclusion, the results in this thesis benefit tool designers, managers setting protocols for the use of decision support tools, healthcare providers, and also future researchers. We hope that this work adds to the understanding of automation bias and is a step forward in showing that automation bias is a

recurring phenomenon in various contexts, that its explanations must include in some cases a normative and rational adaptation of users to automated tools, and that using knowledge of user differences to improve the design of automated tools and to tailor the way information is presented to individual users could prove effective in improving system performance.

References

Alberdi, Eugenio, et al. "Does incorrect computer prompting affect human decision making? A case study in mammography." *International Congress Series*. Vol. 1256. Elsevier, 2003.

Alberdi, Eugenio, et al. "Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography." *Academic radiology* 11.8 (2004): 909-918.

Alberdi, E., et al. "Automation bias and system design: a case study in a medical application." *People and Systems-Who Are We Designing For, 2005. The IEE and MOD HFI DTC Symposium on (Ref. No. 2005/11078)*. IET, 2005.

Alberdi, Eugenio, et al. "CAD in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions." *International Journal of Computer Assisted Radiology and Surgery* 3.1-2 (2008): 115-122.

Alberdi, Eugenio, et al. "Why are people's decisions sometimes worse with computer support?." *Computer Safety, Reliability, and Security*. Springer Berlin Heidelberg, 2009. 18-31.

Alberdi, E., et al. "CAD: risks and benefits for radiologists' decisions." *The handbook of medical image perception and techniques*. By Ehsan Samei et al. Cambridge University Press, 2010.

Alberdi, E., et al. "Use of computer-aided detection (CAD) tools in screening mammography: a multidisciplinary investigation." (2014).

Azar, Beth. "Danger of automation: It makes us complacent." *APA monitor* 29.7 (1998): 3.

Bahner, J. Elin, Anke-Dorothea Hüper, and Dietrich Manzey. "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience." *International Journal of Human-Computer Studies* 66.9 (2008): 688-699.

Balleyguier, Corinne, et al. "Computer-aided detection (CAD) in mammography: does it help the junior or the senior radiologist?." *European journal of radiology* 54.1 (2005): 90-96.

Beam, C. A., Sullivan, D. C., & Layde, P. M. (1996). Effect of human variability on independent double reading in screening mammography. Academic radiology, 3(11), 891-897.

Benish, W. A. "The use of information graphs to evaluate and compare diagnostic tests." *Methods of information in medicine* 41.2 (2002): 114-118.

Berner, Eta S., et al. "Clinician performance and prominence of diagnoses displayed by a clinical diagnostic decision support system." *AMIA Annual Symposium Proceedings*. Vol. 2003. American Medical Informatics Association, 2003.

Bonaccio, Silvia, and Reeshad S. Dalal. "Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences." *Organizational Behavior and Human Decision Processes* 101.2 (2006): 127-151.

Boland, Michael V., Harry A. Quigley, and Harold P. Lehmann. "The impact of risk calculation on treatment recommendations made by glaucoma specialists in cases of ocular hypertension." *Journal of glaucoma* 17.8 (2008): 631-638.

Boland, Michael V., and Harold P. Lehmann. "A new method for determining physician decision thresholds using empiric, uncertain recommendations." *BMC medical informatics and decision making* 10.1 (2010): 20.

Boland, Michael V., Harry A. Quigley, and Harold P. Lehmann. "The impact of physician subspecialty training, risk calculation, and patient age on treatment recommendations in ocular hypertension." *American journal of ophthalmology* 152.4 (2011): 638-645.

Burr, J. M., et al. "Risk prediction tools for development of open-angle glaucoma." (2012).

Cheng, Eric M., et al. "For which glaucoma suspects is it appropriate to initiate treatment?." *Ophthalmology* 116.4 (2009): 710-6.

Cornford, E. J., et al. "The pathological and radiological features of screen-detected breast cancers diagnosed following arbitration of discordant double reading opinions." *Clinical radiology* 60.11 (2005): 1182-1187.

Croskerry, Pat. "Achieving quality in clinical decision making: cognitive strategies and detection of bias." *Academic Emergency Medicine* 9.11 (2002): 1184-1204.

Cuddeback, David, et al. "Towards overcoming human analyst fallibility in the requirements tracing process (NIER track)." *Proceedings of the 33rd International Conference on Software Engineering*. ACM, 2011.

Cummings, M. L. "Automation bias in intelligent time critical decision support systems." *AIAA 1st Intelligent Systems Technical Conference*. Vol. 2. 2004.

Detmer, Don E., Dennis G. Fryback, and Kevin Gassner. "Heuristics and biases in medical decision-making." *Academic Medicine* 53.8 (1978): 682-3.

de Vries, Peter, Cees Midden, and Don Bouwhuis. "The effects of errors on system trust, self-confidence, and the allocation of control in route planning." *International Journal of Human-Computer Studies* 58.6 (2003): 719-735.

Diagnosis and management of chronic open angle glaucoma and ocular hypertension. Manchester, UK: National Institute for Health and Clinical Excellence, 2009.

Dinnes, J., Moss, S., Melia, J., Blanks, R., Song, F., Kleijnen, J. (2001). Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. The Breast, 10, 455-463.

Dreiseitl, Stephan, and Michael Binder. "Do physicians value decision support? A look at the effect of decision support systems on physician opinion." *Artificial intelligence in medicine* 33.1 (2005): 25-30.

Dreiseitl, Stephan, Maja Pivec, and Michael Binder. "Differences in examination characteristics of pigmented skin lesions: Results of an eye tracking study." *Artificial intelligence in medicine* 54.3 (2012): 201-205.

Dzindolet, Mary T., et al. "Misuse and disuse of automated aids." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 43. No. 3. SAGE Publications, 1999.

Elmore, J. G., & Brenner, R. J. (2007). The more eyes, the better to see? from double to quadruple reading of screening mammograms. Journal of the National Cancer Institute, 99(15), 1141-1143.

Food and Drug Administration Center for Devices and Radiological Health Radiological Devices Advisory Panel Meeting. *3 February 2004,* Gaithersburg Marriott Washingtonian Center.

Figueredo, Lauren, and Connie K. Varnhagen. "Detecting a problem is half the battle: The relation between error type and spelling performance." *Scientific Studies of Reading* 8.4 (2004): 337-356.

Figueredo, Lauren, and Connie K. Varnhagen. "Spelling and grammar checkers: are they intrusive?." *British journal of educational technology* 37.5 (2006): 721-732.

Fingeret, Murray. "Risk calculators can help guide treatment decisions for ocular hypertension."

Fischer, F. William, Donald Shankweiler, and Isabelle Y. Liberman. "Spelling proficiency and sensitivity to word structure." *Journal of memory and language* 24.4 (1985): 423-441.

Friedlander, Myrna L., and Susan D. Phillips. "Preventing anchoring errors in clinical judgment." *Journal of consulting and clinical psychology* 52.3 (1984): 366.

Fitzgerald, Jill. "Research on revision in writing." *Review of educational research* 57.4 (1987): 481-506.

For the Screening, Prognosis, Diagnosis, Management and Prevention of Glaucoma 2010. Canberra, Australia: National Health and Medical Research Council, 2010.

Friedman, Charles P., et al. "Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems." *Jama* 282.19 (1999): 1851-1856.

Gadala, Marwa, Strigini, Lorenzo, Povyakalo, Andrey, Ayton, Peter. "Identifying and Discussing Uncertain Cases in Breast Cancer Screening Programs" *The 15th Biennial European Meeting of the Society for Medical Decision Making.* Smdm, 2014.

Gadala, Marwa, Strigini, Lorenzo, Povyakalo, Andrey, Ayton, Peter. "Double Reading of Mammograms: Effectively Pairing Readers with Diverse Skills to Improve Performance" *The 35th Annual North American Meeting of the Society for Medical Decision Making.* Smdm, 2013.

Galletta, Dennis F., et al. "Does spell-checking software need a warning label?." *Communications of the ACM* 48.7 (2005): 82-86.

Gilbert FJ, Astley SM, Gillan MG et al. Single reading with computer-aided detection for screening mammography. N Engl J Med, 2008, 359:1675–1684.
Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. "Automation bias-a hidden issue for clinical decision support system use." *Studies in health technology and informatics* 164 (2010): 17-22.

Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. "Automation bias: a systematic review of frequency, effect mediators, and mitigators." *Journal of the American Medical Informatics Association* 19.1 (2012a): 121-127.

Goddard, Kate (2012b). Automation bias and prescribing decision support – rates, mediators and mitigators. (Unpublished Doctoral thesis, City University London)
Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. "Automation bias: Empirical results assessing influencing factors." *International journal of medical informatics* 83.5 (2014): 368-375.

Goni, F. "The glaucoma risk calculator: practical remarks." *ARCHIVOS-SOCIEDAD ESPANOLA DE OFTALMOLOGIA* 82.1 (2007): 7.

Gray, Cristina (2012). Development of a pilot experiment to test automation bias. (Unpublished Masters thesis, City University London)

Halligan, Steve, et al. "Incremental benefit of computer-aided detection when used as a second and concurrent reader of CT colonographic data: multiobserver study." *Radiology* 258.2 (2011): 469-476.

Hartswood, Mark, et al. "Subjective responses to prompting in screening mammography." *MIUA-96* (1996).

Hartswood, Mark, et al. "Drawing the line between perception and interpretation in computer-aided mammography." *Proceedings of the 1st International Conference on Allocation of FunctionsInternational Ergonomics Association Press, Louisville*. 1997.

Hartswood, Mark, Rob Procter, and Linda J. Williams. "Prompting in practice: How can we ensure radiologists make best use of computer-aided detection systems in screening mammography?." *Digital Mammography*. Springer Netherlands, 1998. 363-370.

Hartswood, M. Procter, R. Institute for Communicating and Collaborating Systems. *Computer-Aided Mammography: A Case Study of Error Management in a Skilled Decision-Making Task*. Scotland, 1999. Web. <http://www.dcs.gla.ac.uk/~johnson/papers/HECS_99/Hartswood_Procter.htm>.

Hartswood, Mark, et al. "'Repairing'the Machine: A Case Study of the Evaluation of Computer-Aided Detection Tools in Breast Screening." *ECSCW 2003*. Springer Netherlands, 2003.

Harvey, Susan C., et al. "Increase in cancer detection and recall rates with independent double interpretation of screening mammography." *American Journal of Roentgenology* 180.5 (2003): 1461-1467.

Hillman, hessel, et al. Improving Diagnostic Accuracy: A comparison of interactive and delphi consultations. 1976.

Ho, Geoffrey, Dana Wheatley, and Charles T. Scialfa. "Age differences in trust and reliance of a medication management system." *Interacting with Computers* 17.6 (2005): 690-710.

Jacoby, Larry L., and Ann Hollingshead. "Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words." *Canadian Journal of Psychology/Revue canadienne de psychologie* 44.3 (1990): 345.

Juhasz, Barbara J., and Keith Rayner. "Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29.6 (2003): 1312.

Karmel, Miriam. "Glaucoma: Calculating the Risk." *American Academy of Ophthalmology*, n.d. Web. 01 Jan. 2015.

Kass, Michael A., et al. "The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma." *Archives of ophthalmology* 120.6 (2002): 701-713.

Kass, Michael A., et al. "Delaying treatment of ocular hypertension: the ocular hypertension treatment study." *Archives of ophthalmology* 128.3 (2010): 276.

Kawamoto, Kensaku, et al. "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success." *Bmj* 330.7494 (2005): 765.

Kim, Yong-Seok. "Effects of Driver, Vehicle, and Environment Characteristics on Collision Warning System Design." (2001).

Khoo, Lisanne AL, Paul Taylor, and Rosalind M. Given-Wilson. "Computer-aided Detection in the United Kingdom National Breast Screening Programme: Prospective Study 1." *Radiology* 237.2 (2005): 444-449.

Kobayashi, Takeshi, et al. "Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs." *Radiology* 199.3 (1996): 843-848.

Lawrence, Edward M., et al. "Colorectal Polyps: Stand-alone Performance of Computer-aided Detection in a Large Asymptomatic Screening Population 1." *Radiology* 256.3 (2010): 791-798.

Liston, J. C., and B. J. G. Dall. "Can the NHS Breast Screening Programme afford not to double read screening mammograms?." *Clinical radiology* 58.6 (2003): 474-477.

Lyell, David, and Enrico Coiera. "Automation bias and verification complexity: a systematic review." Journal of the American Medical Informatics Association (2016): ocw105.

Madhavan, Poornima, and Douglas A. Wiegmann. "Effects of Information Source, Pedigree, and Reliability on Operators' Utilizaton of Diagnostic Advice." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 49. No. 3. SAGE Publications, 2005.

Madhavan, Poornima, and Douglas A. Wiegmann. "Effects of information source, pedigree, and reliability on operator interaction with decision support systems." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49.5 (2007): 773-785.

Mansberger, Steven L. "A risk calculator to determine the probability of glaucoma." *Journal of glaucoma* 13.4 (2004): 345-347.

Mansberger, Steven L., and George A. Cioffi. "The probability of glaucoma from ocular hypertension determined by ophthalmologists in comparison to a risk calculator." *Journal of glaucoma* 15.5 (2006): 426-431.

Mansberger, Steve L., Felipe A. Medeiros, and Mae Gordon. "Diagnostic tools for calculation of glaucoma risk." *Survey of ophthalmology* 53.6 (2008): S11-S16.

Marten, Katharina, et al. "Computer-assisted detection of pulmonary nodules: performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologists." *European radiology* 14.10 (2004): 1930-1938.

Matcham, N. J., et al. "Breast screening: the use of consensus opinion for all recalls." *The Breast* 13.3 (2004): 184-187.

Manzey, Dietrich, Nina Gérard, and Rebecca Wiczorek. "Decision-making and response strategies in interaction with alarms: the impact of alarm reliability, availability of alarm validity information and workload." *Ergonomics* 57.12 (2014): 1833-1855.

McBride, Sara E., Wendy A. Rogers, and Arthur D. Fisk. "Understanding human management of automation errors." Theoretical issues in ergonomics science 15.6 (2014): 545-

McKinlay, Andy, Rob Procter, and Anne Dunnett. "An investigation of social loafing and social compensation in computer-supported cooperative work." *Proceedings of the international ACM SIGGROUP conference on Supporting group work*. ACM, 1999.

Medeiros, Felipe A., and Robert N. Weinreb. "Estimating the risk of developing glaucoma." *The open ophthalmology journal* 3 (2009): 50.

Memmert, Daniel. "The effects of eye movements, age, and expertise on inattentional blindness." *Consciousness and cognition* 15.3 (2006): 620-627.

Meyer, Joachim. "Effects of warning validity and proximity on responses to warnings." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 43.4 (2001): 563-572.

Mithal, Sandeep, et al. "Role of Star Calculator in Ocular Hypertensives."

Moray, Neville, Toshiyuki Inagaki, and Makoto Itoh. "Adaptive automation, trust, and self-confidence in fault management of time-critical tasks." *Journal of Experimental Psychology: Applied* 6.1 (2000): 44.

Mosier, Kathleen L., et al. "Automation bias: Decision making and performance in high-tech cockpits." *The International journal of aviation psychology* 8.1 (1998): 47-63.

Mosier, Kathleen L., et al. "Aircrews and automation bias: the advantages of teamwork?." *The International Journal of Aviation Psychology* 11.1 (2001): 1-14.

Mucci, B., G. Athey, and G. Scarisbrick. "Double reading of screening mammograms: the use of a third reader to arbitrate on disagreements." *The Breast* 8.2 (1999): 63-65.

Nenkov, G. Y., Morrin, M., Ward, A., Schwartz, B., & Hulland, J. (2008). A short form of the Maximization Scale: Factor structure, reliability and validity studies. Judgment and Decision Making, 3(5), 371-388.

Ocular Hypertension Treatment Study Group, and European Glaucoma Prevention Study Group. "A validated prediction model for the development of primary open angle glaucoma in individuals with ocular hypertension." *Ophthalmology* 114.1 (2007): 10.

Official Report: Grounding of Cruise Ship (Report No. 7042OR). London, UK: The Nautical Institute, 2012. Accessed January 2015. Web.
< http://www.nautinst.org/en/forums/mars/mars-2012.cfm/cruiseshipgrounding>

Oron-Gilad, T., et al. "Incorporating individual differences into the adaptive automation paradigm." *Human factors in organizational design and management VIII* (2005): 581-586.

Parasuraman, Raja. "Memory load and event rate control sensitivity decrements in sustained attention." *Science* 205.4409 (1979): 924-927.

Parasuraman, R., P. A. Hancock, and O. Olofinboba. "Alarm effectiveness in driver-centred collision-warning systems." *Ergonomics* 40.3 (1997a): 390-399.

Parasuraman, R., Riley,V. (1997b) Humans and automation: use, misuse, disuse, abuse. *Human Factors*, 39, 230–253.

Parasuraman, Raja, and Dietrich H. Manzey. "Complacency and bias in human use of automation: An attentional integration." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52.3 (2010): 381-410.

Parasuraman, Raja, et al. "Dopamine beta hydroxylase genotype identifies individuals less susceptible to bias in computer-assisted decision making." *PloS one* 7.6 (2012): e39675.

Petrick, Nicholas, et al. "CT Colonography with Computer-aided Detection as a Second Reader: Observer Performance Study 1." *Radiology* 246.1 (2008): 148-156.

Philpotts, Liane E. "Can Computer-aided Detection Be Detrimental to Mammographic Interpretation? 1." *Radiology* 253.1 (2009): 17-22.

Pitman, Alexander Gregory, et al. "Intrareader variability in mammographic diagnostic and perceptual performance amongst experienced radiologists in Australia." *Journal of medical imaging and radiation oncology* 55.3 (2011): 245-251.

Popov, Peter, et al. "Software diversity as a measure for reducing development risk." Dependable Computing Conference (EDCC), 2014 Tenth European. IEEE, 2014.

Povyakalo, Andrey A., et al. "How to Discriminate between Computer-Aided and Computer-Hindered Decisions A Case Study in Mammography." *Medical Decision Making* 33.1 (2013): 98-107.

Primary Open-Angle Glaucoma Suspect Summary Benchmarks for Preferred Practice Pattern Guidelines. American Academy of Ophthalmology, October 2013.

Purcell, Gretchen P. "What makes a good clinical decision support system: We have some answers, but implementing good decision support is still hard." *BMJ: British Medical Journal* 330.7494 (2005): 740.

Quality Assurance Guidelines for Breast Cancer Screening Radiology. Sheffield, UK: NHS Cancer Screening Programmes, 2011.
Quekel, L. G. B. A., et al. "Detection of lung cancer on the chest radiograph: impact of previous films, clinical information, double reading, and dual reading." *Journal of clinical epidemiology* 54.11 (2001): 1146-1150.

Ramnarayan, Padmanabhan, et al. "Diagnostic omission errors in acute paediatric practice: impact of a reminder system on decision-making." *BMC medical informatics and decision making* 6.1 (2006): 1.

Reichenbach, Juliane, Linda Onnasch, and Dietrich Manzey. "Human performance consequences of automated decision aids in states of sleep loss." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53.6 (2011): 717-728.

Reichle, Erik D., Andrew E. Reineberg, and Jonathan W. Schooler. "Eye movements during mindless reading." *Psychological Science* 21.9 (2010): 1300-1310.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77.

Roehrig, Jimmy, and Ronald A. Castellino. "The promise of computer aided detection in digital mammography." *European journal of radiology* 31.1 (1999): 35-39.

Salako, Kizito, and Lorenzo Strigini. "When does" Diversity" in Development Reduce Common Failures? Insights from Probabilistic Modelling." (2014): 1-1.

Schwartz, Barry, et al. "Maximizing versus satisficing: happiness is a matter of choice." *Journal of personality and social psychology* 83.5 (2002): 1178.

Schwartz, Barry. "The tyranny of choice." *SCIENTIFIC AMERICAN-AMERICAN EDITION*- 290.4 (2004): 70-75.

Shapiro, Sam, et al. "Breast cancer screening programmes in 22 countries: current policies, administration and guidelines." *International Journal of epidemiology* 27.5 (1998): 735-742.

Skitka, Linda J., Kathleen L. Mosier, and Mark Burdick. "Does automation bias decision-making?." *International Journal of Human-Computer Studies* 51.5 (1999): 991-1006.

Skitka, Linda J., Kathleen Mosier, and Mark D. Burdick. "Accountability and automation bias." *International Journal of Human-Computer Studies* 52.4 (2000): 701-717.

Skitka, Linda J., et al. "Automation bias and errors: Are crews better than individuals?." *The International journal of aviation psychology* 10.1 (2000): 85-97.

Smith-Bindman, R., Chu, P. W., Miglioretti, D. L., Sickles, E. A., Blanks, R., Ballard-Barbash, R., & Kerlikowske, K. (2003). Comparison of screening mammography in the United States and the United Kingdom. JAMA: the journal of the American Medical Association, 290(16), 2129-2137.

Sniezek, Janet A., and Timothy Buckley. "Cueing and cognitive conflict in judge-advisor decision making." *Organizational Behavior and Human Decision Processes* 62.2 (1995a): 159-174.
Sorkin, Robert D., and David D. Woods. "Systems with human monitors: A signal detection analysis." *Human-Computer Interaction* 1.1 (1985): 49-75.

Southern, William N., and Julia Hope Arnsten. "The effect of erroneous computer interpretation of ECGs on resident decision making." *Medical Decision Making* (2009).

Stewart, William C., et al. "Cost-effectiveness of treating ocular hypertension." *Ophthalmology* 115.1 (2008): 94-98.

Strigini, Lorenzo, Andrey Povyakalo, and Eugenio Alberdi. "Human-machine diversity in the use of computerised advisory systems: a case study." *43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE Computer Society, 2003.

Swensson, Richard G., Samuel J. Hessel, and Peter G. Herman. "The value of searching films without specific preconceptions." *Investigative Radiology* 20.1 (1985): 100-107.

Swets, John A. "The science of choosing the right decision threshold in high-stakes diagnostics." *American Psychologist* 47.4 (1992): 522.

Takwoingi, Yemisi, et al. "External validation of the OHTS-EGPS model for predicting the 5-year risk of open-angle glaucoma in ocular hypertensives." *British Journal of Ophthalmology* 98.3 (2014): 309-314.

Taylor, P. "Computer aids for detection and diagnosis in mammography." *Imaging* 14.6 (2002): 472-477.

Taylor, Paul, Richard Lee, and Eugenio Alberdi. "Incorporating Explanations into a CAD tool for the classification of calcifications." *Digital Mammography*. Springer Berlin Heidelberg, 2003. 372-376.

Taylor, P. M., et al. "An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms." (2004).

Taylor P, Champness J, Given-Wilson R, Johnston K, Potts H. Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography. Health Technol Assess 2005; 9 (6).

Taylor, Stuart A., et al. "CT Colonography and Computer-aided Detection: Effect of False-Positive Results on Reader Specificity and Reading Efficiency in a Low-Prevalence Screening Population 1." *Radiology* 247.1 (2008a): 133-140.

Taylor, Stuart A., et al. "CT Colonography: Investigation of the Optimum Reader Paradigm by Using Computer-aided Detection Software 1." *Radiology* 246.2 (2008b): 463-471.

Taylor, Paul, and Henry WW Potts. "Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate." *European Journal of Cancer* 44.6 (2008c): 798-807.

Taylor, P., et al. "Evaluation of CAD with Full Field Digital Mammography in the NHS Breast Screening Programme." (2009).
The Smart Choice is R2 Digital CAD. USA: Hologic Incorporation, 2007. Print.

Thurfjell, Erik, et al. "Sensitivity and specificity of computer-assisted breast cancer detection in mammography screening." *Acta Radiologica* 39.4 (1998): 384-388.

Tsai, Theodore L., Douglas B. Fridsma, and Guido Gatti. "Computer decision support as a source of interpretation error: the case of electrocardiograms." *Journal of the American Medical Informatics Association* 10.5 (2003): 478-483.

Tseng, Shawn, and B. J. Fogg. "Credibility and computing technology." *Communications of the ACM* 42.5 (1999): 39-44.

Van Dongen, Kees, and Peter-Paul Van Maanen. "A framework for explaining reliance on decision aids." *International Journal of Human-Computer Studies* 71.4 (2013): 410-424.

Verheyen, Steven, and Gert Storms. "A Mixture Approach to Vagueness and Ambiguity." *PloS one* 8.5 (2013): e63507.

Walsham, Anna C., et al. "The use of computer-aided detection for the assessment of pulmonary arterial filling defects at computed tomographic angiography." *Journal of computer assisted tomography* 32.6 (2008): 913-918.

Wang, Lu, Greg A. Jamieson, and Justin G. Hollands. "Selecting methods for the analysis of reliance on automation." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 52. No. 4. SAGE Publications, 2008.

Weinreb, Robert N., et al. "Risk assessment in the management of patients with ocular hypertension." *American journal of ophthalmology* 138.3 (2004): 458-467.

Wickens, Christopher, et al. *Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis*. ILLINOIS UNIV AT URBANA SAVOY, 2005.

Wickens, Christopher D., and Stephen R. Dixon. "The benefits of imperfect diagnostic automation: A synthesis of the literature." *Theoretical Issues in Ergonomics Science* 8.3 (2007): 201-212.

Wiegmann, Douglas A. "Agreeing with automated diagnostic aids: A study of users' concurrence strategies." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 44.1 (2002): 44-50.

Wiredu, E. (2014). *Introduction to IBM SPSS with Statistics for Healthcare Professionals and Academics.* Lancashire: Data Solutions Services.

Wivell, G., et al. "Can radiographers read screening mammograms?." *Clinical radiology* 58.1 (2003): 63-67.

Woloshin, Steven, Lisa M. Schwartz, and Andrew Ellner. "Making sense of risk information on the web." *BMJ* 327.7417 (2003): 695-696.

Youden W. J., "Index for rating diagnostic tests.", Cancer 1950; 3: 32-35

Zheng, Bin, et al. "Soft-Copy Mammographic Readings with Different Computer-assisted Detection Cuing Environments: Preliminary Findings 1." *Radiology* 221.3 (2001): 633-640.

Zheng, Bin, et al. "Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments" *Academic radiology* 11.4 (2004): 398-406.

Appendices

Appendix 2A: Email Invitation to Participate in Glaucoma Experiment and Participant Information Sheet

Hello,

Researchers at City University London are running a web-based study exploring factors that affect healthcare providers' management decisions of ocular hypertensive (OHT) patients.

The study is an online questionnaire which asks you for your treatment decision of 20 OHT patients, and is expected to take approximately 30 minutes of your time.

If you have the opportunity to kindly complete the questionnaire, we will be able to send you a personalized report containing a large subset of the cases you viewed together with: (1) your management decision, (2) an anonymized distribution of other participants' decisions, and (3) the decision of an 11-member expert panel nominated by international professional ophthalmologic and optometric societies for their clinical expertise and professional reputation. We hope that this information will be beneficial to participants who kindly share their time and expertise to inform this research.

To participate, simply read the participant information at the end of this email, and then click the survey link below. However, **please only click the survey link when you have a convenient 30 minute time slot to complete the questionnaire in a single sitting because although you can log out at any time, you cannot log in again**. Please also log in from a **desktop/laptop, not a phone** to ensure all images are properly visible.

Survey link: https://cityunilondon.eu.qualtrics.com/SE/?SID=SV_3DbsSvqWZv0d11H

Many thanks for taking the time to read this information. We sincerely hope you are able to participate and we welcome any questions or comments you may have.

Kind Regards,

Mrs. Marwa Gadala
marwa.gadala.1@city.ac.uk

Important Factors in the Management Decision of Ocular Hypertensive Patients

Introduction

Researchers at City University London would like to invite practicing ophthalmologists and optometrists to take part in a multi-disciplinary research study exploring healthcare providers' management decisions of ocular hypertensive (OHT) patients. Our aim is that this study will shed light on important factors that shape these often difficult decisions, and your participation will be valuable in contributing to this knowledge.

Before you decide whether to take part in this study, it is important that you understand why the research is being done and what it would involve for you. Please take the time to read the

following information carefully and to discuss it with others if you wish. Also, please ask us if there is anything that is unclear or if you would like more information.
The study is an online questionnaire which asks you for your treatment decision of 20 OHT patients, and is expected to take approximately 30 minutes of your time.

Procedure

Participation is voluntary and you can withdraw from the study at any time. It is up to you to decide whether or not to take part. If you do decide to take part, you will be asked to complete a consent form. If you decide to take part you will still be free to withdraw at any time and without giving a reason.

Please do not click the survey link until you have a convenient 30 minute time slot to complete the questionnaire in a single sitting because although you can log out at any time, you cannot log in again. Please also log in from a **desktop/laptop, not a phone** to ensure all images are properly visible.

In each case, you will be provided with certain patient information which may include age, intraocular pressure, central corneal thickness, cup-to-disc ratio, pattern standard deviation, family history, disc size, life expectancy, or the patient's risk of developing primary open angle glaucoma within the next five years calculated using a glaucoma risk estimator. Then you will be asked to tick the management option that best represents your decision for that patient.

Benefit to Participants

Within a few months of completing the study, you will be sent a report containing a large subset of the cases you viewed together with your treatment decision, as well as an anonymized distribution of other participants' decisions and the decision of an expert panel on these cases, and we hope that this information will be beneficial to you.

Research Ethics

All data stored will be anonymous and confidential and will be used solely for the purpose of this study. Results may form sections of submitted papers to peer reviewed journals. Participants will be contacted only to provide feedback on the study results. After that, all email addresses will be destroyed.

This study has been reviewed and given a favourable opinion by the City University London Research Ethics Committee.

If you have any questions or comments about the research, please contact Mrs. Marwa Gadala at marwa.gadala.1@city.ac.uk.

If you would like to complain about any aspect of the study, City University London has established a complaints procedure via the Secretary to the Senate Research Ethics Committee. To complain about the study, phone 020 7040 3040. You can then ask to speak to the Secretary of the Senate Research Ethics Committee and inform them that the name of the project is: Important Factors in the Treatment Decision of Ocular Hypertensive Patients.

You could also write to the Secretary at:

Anna Ramberg
Secretary to Senate Research Ethics Committee
CRIDO
City University London
Northampton Square
London
EC1V 0HB
Email: Anna.Ramberg.1@city.ac.uk

Thank you for taking the time to read this information.

Appendix 2B: Consent Form for Glaucoma Experiment

Consent Form

Title of Project:
Important Factors in the Management Decision of Ocular Hypertensive Patients

Researcher:
Mrs. Marwa Gadala, City University London, marwa.gadala.1@city.ac.uk

1. I confirm that I have read the participant information (in the email) for the above study, and that I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.

2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without being penalized or disadvantaged in any way.

3. I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports, any publications, or to any other party.

☐ **I have read and understand the participant information and consent pages and agree to take part in this study.**

Appendix 2C: Maximization Scale Survey

I would describe my agreement with the following statements as:

	Strongly Agree	Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Disagree	Strongly Disagree
When I am in the car listening to the radio, I often check other stations to see if something better is playing, even if I am relatively satisfied with what I'm listening to.	○	○	○	○	○	○	○
No matter how satisfied I am with my job, it's only right for me to be on the lookout for better opportunities.	○	○	○	○	○	○	○
I often find it difficult to shop for a gift for a friend.	○	○	○	○	○	○	○
Renting/Choosing videos is really difficult. I'm always struggling to pick the best one.	○	○	○	○	○	○	○
No matter what I do, I have the highest standards for myself.	○	○	○	○	○	○	○
I never settle for second best.	○	○	○	○	○	○	○

Appendix 2D: Cases Used in the Glaucoma Experiment

Table 3-13: Patient Cases Used in the Glaucoma Experiment											
Case ID*	**Age** (years)	**IOP** (mm Hg)	**CDR**	**CCT** (μm)	**PSD**	**Disc Size**	**Family History**	**Life Expectancy****	**Calculator's Risk**	**Expert Panel's Decision**	**NICE Guidelines' Suggestion**
1	85	25	0.6	500	1.5	large	negative	abnormal	50%	No treat	No treat
1A	80	24	0.7	500	1.4	large	negative	abnormal	50%	No treat	No treat
2	85	22	0.8	500	1.5	large	negative	normal	52%	No treat	No treat
2A	80	22	0.9	496	1.4	large	Negative	normal	52%	No treat	No treat
3	50	22	0.6	600	1.5	medium	positive	normal	4%	No treat	No treat
3A	55	23	0.5	604	1.6	medium	positive	normal	4%	No treat	No treat
4	50	25	0.3	600	1.5	small	positive	normal	3%	No treat	No treat
4A	45	24	0.4	596	1.4	small	positive	normal	3%	No treat	No treat
5	50	28	0.6	500	2.3	large	negative	normal	52%	Treat	Treat
5A	55	29	0.5	504	2.4	large	negative	normal	52%	Treat	Treat
6	85	28	0.6	500	2.3	small	positive	normal	77%	Treat	No treat
6A	81	27	0.7	500	2.2	small	positive	normal	77%	Treat	No treat
7	62	30	0.8	600	2.3	medium	positive	normal	24%	Treat	No treat
7A	67	31	0.7	604	2.4	medium	positive	normal	24%	Treat	No treat
8	85	28	0.8	500	2.3	small	negative	normal	87%	Treat	No treat
8A	81	27	0.9	496	2.3	small	negative	normal	87%	Treat	No treat
9	62	25	0.6	550	1.5	large	positive	abnormal	17%	No treat	Treat
9A	57	24	0.7	546	1.4	large	positive	abnormal	17%	No treat	Treat
10	62	25	0.6	500	1.5	large	negative	abnormal	36%	No treat	Treat
10A	57	24	0.7	496	1.4	large	negative	abnormal	36%	No treat	Treat

*Case IDs ending in "A" represent matched cases

**Abnormal life expectancy is defined as >90% chance of death within five years

Appendix 3A: Abstracts Presented at the Medical Decision Making Conferences in 2013 and 2014

The following two pages contain copies of the two abstracts addressing the research questions: (1) which readers should be paired with whom, and (2) which cases should be double read? The abstracts were accepted and presented at the Society for Medical Decision Making Conference in Baltimore, USA in October 2013 and Antwerp, Belgium in June 2014, respectively.

Gadala, Marwa, Strigini, Lorenzo, Povyakalo, Andrey, Ayton, Peter. "**Double Reading of Mammograms: Effectively Pairing Readers with Diverse Skills to Improve Performance**" *The 35th Annual North American Meeting of the Society for Medical Decision Making.* Smdm, 2013.

Purpose: Double reading is standard practice in breast cancer screening programs in over 12 countries. We retrospectively investigated whether its benefits can be increased by forming complementary reader pairs according to indicators of ability, as per published guidelines.

Method: We used data from an independent UK clinical trial where 50 readers each read 180 mammograms - 60 with cancer and 120 normal. We selected four groups of *complementary* reader pairs in which a member of a Group A, expected to be more effective, is paired with a member of a Group B, expected to be less effective. The complementary AB groups are: (1) high and low experience (recommended by UK NHS), (2) high and low specificity, (3) high and low sensitivity, and (4) high sensitivity and low specificity readers. For each group, all possible AB double reading pairs were simulated using the OR recall rule. We compared sensitivities and specificities of these complementary pairs first to those of homogeneous (AA, BB) pairs, and then to each other. Statistical significance was determined using Welch's t-test and 95% confidence intervals. To weigh sensitivity and specificity benefits, ROC curves, Youden's indices, and positive likelihood ratios were compared.

Results: Grouping according to sensitivity and according to specificity significantly increased sensitivity by 3.5% (p=0.0009) and 1.7% (p=0.037) respectively, compared to homogeneous pairings, with no significant effects on specificity. Grouping high sensitivity and low specificity readers produced a sensitivity of 0.918 (95%CI: 0.913, 0.924), significantly higher than all other groups. Grouping according to experience produced a sensitivity of 0.852 (95%CI: 0.844, 0.859), significantly lower than all other groups, but also the significantly highest specificity, 0.722 (95%CI: 0.707, 0.737). The bootstrap method for ROC comparisons applied to pAUC (sensitivity >0.8) shows that Group (1)'s most extreme readers have the highest performance (non-significant). However, Youden's indices and positive likelihood ratios show significant differences between the complementary groups with Group (1) still being the highest, followed by Group (3).

Conclusion: Some forms of pairing by complementary ability levels can significantly improve sensitivity, with an insignificant effect on specificity, compared to homogeneous pairings. These preliminary results suggest that pairing by sensitivity yields the best clinical performance, and should be further investigated. Pairing readers simply according to convenience could be significantly less effective.

Gadala, Marwa, Strigini, Lorenzo, Povyakalo, Andrey, Ayton, Peter. **"Identifying and Discussing Uncertain Cases in Breast Cancer Screening Programs"** *The 15th Biennial European Meeting of the Society for Medical Decision Making.* Smdm, 2014.

Purpose: We retrospectively investigated whether double reading benefits can be increased by isolating difficult cases for discussion.

Method: We used data from a UK clinical trial where 50 readers each read 180 mammograms (60 with cancer and 120 normal) with and without CAD. Readers marked cases as 1:Recall, 2:Discuss but probably recall, 3:Discuss but probably not recall, and 4:Not recall. We simulated 49 pairs using the OR recall rule by pairing each reader with "AvgReader", an average reader according to the group's Youden's index. These pairs' performance was used to represent traditional double reading of all cases. In comparison, we simulated a protocol where cases marked sure (1 or 4) by the first reader, were double read by AvgReader, but those marked unsure (2 or 3) were *discussed* with another reader. Discussion was modelled according to four alternative scenarios: Ideal (correct decision), Ideal limited (correct decision when readers disagreed), Authoritarian (decision of more experienced reader), and Confident (decision of more confident reader else OR recall). Statistical significance was determined using Fisher's Exact Test and 95% confidence intervals (reported in square brackets).

Results: 16% of opinions were unsure. Compared to sure cases, unsure cases were nearly five times significantly more likely to change after seeing CAD and nearly five times significantly less likely to be correct, with 0.465 probability of being incorrect. Double reading all cases with AvgReader yielded average sensitivity 0.834 [0.822, 0.847] and average specificity 0.719 [0.698, 0.739]. In comparison, discussing unsure cases with the best reader yielded: (1)Ideal: sensitivity 0.857 [0.844, 0.869], specificity 0.829 [0.807, 0.850]; (2)Ideal limited: sensitivity 0.842 [0.830, 0.854], specificity 0.820 [0.800, 0.841]; (3)Authoritarian: sensitivity 0.814 [0.804, 0.825], specificity 0.799 [0.778, 0.819]; (4)Confident: sensitivity 0.812 [0.802, 0.822], specificity 0.811 [0.791, 0.830]. Modelling with poor or average readers improved sensitivity in ideal and ideal limited discussion, and always increased specificity with varying significance and non-significance.

Conclusion: Readers' statements of uncertainty are strong predictors of error and change, and using them to select cases to discuss with another, good reader is attractive. Our simplistic modelling shows significant improvements, especially in specificity (up to 11%), suggesting this protocol is worth experimental exploration.

Appendix 4A: List of Target Words

These words and their division are part of the publication by Fischer (1985) and were used as targets in our passages.

Level 1: phonetic realization is reasonably close to the orthographic representation

Examples: Yam inflate adverb vortex cameo harp terminated trump vacate update vibrated mandated compensates delimit zebra blunder emit boxer repent intertwined uncover diplomat retort canister clustering undiminished terminology mask manifestation definitions frustrated expectation alternate stimulation examiner preventive unemployment punishment establishing electronics

Level 2: contain an ambiguous segment involving some departure from straightforward phonetic mapping that may require application of established orthographic conventions

Examples: Strapped skilful cancelled picnickers abridgment flier changeable sincerely echoes disbelieve sobbing beige skies unperceived clannish noticeable ninety thinned basically definitely misspell aggravate commemorate defensible grammar clearance inexhaustible utterance continuance prevalent dissimilar preferring inspiration omitted repetition indigestible recommend regrettable equipped commiserate

Level 3: contain segment(s)/pattern(s) that don't generally occur in English or occur with low frequency sometimes because they are borrowed from other languages

Examples: Chihuahua onomatopoeia Fahrenheit plagiarism sarsaparilla haemorrhage sergeant connoisseur mnemonic reveille desiccate syphilis hieroglyphic thumb lengthen Wednesday soldered subpoena kaleidoscope thigh slaughter indebted climb answering knock beautifully laugh tongue folk bourgeois listener gnaw talker rhododendron diphtheria annihilate pyorrhoea pygmy sacrilegious

CITY UNIVERSITY LONDON

Appendix 4B: Spell Checker Experiment Recruitment Poster

Decision-Making with the Aid of an Automated Tool

Researchers at City University London would like to invite you to take part in a study exploring decision-making with automated tool support. This study has been granted favorable ethical opinion by the City University London Ethics Committee.

What is involved in the study?

In the study, you are asked to find and correct all spelling errors in a series of short passages of text with varying support from a unique spell checking tool. People of all spelling abilities are welcome to participate.

How long will the study take?

The study is expected to take **no more than one hour** to take place at a convenient time of your choice in a computer lab on City University campus. If at any point you wish to withdraw from the study you can do so without providing any reasons for withdrawal.

As a token of appreciation, you will be offered an **£8 Amazon voucher**.
Furthermore, all participants who correctly edit all errors in the passages will be entered in a **draw for an £80 Amazon voucher**.

How will my answers be used?

All decisions will be confidential, anonymized and used only for the purposes of this study. The objective of the study is to measure the effect of the spell checking tool, not the level of your spelling ability.

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Decision-Making with the Aid of an Automated Tool
Register at
Contact Marwa Gadala at [redacted] for questions

Appendix 4C: Spell Checker Participant Information Sheet

CITY UNIVERSITY LONDON

Decision-Making with the Aid of an Automated Tool

Introduction

Researchers at City University London would like to invite you to take part in a multi-disciplinary research study exploring decision-making with automated tool support. Our aim is that this study will shed light on important factors in the design of automated tools, and your participation will be valuable in contributing to this knowledge.

Before you decide whether you would like to take part in this study, it is important that you understand why the research is being done and what it would involve for you. Please take the time to read the following information carefully and discuss it with others if you wish. Also, please ask us if there is anything that is unclear or if you would like more information.

Procedure

Participation is voluntary and you can withdraw from the study at any time. It is up to you to decide whether or not to take part. If you do decide to take part, you will be asked to complete a consent form. If you decide to take part you are still free to withdraw at any time and without giving a reason. People of all spelling abilities are welcome to participate, and the objective of the study is to measure the effect of the spell checking tool, not the level of your spelling ability.

The study is an online application which asks you to find and correct all spelling errors in four one-page passages. Following this, you will be asked to answer a short questionnaire and to spell a list of words, of varying difficulty. The entire experiment is expected to take between 45 - 60 minutes of your time. You will be aided by a spell checking tool. However, unlike conventional spell checkers, this tool is primarily built to catch homophone errors (words that are pronounced the same but differ in meaning and spelling such as "to"/"two"/"too").

Benefit to Participants

After signing the consent form you will be allowed to choose a suitable time to come in to do the experiment. It takes approximately **45-60 minutes** to complete and in appreciation of your time, you will be offered an **£8 Amazon voucher**. Furthermore, all participants who correctly edit all errors in the passages will be entered in a **draw for an £80 Amazon voucher**. At the end of the study, we can also send you a **copy of the corrected passages**, which we hope you will find interesting to compare your performance to.

Research Ethics

All data stored will be anonymous and confidential and will be used solely for the purpose of this study. Results may form sections of submitted papers to peer reviewed journals. Participants will be contacted only to set up lab time and to provide a copy of the corrected passages. After that, all email addresses will be destroyed.

This study has been reviewed and given a favorable opinion by the City University London Research Ethics Committee.

If you have any questions or comments about the research, please Mrs. Marwa Gadala at marwa.gadala.1@city.ac.uk.

If you would like to complain about any aspect of the study, City University London has established a complaints procedure via the Secretary to the Senate Research Ethics Committee. To complain about the study, phone 020 7040 3040. You can then ask to speak to the Secretary of the Senate Research Ethics Committee and inform them that the name of the project is: Decision-Making with the Aid of an Automated Tool.

You could also write to the Secretary at:

Anna Ramberg
Secretary to Senate Research Ethics Committee
CRIDO
City University London
Northampton Square
London
EC1V 0HB
Email: Anna.Ramberg.1@city.ac.uk

Thank you for taking the time to read this information.

[Button to proceed to the Consent Page]

Appendix 4D: Spell Checker Experiment Consent Form

Consent Form

Title of Project: Decision-Making with the Aid of an Automated Tool
Researcher: Mrs. Marwa Gadala, City University London, marwa.gadala.1@city.ac.uk

I confirm that I have read the participant information for the above study (located on the previous page), and that I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.

I understand this study will involve:

* using a computer to edit a series of passages for spelling errors,
* using a computer to listen to a list of words and spell them to the best of my ability,
* completing questionnaires asking me about my linguistic background and then my perception of the tool.

I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without being penalized or disadvantaged in any way.

I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports, any publications, or to any other party.

□ I have read and understand the participant information and consent page and agree to take part in this study

[Button to return to the Participant Information Page]

[Button to proceed to demographics only if above box is ticked]

Appendix 4E: List of Assessment Words

The following words make up the dictation test. There are 10 words from each level and they are based on the division and lists published in Fischer's paper (1985).

Level 1 Words: carpentry, compendium, gastronomy, infiltration, refreshment, conjugate, dogmatic, hollow, inactive, profile

Level 2 Words: asymmetric, gracious, homogeneous, indefinite, misshapen, clannish, metallic, recurring, relieve, youthful

Level 3 Words: bureaucracy, chandelier, diaphragm, gauge, jeopardy, manoeuvre, marijuana, plagiarism, rendezvous, dialogue

Appendix 4F: Copy of Passages

The following four passages were used as part of the spell checker experiment. They were written by Marwa Gadala and target words and homophones were deliberately inserted. We highlight in bold all target words and underline homophones. However, this highlighting does not mean they were highlighted by the tool. We present the passages as they appeared to the participants, including spelling errors and incorrect usage. The first two passages were considered easy while the second two were considered the more difficult ones.

How to Save a Wet Phone

Have you accidentally left your cell phone in your pocket and run it through the washing machine? Have you gone for a swim and forgot it in your pocket? Maybe you've managed to knock it off your bed into a glass of water. Or perhaps it fell out of your pocket and into the water bowl of your **Chuahua** pet? Well, acting quickly can make all the difference. Don't let the gilt eat at you just yet because there is a chance you can still save it. Here's how:

1. Retrieve your phone and quickly turn it off.

Resist **establishing** whether the phone is still working by trying different keys because this **stimulation** can further **inflate** your problem by shifting liquid deeper into the device. Detach any chords and remove the battery (hopefully it's not **soldered** in). This is important because many **electeronics** inside the phone will survive immersion in water provided they are not attached to a power source (battery) when wet. Also remove the SIM card because some or all your valuable data, such as contacts, is stored there. For many people, these might be more valuable than the phone itself.

2. Remove any **noticable** water.

Take the phone apart as far as you can, then dry it and all its parts with a soft towel. The key is to do this quickly.

3. Remove the water you can't see.

To do this, try using a vacuum cleaner. Be careful not to hold the vacuum too close to the phone, as a vacuum can create static electricity, which is essentially worse for the phone than water. You may have heard suggestions to use a hair dryer to **vacate** the moisture, but we do not **recomend** this, because even on the "cold" mode, a hair dryer can force moisture further inward. Other suggestions include placing the phone (minus the battery) in an oven, but this can altar components and melt the ceiling and adhesives. As an alternate solution, place the phone in a substance with a high affinity for drawing out moisture. For example, place it in a **canister** with silica gel packs - the kind used in shoeboxes - to **dessicate** the moisture. If you don't have any, a similar but inexpensive option is to place the phone in a bowl of uncooked rice.

4. Now you need to wait.

The desired waiting time is 72 hours, but **definitely** wait at least 24 hours before attempting to switch the phone on. This may be the hardest step of all. In our day and age, it's a **regrettable** situation that may **aggravate** you, but you'll survive - even if you're an avid **talked** or sincere Facebook user.

Unfortunately, even if you are able to revive your phone, there's still a chance that the water contact has made some **unperceived** changes. So if a year later, your phone begins to die a slow and painful death, you'll know it's because of some of the undiminished effects. You also cannot claim for breech of warranty as your warranty is now void. Many companies have mandated that if you've somehow immersed your phone in water, the blunder is all on you, and you can't mask the truth as all phones have a water sensor that will uncover if your phone has gone for a swim. In the end, if all your effort is lost in vein and you need to by a new phone, perhaps consider one of the **beautifully** crafted waterproof ones?

Dear colleagues,

We are organizing a follow-up meeting this Wednesday to discuss the recent complaints we received regarding the new **beige** child pushchair model (#A324) we released last March; so please be sure to mark your calenders. As most of you are aware, to date, there have been five recorded incidents where children have caught their fingers in the buggies' hinge mechanisms, resulting in three fingertip amputations, one **thump** amputation, and one minor **haemorrage**.

The principal reason for the follow-up meeting will be to discuss how we will handle questions from the press. We are already receiving a large number of calls from frustrated and concerned clients as well as from the media, and we want to insure that our statements and terminology compliment our print strategy. As a distinguished and well-recognized manufacturer, we must discretely manage information and the affects of any negative publicity.

We are certainly treating these reports with utmost priority and we **comiserate** strongly with the families of the injured children. The family claims are currently being investigated and we will update you regarding results of the investigations as they unfold. There are a few reluctant families, and the court may serve a **subpena** on them to disclose the root of the accidents, and our expectation is that this may **lengthen** the process.

We have currently **cancelled** all future orders of this model, including fifty thousand orders that were due for distribution next week. We also propose to disperse a compinsation to affected clients, and that we delimit or possibly terminate the production of this model. In a **preventive** approach of future accidents we may offer free hinge covers to customers, hand out a **flier** to purchasers, add a label to warn parents to prevent children's fingers from clustering around the hinge as the buggy is being folded or unfolded, or go as far as issuing a complete recall of the product. There are many **intertwined** issues, and until we are **equipped** with the results of the investigations, it is difficult to determine weather a complete recall is the best course of action. We need to keep the number of accidents in perspective, especially since we have sold more than a million of these buggies around the world. Furthermore, in a logical **retort**, we strongly believe that if a buggy is folded or unfolded in line with our operating instructions and the same level of care is taken as when opening or closing a car door, then the risk of injury is practically non-existent.

If until we meet, you receive any calls you cannot handle, please forward them to Susan or me. I take this opportunity to thank Susan, a natural **diplomat**, whom our administration is **indebted** to for her competence in responding to all queries related to this issue.

Sincerly,

John Smith

The Battle Against Malaria

It is a serious threat to human life having claimed more lives than any ruthless **sargent** or **sacrilegious** invasion of troops. In fact, scientists theorize that it may have killed around half of all humans who have ever lived. The troops themselves are so miniscule in size, although they pose no **pigmy** problem robbing families, communities and whole continents of their livelihoods and prosperity. In their raid, these troops have mostly killed young children, especially those in poor and remote areas, but they have also affected **bourgeoisie** citizens and even prominent individuals in history such as four famous American Presidents. In fact, the faint buzz of a single member of these troops is no trivial **onomatopoeia** but a harsh cue of a deadly threat.

This threat is malaria: an infectious disease caused by the plasmodium parasite and transmitted by the female anopheles mosquito. It is both preventable and treatable, yet to this day, it continues to afflict some 200 million people each year, with a current death toll of half a million per year. Symptoms include high fever - over 100.4 **Farenheit** - along with chills, headaches, vomiting, and diarrhoea generally immerging between 7 and 18 days after the mosquito **gnaw**.

The battle against malaria is in no way easy because of scientific, economic, and cultural challenges. Scientifically, it's a complicated disease in comparison to other diseases managed in the last century such as **diphtheria**. It's true that much has happened since the early use of **sarsaparila** to fight malaria, with latest test results showing an inoculation vaccine currently in development to be partially effective. However, over the history of malaria, two of the drugs formerly used to treat it became useless due to mutating parasites; so it's also important to ensure that the drugs used remain effective. Interestingly, recent research established that sleeping under bed nets is "by far the most important intervention" to date, but new weapons in the form of scientific breakthroughs are still needed, and lately even **defensible** ideas of embracing mosquito eaters - such as ticks and lizards - have not been **omitted**.

Economically, malaria can be difficult to tackle as it often affects people in poor and remote areas. Recently, cultural challenges although previously prevalent, have also come into light. Often those that stand to benefit the most from the fight against malaria care about it the least. Many of them perceive malaria as a normal everyday problem and view the bed net as an inconvenience, thus preferring to sleep in the outdoor breeze. Their reaction is not too **dissimilar** to our response to educational campaigns about **pyorrhea**. That is why more money is currently being spent in education and awareness to advise people, council them, incite them to use the available resources, and overall illicit their help to fight the disease.

In September 2015, a new ambitious goal was set - to reduce malaria incidence and mortality by a further 90% by 2030. With this **ninety** percent reduction goal in cite, experts estimate 64 billion British pounds are needed by 2020. Although an ambitious aim, it is a goal with a clear plan, and the past fifteen years are a true **insperation** showing how colossal commitment and funding have **thined** the malaria map and saved millions of lives.

So this is no time to **commemorate** our success or to stand idol; instead, we need to redouble our efforts to defeat malaria. History has repetitively shown that if our will starts to falter, we can lose all the progress gained in the past fifteen years in a single year of resurgence. Although we should not **disbeleive** that success is possible, the plan to **anihilate** malaria will only materialise through **continuence**: if countries commit to sustaining current levels of funding, new donors step up their commitment, and we grab the reigns in search of new advances to tackle the killer. But perhaps with

so much at steak, the most lasting way forward is attacking the disease according to the priorities of the **folks** who live with it. We need to encourage their help by launching proper education, and attacking poor roads, bad houses, deficient drainage and rural poverty in an effort to terminate not only malaria, but any malarious ways of life.

Four Habits of Highly Effective People

Your life doesn't just happen - it is carefully designed by you. You choose happiness or sadness. You choose decisiveness or ambivalence. You choose success or failure. And within these choices lies the ideal opportunity to do things differently to get better results. But how do we achieve the success so many of us choose? Some think it depends on a combination of desirable circumstances. But sometimes people are better placed, smarter, more **skilful** - yet less successful. Instead, success seems to be linked more to character than circumstances. In turn, character is basically a composite of habits. Thus, one licence to success is identifying the habits that can help lead to it. As the maxim goes: "Sow a thought, reap an action; sow an action, reap a habit; sow a habit, reap a character; sow a character, reap a destiny." As a start, here is a short summary of four of the seven habits published by Stephen Covey in his top-seller book, "The Seven Habits of Highly Effective People".

Firstly, highly effective people are proactive, focussing on the "Circle of Influence" that lies within their "Circle of Concern". They focus on what is personally **changeable** which results in **inexhaustible** energy, hard work, responsibility, and dedication - all factors which really pay off. It's not that these people have a small "Circle of Concern" which does not encompass big issues - for example, terrorism or the relentless **sloughter** of innocent people - but that they ask what they can do and then flex their proactive mussels rather than stand stationery and dwell on what they can't control. Reactive people, on the other hand, respond to external factors by focussing on things that are in their "Circle of Concern" but outside their "Circle of Influence". Their frequent **sobing** about external factors, such as genetics or circumstances causes them to become **eunuchs**, to feel increasingly victimized, and to end up passively **strapped** in their seats believing that the problem is "out there" - when that in itself is their problem.

Secondly, highly effective people begin with the end in mind. The easiest way to understand this is to imagine yourself at your own funeral: what do you want to hear about your life? If you aren't envisaging what you would like to hear, then you need start paving a new road for your life. Most of us find it rather easy to be busy, but it's possible to be very busy without being effective. It's equivalent to working harder and harder to climb a ladder only "to discover that it's leaning against the wrong wall". That is why it's important to drop the **kaleidoscope** image, and take the time to create a clear vision of who we truly are and what we really want in life; otherwise, we wave the right to shape our life and give other people and circumstances clearence and ascent to shape it - a threat of immanent dissatisfaction and displeasure.

Thirdly, highly effective people put first things first. They are able to organize and manage time in a manner that helps them move forward, closer to the priorities they've **established**. Because they have a clear vision of their desired destination, their day-to-day decisions are not based on a **hieroglyphic** map of life. They've trained their **tongs** to be comfortable **ansering** "no" to matters that don't contribute to their mission. The simple **uterance** of this two letter word helps them filter through the frequent noise in the world. They successfully minimize distractions, easily laugh off dispiriting people's opinions, and avoid habits of lazy **picnickers** or regular gossipers. In essence, they practise the art of discipline until they master it, which allows them to manage their time based not just on what is most urgent, but also what is most important.

Fourthly, highly effective people invest in themselves - the greatest asset they have. They are strong **liseners** to their needs. Covey terms it "sharpening the saw" meaning devoting time to renew yourself which synergistically increases your ability to exercise each of the other six habits and creates the upward spiral of growth that progressively lifts you to higher **skyes**. It's about taking the time each

day to improve yourself spiritually, physically, mentally, and socially. Although this habit is in no way **indigestible** or difficult to understand, you would be surprised how many people complain they don't have the time or energy to make changes such as: reading for 15 minutes a day, going to the gym, learning a marshal art, broadening their musical pallet, becoming a **connoisseur** of art, etc.

Perhaps this **abridgement** inspires you to read about Covey's other three habits or even his entire book, but more importantly perhaps it comes as a **revile** to many of us to make active changes in our lives. And hopefully we won't forget these four habits - maybe a **nemonic** will help us recall them!